Modelling Spatio-Temporal Comprehension in Situated Human-Robot Dialogue as Reasoning about Intentions and Plans

Geert-Jan M. Kruijff & Michael Brenner
EU FP6 "Cognitive Systems for Cognitive Assistants" (CoSy)
⟨gj@dfki.de, brenner@informatik.uni-freiburg.de⟩

Abstract

The article presents an approach to modeling spatio-temporal comprehension in situated dialogue. The approach combines linguistic reasoning with reasoning about intentions and plans during incremental interpretation of dialogue moves. The article explores how intention-directed planning can prime selectional attention in utterance comprehension by disambiguating linguistic analyses on the basis of plan availability, and by raising expectations what action(s) may be mentioned next. Also, planning can complement linguistic analyses with details on spatiotemporal-causal structure established in planning inferences. This makes such inferences available for future referencing in the discourse context.

1 Introduction

From a Gricean perspective, all dialogue is about conveying and comprehending *intentions*. Intentions, however, are meant to be realized and therefore are associated by both speaker and listener, explicitly or implicitly, with possible *plans* for their achievement. In this article, we investigate how reasoning about the corresponding intentions and plans can extend the linguistic interpretation of utterances. Particularly, we are interested in how such a cross-modal integration can be established *during* incremental utterance analysis, so that we can use preferred plan interpretations to guide the parsing process. The approach we propose is being integrated into an implemented system for collaborative human-robot interaction. While the analysis of an utterance incrementally unfolds, we want to establish a relation between the already constructed linguistic meaning, and the spatio-temporal planning context. Spatio-temporal context, as an aspect of situated awareness, has a demonstrated attentional effect on the ways *humans* process utterances in situated contexts (Altmann & Kamide 2004; Knoeferle & Crocker 2006). We want to explore whether cross-modal integration between incremental utterance processing and intention-directed action planning can confer similar attentional effects on situated dialogue processing for human-robot interaction. There are several potential advantages in this kind of integrated approach: Planning can help to disambiguate analyses, based on whether an interpretation corresponds to a construable (and likely) plan or not. Furthermore, planning can use possible plan continuations to raise expectations about what actions or objects may be mentioned next, as the analysis of the utterance unfolds. Finally, planning has the potential to contribute information to the linguistic representation.

Our approach to representing linguistic tense and aspect is based on an *event nucleus*, being a representation of the spatiotemporal-causal structure of a process (Moens & Steedman 1988). One option is to construct event nuclei solely during parsing. This would have the disadvantage that the necessary "world knowledge" about real world processes would have to be represented directly in the linguistic resources we use during parsing. The alternative we opt for here, is to use planning to provide such world knowledge. The cross-modal integration of this information lessens the burden on linguistic processing and representational resources, as inferences using world knowledge happen outside the language system. Furthermore, by integrating the relevant results of these inferences into the linguistic representation, we can make that information about the spatio-temporal context available within the discourse context. This means that we create the possibility for contextual references to e.g. processes whose presence is inferred, but which have not been explicitly mentioned in the discourse (bridging anaphora). Finally, because we are construing a representation of the linguistic meaning of an utterance in an *incremental* fashion, we can use information about the (in)availability of plans for a partial interpretation to create a preference order over available partial analyses, and so guide the parsing process.

In this article, we describe the conceptual basis for our ongoing work on an integration of incremental parsing and continual planning in the EU project CoSy. While both individual components have already been implemented (Baldridge & Kruijff 2002; Brenner & Nebel 2006), our current work is their integration on a robot. The article is structured as follows. §2 discusses a model for the representations of some basic linguistic phenomena, and §3 discusses the cross-modal integration of the linguistic representations with planning. We end with conclusions.

2 Linguistic Model

In this section we describe the basic *representational* model, covering the basic linguistic phenomena of *tense* and *as-*

pect (Steedman 2005). Tense and aspect provide a linguistic reflection of the temporal-causal structure for the events that are being talked about. We build these representations compositionally, and incrementally, during utterance comprehension. They provide the basis for interpretation against the planning context, and for the integration of planning inferences into the linguistic context. As formal framework we use ontologically sorted, relational structures (Kruijff 2001) in combination with Combinatory Categorial Grammar (Steedman 2000). The framework is implemented using OpenCCG[1] The relational nature of the representations we build make it possible to include *modal* information into account as well, to reflect intentions (Steedman 2005).

Representing linguistic meaning We represent linguistic meaning as a richly ontologically sorted, relational structure. We encode such structures using a description logic-based framework (Kruijff 2001). We build these structures during parsing, in a monotonic (compositional) way on the basis of the lexical meanings specified for the words in the utterance (Baldridge & Kruijff 2002).

(1) "I moved the block"

$$
\begin{aligned}
&@_{\{m1:action\}}(\mathbf{move} \\
&\quad \wedge\ \langle Tense\rangle\mathbf{past} \\
&\quad \wedge\ \langle Mood\rangle\mathbf{ind} \\
&\quad \wedge\ \langle Actor\rangle(i1 : person\ \wedge\ \mathbf{i}) \\
&\quad \wedge\ \langle Patient\rangle(b1 : thing\ \wedge\ \mathbf{block}))
\end{aligned}
$$

Example 1 gives an illustrative representation of the linguistic meaning for the utterance "I moved the block". The representation consists of a conjunction of *elementary predications*. We basically have three different types of elementary predications. $@_{\{m1:action\}}(\mathbf{move})$ specifies that we have a proposition $\mathbf{move}$ of sort *action*. We can identify the proposition by its handle $m1$. The embedded conