

Continual Planning for Cross-Modal Situated Clarification in Human-Robot Interaction

Geert-Jan M. Kruijff

DFKI GmbH
Saarbrücken, Germany
gj@dfki.de

Michael Brenner

Albert Ludwig University
Freiburg i.B., Germany
brenner@informatik.uni-freiburg.de

Nick Hawes

University of Birmingham
Birmingham, U.K.
n.a.hawes@cs.bham.ac.uk

Abstract—Robots do not fully understand the world they are situated in. This includes what humans talk to them about. A fundamental problem is thus how a robot can clarify such a lack of understanding. This paper addresses the issue of how a robot can create a plan for resolving a need for clarification. It characterises situated clarification as an information need which may arise in any sensory-motoric modality required to interpret the situated context of the robot, or any deliberative modality referring to that context. It then focuses on how, once a clarification need has been identified, the robot can create a plan in which one or more modalities are used to resolve it. Modalities are involved on the basis of the types of information they can provide. These information types are identified in the ontologies the modalities use to interconnect their content with content of other modalities (via information fusion). We take a continual approach to planning and execution monitoring. This provides the ability to re-plan depending on modality availability and success in resolving (part of) a clarification need. We illustrate the implementation on several examples.

I. INTRODUCTION

The real world is a rich and complex environment. A robot in this environment may not always fully understand what it is seeing or hearing, what it should do, or how it should do it. Our stance is that an intelligent system should have the ability to try to figure such things out. If a robot can try to understand the world on its own terms, or fail gracefully whilst trying, it will be a less brittle system than those currently presented in the literature.

This is where *clarification* comes in – and where this paper starts. We focus here on the clarification problem, which we consider a sub-problem of the problem of a robot satisfying its general *information needs*. In clarification tasks we assume that the robot is able to identify that it lacks information to *complete* or *disambiguate* something it already partially understands. The robot is able to identify *what* it needs to clarify.

In this paper we address how the robot can develop, and flexibly execute, a plan to resolve a clarification need. We discuss how a clarification need can arise in a robot’s cognitive architecture when it is trying to understand something about the situation. We are interested both in clarifying aspects of the situation itself, and aspects of the situation relative to a plan or an utterance. We present an approach which conceives of such a need for clarification as a *clarification request* (CR; cf. [8], [20]). There are two principle ideas behind a CR. First, it captures what is requested about the perception of the situation, or the relation between the

perception and deliberative content tied to it. Generally this can be considered to represent what information the robot is missing about the situation. We propose a compositional way for formulating these requests, so they can have any level of complexity in combining aspects to be clarified. Second, a CR is built such that we can translate the description of missing information into a high-level plan for obtaining that information: a CR captures hints for its own resolution.

Given this description of a CR, we approach planning for clarification resolution using the following process. Initially a high-level plan specifies what modalities could be queried to yield particular pieces of information. This enables us to use any combination of modalities to resolve a clarification request. What modalities are used on a particular occasion depends on the status (content, availability, etc.) of those modalities at that point. The planner inspects each prospective modality to determine whether it could take care of a part of the resolution of the clarification request. This eventually yields an instantiation of the high-level plan in terms of several modality-specific plans, which are subsequently executed to yield an answer to the request.

Contributions. Clarification has been defined primarily for dialogue, as a means to overcome a breakdown in communication [7], [21]. It has been investigated in human-robot interaction (HRI), to clarify aspects of spoken dialogue [18] and, to a limited degree, aspects of the situated context [16]. This paper presents several extensions over this work. It generalises CRs to the different levels at which clarification needs can arise with respect to situated meaning [7], [21], [22], tying it into an ontology-based approach to “grounding” [15], [13]. This makes it possible to plan the use of any viable combination of modalities to resolve a CR on the basis of the ontological characterisations of the types of information these modalities can provide. [20], [16], [18] only use dialogue to clarify.

Overview. §II discusses the concept of clarification, relating it to current notions of situated meaning and symbol grounding. §III presents our approach, including a definition of the notion of situated clarification, and how the resolution of a CR can be planned for using a continual approach to planning and plan execution. §IV illustrates the implementation of the approach on real-life examples.

Acknowledgements. The research reported of in this paper was supported by the EU FP6 IST Cognitive Systems Integrated project *Cognitive Systems for Cognitive Assistants*

II. BACKGROUND

Clarification is a relatively common device to overcome a breakdown in communication [21]. It is used when one of the dialogue participants fails to understand what another one said. Only once the participants have clarified misunderstood issues, can the dialogue proceed.

Clarification can refer to different aspects of what is being communicated. It is one of many ways in which people try to manage communication, acting as part of a *grounding* process which interacts with linguistic understanding (e.g. [19]). With clarification, a dialogue participant provides a form of *back-channelling* [25], [24] indicating that she fails to ground. This grounding can be relative to the different *levels* at which an utterance is interpreted in a dialogue [30], [7], [1]. The exact definitions of these levels vary across theories. Table I gives the division we adopt, based on [7], [1], [22].

	Level	Kind of problem	CR example
1	ATTENTION	Channel	“Excuse me”
2	IDENTIFICATION	Acoustic problem	“What did you say?”
3	RECOGNITION	Lexical problem	“What is a PTU?”
		Grammatical problem	“Is the ball near the mug, or do you want me to put it there?”
		Referential problem	“Which?” “Where?”
4	CONSIDERATION	Problem with recognising/evaluating intention	“Why?”

TABLE I
LEVELS OF CLARIFICATION; ADOPTED FROM [22]

Table I discerns different form/function relations at which an utterance can be considered, and provides examples of the kinds of CRs that can arise at each level. ATTENTION establishes the communication channel between the participants – e.g. through eye contact. IDENTIFICATION recognises words in the acoustic signal, and provides the first mapping to a symbolic level that can be linguistically analysed. RECOGNITION analyses the structure of an utterance, to build up a (linguistic) interpretation. This covers the grammatical structure of the utterance, and the resolution of how the content of this utterance relates (co-refers) to content mentioned in the preceding dialogue. CONSIDERATION interprets the utterance meaning in terms of why it was said, and how it furthers the dialogue as engagement [7], [1], [26].

The relation between these levels is *functional*. A representation built at one level is interpreted at the next as having a particular function there. We use syntactic structure to contribute to forming linguistic meaning, and we interpret linguistic meaning further to see how it fits into the dialogue context. Furthermore, levels interact in parallel, mutually guiding and constraining the interpretations that should be considered in the current context (e.g. [2], [29]).

We can extend this view on levels of *linguistic* interpretation to any form of *information processing*. Semiotically speaking, each level acts as a *sign* for the formation of an *interpretant* at the next level, relative to an external *object* of interpretation. Clarification needs can then arise relative

to any level of form (the sign), its interpretation (internal to the robot), or object (reference to a situated context). This is crucial to situated dialogue in human-robot interaction, as it extends the need for grounding communicated meaning not only in linguistic understanding, but also in how the world is understood and what is to happen in it.

Several approaches have been recently proposed to model “symbol grounding” of linguistic meaning, to arrive at the situated meaning of a linguistic expression [23], [27], [28], [10]. These approaches all try to capture the relation between sensory experiences, and symbols used in deliberative processes. Steels proposes a notion of *semiotic networks* [27]. These networks are dynamic systems in which multiple levels of sensory experience are related to categorical and linguistic interpretation. What is important about these networks is that they capture a notion of *affordance*. Experiences and symbols are related to make clear what is possible in a given context. This thus lends the symbols not only an intensional dimension, but also an *intentional* one. Similar to Steels’ semiotic networks are Roy’s semiotic schemata [23], [10].

None of these approaches, however, have studied how clarification can enter situated dialogue – particularly, in connecting language to the world. Clarification and models of linguistic grounding have been studied in detail in non-situated dialogue systems, e.g. [30], [17], [20]. In HRI they have only been studied to a limited degree. Li et al. [18] present a basic model for communicative grounding in HRI, similar to [30]. Kruijff et al. [16] discuss an approach to clarification in human-augmented mapping, based on [17].

In this paper, we extend the idea of clarification to situated meaning. Similar to Table I we can apply the aspects of channel, signal, extension/intension, and intention to how content of a particular modality (sensory-motoric or deliberative) can be grounded in the situated context. Abstractly put, the situated meaning of a sign is the interpretation we can give to it in the current situated context. What counts as a sign depends on what level of understanding we are considering [27]. Situated meaning is thus not a “monolithic” concept, but a *holistic* concept [23]. It arises from, and is constituted by, how different levels of understanding interact and eventually establish mutually supported interpretations.

A fundamental aspect of situated meaning is that signs can be interpreted across levels or modalities that may have different representational *forms*. Like [27] we adopt a model for relating representations that is based on *mediation* [13]. Content from different representations are not linked directly. Instead, they are linked through ontologies that capture shared characteristics of the the linked representations. This has two distinct advantages. First, the ontologies provide *a-modal* characterisations of content. This is an important level of abstraction which enables us to connect any number of modalities, as long as they provide mappings from their modality-specific representational forms into the a-modal format. Second, we *link* representations rather than fusing them. This means that we have both the a-modal representation of a sign, and its modality-specific representation (cf. also [3]).

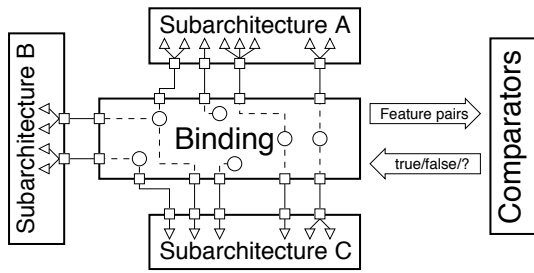


Fig. 1. The binder operates between the other subarchitectures by interpreting the contents on their working memories. The modal representations are translated into proxies (squares), described in detail by a set of features (triangles). Some pair of features are comparable and proxies with comparable and matching features may be combined into unions (circles). The unions are a-modal representation of objects, relations and actions, while maintaining references to the modal representation of the features.

Figure 1 illustrates this idea in more detail [15], [13]. In a cognitive architecture, each subarchitecture (a modality- or function-specific subsystem of the entire architecture) maintains an ontology that specifies the mapping between its own representations and a-modal representations. In addition, each subarchitecture has a binding monitor, a process which monitors the subarchitecture’s working memory. Every time the working memory contains content that could be connected to content in other modalities, the binding monitor translates this content using a mapping between the subsystem’s own representational formalism, and the a-modal format used in the binding subarchitecture.

III. APPROACH

In §II we discussed how situated meaning arises through an interplay of sign interpretation processes at different levels of comprehension. Signs can be interpreted across modalities through ontology-based mediation, dynamically giving rise to interpretations that may be formed over time. We can extend clarification from the purely linguistic domain to situated meaning by seeing it as part of this more general, bi-directional grounding process. We have the top-down connection of symbols to sensory experience, and we have the bottom-up abstraction of sensory experiences to symbols that represent them (cf. [27]). Clarification needs may arise along both lines to overcome a problem in relating signs and representations across levels of comprehension.

A. Situated clarification

Figure 2 specifies the different dimensions along which issues in need of clarification can arise when understanding an observation (bottom-up, or top-down) in the process of building up models of the situated context. We divide the situated understanding of an observation into *comprehension* (extensional and intensional understanding), and *purpose* (intention). In this paper we focus on comprehension. Comprehension captures what we are trying to understand about the sign itself, including both intensional and extensional aspects. We characterise these aspects as the “what?”, “how/where?” and “whether?” properties.

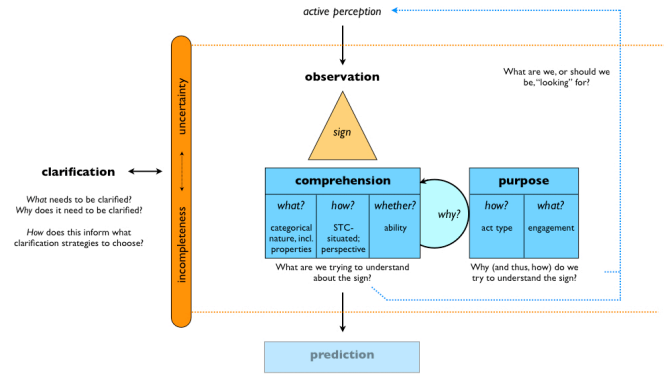


Fig. 2. Clarification dimensions

The “what?” covers the categorical nature of the sign. Clarification regards the type and associated properties of the sign, and can arise at three basic phases. From the sensory projection to the (modality-specific) analog interpretation, there may be unclarity in the basic sensory properties of the sign. For example, audio processing may be unsure whether the input is noise or a sequence of phonemes – or, which phonemes are recognisable. Similarly, colour in visual perception may vary strongly under lighting conditions. The next phase is at the abstraction from analog interpretation to discrete symbolisation, following a modal categorisation. Here, the system may be unable to form a symbolic representation (e.g. unable to parse a sequence of words), or be uncertain how to categorise a sensory experience (e.g. uncertainty in visual categorisation). Finally, we have the step in which a modal categorisation is translated into an a-modal binding proxy. An example here is the translation of a modality-internal soft categorisation of visual perception into a discrete “hard” categorisation used in a-modal binding.

Comprehension also includes the spatio-temporal-causal dimensions of sign interpretations, and the ability to perceive, or perform an action. These dimensions are inherently cross-modal. They relate modality-specific information with information about the larger spatio-temporal context in which the modal interpretations are set. This crucially relies on the availability of a modality-internal history [15]. Clarification needs regarding the spatio-temporal-causal (STC) nature of signs may arise at the same phases as categorical clarification needs do. Spatio-temporal clarification primarily concerns re-identification of signs (and their interpretations) across spatio-temporal contexts. Visually, this may yield requests such as “is this the same object as seen before?”, and “where was the object some time ago?” Linguistically, the history is the dialogue model, and internal spatio-temporal resolution means relating the current utterance to the preceding context.

Typically, STC clarification requests involve cross-modal information processing, e.g. dialogue requests about the location of items or rooms. Here we need a reference to a spatio-temporal context, in which the request can be placed. This reference is combined with the modality-internal history to determine the spatio-temporal context for resolving the

clarification request. Currently, we assume that the ability aspect of comprehension in principle follows the same information processes as those of the STC-based resolution.

Given this characterisation of how clarification may arise in situated meaning (Figure 2), we now present a formalisation for stating such clarification requests. Ginzburg [8] models the content of an “utterance” (i.e. *sign*) u for an agent A to consist of the literal meaning $\mu(u)$ and the purpose A assigns to u , $GOALS(A, u)$. The literal content of u models the intension and extension of the sign, whereas the purpose captures its intentional aspect.

$$content(u, A, \mu(u)) =_{def} \{(u, \mu(u)), GOALS(A, u)\} \quad (1)$$

(1) captures the (composite) relation between the literal meaning and the purpose of an utterance. When clarifying an utterance, this means questioning one or more aspects of content. Representing “questioning” as an operator $?$ (question mark), then $u?\mu(u)$ questions whether the literal meaning is applicable, and $GOALS(A, u)?$ indicates the need to clarify the intentional aspect. (2) defines the composite question.

$$content(u, A, \mu(u))? =_{def} \{(u?\mu(u)), GOALS(A, u)?\} \quad (2)$$

Following [9] we formulate a compositional way in which aspects of content can be questioned. We define a property x as $x : property(x, P)$, and a proposition as $prop(X)$ with $X = x_1, \dots$ a finite number of bound variables over such properties. A question about the polarity (“truth”) of a proposition is defined as $?prop(X)$ (“is it the case that the proposition holds?”), and about a value assignment to a property as $?x : property(x, P \in \{v_1, \dots, v_n\}).prop(X)$ with $x \in X$. ($n > 1$ models an ambiguous value assignment.) Factual issues are raised by taking the variable or property as argument of $?$: $?x.prop(X)$, or $?x : property(x, P).prop(X)$.

We extend $?$ to also range over binding proxies and unions, to deal with symbol grounding [13] (Figure 1). A proxy is a structure $prx = \langle P_p, \mathbf{Un} \rangle$ with P_p a set of one or more properties (“features”) and \mathbf{Un} the union(s) in which p participates. A union is a structure $un = \langle P_{un}, \mathbf{Prx} \rangle$, with P_{un} the properties represented in the union, and \mathbf{Prx} the set of proxies included in the union. We define the polar and factual uses of $?$ over prx, un analogous to the constructions for content. The situated meaning of u is thus:

$$sitcontent(u, A) =_{def} \{(u, \mu(u)), GOALS(A, u), \\ prx(u, \mathbf{Un}), \\ \{un \in \mathbf{Un} | un = \langle P_{un}, \mathbf{Prx} \rangle\}\} \quad (3)$$

We represent content as an ontologically richly sorted, relational structure. A statement $x : property(x, P)$ can thus be further specified into relational and ontological components: $x : ontologicalsort : \sigma \wedge x \langle relationtype : \rho \rangle y$. Depending on whether we are questioning y , or y is bound in the proposition, we can express clarification over relational structure. We can similarly question the ontological sort of x ($?\sigma$) or the relational type ($?\rho$).

Clarification over categorical properties of a piece of content can now be modelled as (polar or factual) questions over ontological sort and (material or contrastive) features represented in the proposition. STC clarification ranges over the variables and features which represent time (e.g. following [4] $@t1(Past)t2$, including both the Priorian past tense relation and the explicit time points), and the relations that qualitatively model spatial and causal relations [14]. Purposive clarification uses the $GOALS$ construct, including explicit reference to relations in the structured history that the modality employs (cf. [14] for an example of a purely relational dialogue model).

B. Planning clarification processes

Using multiple $?$ -quantifications and simple conjunctions of propositions we can subsequently build up clarification requests. The clarification request is then transformed into a goal for the *clarification planning component*. This component is a two-level planning process. On the high level it determines appropriate modalities that could (according to the properties in $\mu(u)$) potentially help resolving one or more $?$ -bound variables or propositions. On the low level it then plans modality-specific solutions to the clarification goal. On both levels the goal to achieve is an *epistemic* one, i.e. clarification planning is essentially planning for *information gathering* from appropriate modalities and/or other agents. The high-level planner does *not* distinguish between (internal) sensory modalities and other agents: both can be queried to provide additional information that may be used to bind proxies from different modalities.

For example, in order to clarify to which object in a scene the human user referred to, the robot can query the human directly (lower-level planning for this would be performed by the *dialogue* planner) or plan for its *vision* subarchitecture to perform additional scene analysis. The high-level clarification goal could potentially be resolved by both modalities. However, based on the current *situation* the clarification planning process selects the one assumed to provide the missing information most directly and then, on the lower level, plans the modality-specific execution of the corresponding request.

We model clarification as a *continual planning* process, i.e. the execution of a clarification plan is monitored and, in case of failure, adapted or rebuilt [6]. Thus, unhelpful results from one modality may lead to new clarification plans employing different ones. For example, when trying to obtain extra information about a visual referent the system could query its visual subarchitecture. If this failed, then the planner could generate a direct linguistic request to the user.

IV. IMPLEMENTATION

Our approach to clarification is developed as part of the EU-funded CoSy project¹ and will be used in several scenarios featuring human-robot interaction (for an overview of the system architecture see [11]). However, in order to

¹www.cognitivesystems.org

- (1) Anne: "Please bring me the coffee, R2D2."
- (2) R2D2 senses: *the coffee is not in the living room.*
- (3) R2D2: "Where is the coffee, Anne?"
- (4) Anne: "The coffee is in the kitchen."
- (5) R2D2: "Thanks, Anne."
- (6) R2D2 moves to the kitchen.
- (7) ...

Fig. 3. Clarification subdialogue in MAPSIM.

easily evaluate and compare variants of clarification planning with different modalities we also employ a *simulation testbed* for continual multiagent planning called MAPSIM [5]. In particular, MAPSIM is able to interleave dialogue with physical action and sensing, i.e. it is able to take the *situatedness* of the dialogue into account.

MAPSIM simulations are generated automatically by interpreting a formal multiagent planning domain description. In other words, MAPSIM treats a planning domain as an executable model of the environment itself. Crucially, the same formal model is used by the agents in the simulation to generate their actions plans or, for this paper, their clarification plans. As a result, agents not only to plan, but also *execute* their plans in the simulation. This is essential for studying *continual* planning where planning, acting and sensing is interleaved. In the case of continual *clarification* planning a (simulated) robot can thus deliberately suspend its planning process to gather additional information (e.g. to ask a clarifying question).

Figure 3 shows a clarification subdialogue in a MAPSIM scenario that involves two artificial agents who assume the roles of a robot and a human user, respectively. A clarification subdialogue arises when the robot R2D2 cannot resolve the reference to the "coffee" object mentioned by the user. In particular, the high-level plan generated by the agent specifies that the *position* of the object must be determined in order to be able to bring it to the user. This corresponds to an epistemic goal that the agent then plans to satisfy. Since "clarification" by sensing does not provide the missing information (line 2) the continual planner resorts to the subdialogue of lines 3–5 which satisfies the clarification goal and enables the robot to achieve the overall goal given by the user.

Situated clarification often involves *cross-modal reference disambiguation*, i.e. a unique mapping between object references in different modalities must be found. For example, in line 1 of Figure 4 the expression "the mug" used by Anne is stored in a linguistic binding proxy. However, it does not provide enough distinguishing features to be bound to one of the two possible visual proxies (mug9 or mug17). The high-level clarification planner therefore determines that clarifying information must be obtained from the user. A plan for this clarification goal is realised by the modality-specific (in this case: linguistic) clarification request in line 2 of Figure 4.

Figure 6 shows the definition of a clarification operator in the MAPSIM *clarification* domain, as used by the agent R2D2 in Figure 4. Informally, it can be described as follows:

- (1) Anne: "Please give me the mug, R2D2."
- (2) R2D2: "Do you mean the blue or the red mug, Anne?"
- (3) Anne: "The blue mug."

Fig. 4. Cross-modal reference disambiguation (1).

- (1) Anne: "Please give me the blue mug, R2D2."
- (2) R2D2 queries VisionSA for colour(mug9).
- (3) VisionSA: colour(mug9) = "red".
- (4) R2D2 queries VisionSA for colour(mug17).
- (5) VisionSA: colour(mug17) = "blue".
- (6) R2D2 takes mug17.

Fig. 5. Cross-modal reference disambiguation (2).

When an agent (the *speaker*) has access to information about a particular feature $?f$ of a binding proxy $?p$, then the agent can tell another one (the *hearer*) about it. In the MAPSIM run of Figure 4, R2D2 needs additional information about the feature *colour* of a proxy which was created by Anne using *language* as the modality (the expression "the mug"). R2D2 consequently plans for Anne to execute the action "*tell_val_feature Anne R2D2 colour langprox1*". This abstract request is concretised by the low-level clarification planner for *linguistic* clarification request into the form shown in line 2 of Figure 4.

Interestingly, the same operator is also used to clarify a visual referent in Figure 5. In this MAPSIM run, the linguistic information would be sufficient to identify the object. However, the robot has not yet determined the colours of all objects in the scene which, again, renders Anne's request in line 1 of Figure 5 ambiguous. This time the planner can determine that the information necessary for unambiguous binding must be provided by the robot's vision subarchitecture. The low-level planner for vision realises this request by actively running its colour detection algorithm. Since for the first object tested (mug9) the result cannot be bound to the linguistic proxy generated for Anne's expression "the blue mug" the continual planner replans, this time trying to bind to mug17 which indeed turns out to be blue. Since there is now a valid, unambiguous interpretation for Anne's command, the robot can now easily execute it (line 6).

Clarification is also essential for clarifying linguistic concepts that still need to be firmly associated with perceptions, i.e. in *sensory learning*. For example, the vision subarchitecture of the robot may need linguistic clarification of a colour that has been learnt but only partly recognised ("what is the colour of this object?"). Similar clarification requests arise during human-assisted *mapping* of buildings ("what room is this?"). We will explore these examples in future work.

V. CONCLUSIONS

This paper has discussed the problem of cross-modal situated clarification in HRI. We have characterised situated clarification as an information need which may arise in any sensory-motoric modality interpreting the situated context

```

(:action tell_val_feature
:agent (?speaker - agent)
:parameters
  (?hearer - agent
   ?f - feature_type
   ?p - binding_proxy)
:precondition (and
  (exists (?m - modality) (and
    (has_access_to_modality ?speaker ?m))
    (has_access_to_modality ?p ?m)
  )
  (K ?speaker (feature_val ?f ?p))
)
:effect (and
  (K ?hearer (feature_val ?f ?p))
))

```

Fig. 6. Planning operator used in the MAPSIM *clarification* domain

of the robot, or any deliberative modality referring to that context. For resolving such an information need we have proposed a continual planning approach where a robot creates a clarification plan involving multi-modal information-gathering actions, i.e. the planning process automatically determines which modalities are best used to clarify the ambiguities of the current situation. We have shown how this approach can be applied in a simulation environment where a virtual robot can plan and engage in clarification interactions both with other agents and its own sensory modalities.

We are currently applying our approach to the full robotic systems developed in the CoSy project. The conceptual framework and the continual planning algorithm used for clarification planning used within the CAST robotic architecture [12] is identical to the one described in this paper. However, noisy sensor data and interaction with real human users provide many additional possibilities for failed comprehension and, thus, clarification needs. In future work, we will evaluate our approach in these real-world applications. We expect that interactions with robots who can actively strive for clarification in case of failed understanding will be significantly more natural, robust and successful.

REFERENCES

- [1] 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.
- [2] G.T.M. Altmann and M. Steedman. Interaction with context during human sentence processing. *Cognition*, 30(3):191–238, 1988.
- [3] L.W. Barsalou. Perceptual symbol systems. *Behavioral & Brain Sciences*, 22:577–660, 1999.
- [4] P. Blackburn. Fine grained theories of time. In H. Wansing, editor, *Essays on Non-Classical Logic*, pages 1–36. World Scientific Publishing, 2001.
- [5] M. Brenner. Continual collaborative planning for mixed-initiative action and interaction (short paper). In *Proc. of the 7th Int. Conf. of Autonomous Agents and Multiagent Systems (AAMAS 2008)*, Estoril, Portugal, 2008.
- [6] M. Brenner and B. Nebel. Continual planning and acting in dynamic multiagent environments. In *Proc. Int. Symposium on Practical Cognitive Agents and Robots*, Perth, Australia, 2006.
- [7] H. Clark. *Using Language*. Cambridge University Press, Cambridge, UK, 1996.
- [8] J. Ginzburg. Clarifying utterances. In J. Hulstijn and A. Nijholt, editors, *Proceedings of the 2nd Workshop on the Formal Semantics and Pragmatics of Dialogue*, Enschede, The Netherlands, 1998.
- [9] J. Ginzburg and R. Cooper. Clarification, ellipsis, and the nature of contextual updates in dialogue. *Linguistics and Philosophy*, 27(3):297–365, 2004.
- [10] P. Gorniak and D. Roy. Situated language understanding as filtering perceived affordances. *Cognitive Science*, 31(2):197–231, 2007.
- [11] N. 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 *Proceedings of the Twenty-Second Conference on Artificial Intelligence (AAAI’07)*, 2007.
- [12] N. Hawes, M. Zillich, and J. Wyatt. BALT & CAST: Middleware for cognitive robotics. In *Proceedings of the 16th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2007)*, pages 998 – 1003, 2007.
- [13] 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.
- [14] G.J.M. Kruijff. *A Categorical-Modal Logical Architecture of Informativity: Dependency Grammar Logic & Information Structure*. PhD thesis, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic, April 2001.
- [15] G.J.M. Kruijff, J.D. Kelleher, and N. Hawes. Information fusion for visual reference resolution in dynamic situated dialogue. In E. André, L. Dybkjaer, W. Minker, H. Neumann, and M. Weber, editors, *Perception and Interactive Technologies (PIT 2006)*. Springer Verlag, 2006.
- [16] 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.
- [17] S. Larsson. *Issue-Based Dialogue Management*. Phd thesis, Department of Linguistics, Göteborg University, Göteborg, Sweden, 2002.
- [18] 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.
- [19] M. Poesio and D. Traum. Conversational actions and discourse situations. *Computational Intelligence*, 13(3):309–347, 1997.
- [20] M. Purver. *The Theory and Use of Clarification Requests in Dialogue*. PhD thesis, King’s College, University of London, August 2004.
- [21] 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.
- [22] K.J. Rodriguez and D. Schlangen. Form, intonation and function of clarification requests in German task oriented spoken dialogues. In *Proceedings of the 8th Workshop on the Semantics and Pragmatics of Dialogue (Catalog ’04)*, 2004.
- [23] D.K. Roy. Semiotic schemas: A framework for grounding language in action and perception. *Artificial Intelligence*, 167(1-2):170–205, 2005.
- [24] H. Sacks. *Lectures on conversation*. Blackwell, 1992.
- [25] E. Shegloff. Some sources of misunderstanding in talk-in-interaction. *Linguistics*, 25:201–218, 1987.
- [26] C. L. Sidner, C. Lee, C. Kidd, N. Lesh, and C. Rich. Explorations in engagement for humans and robots. *Artificial Intelligence*, 166(1-2):140–164, 2005.
- [27] L. Steels. Semiotic dynamics for embodied agents. *IEEE Intelligent Systems*, 21(3):32–38, 2006.
- [28] L. Steels. The symbol grounding problem has been solved. so what’s next? In M. De Vega, G. Glennberg, and Graesser G., editors, *Symbols, Embodiment and Meaning*. Oxford University Press, Oxford, UK, 2007.
- [29] M. Stone and C. Doran. Sentence planning as description using tree-adjointing grammar. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL’97)*, pages 198–205, 1997.
- [30] D.R. Traum. *A Computational Theory of Grounding in Natural Language Conversation*. PhD thesis, Computer Science Department, University of Rochester, December 1994.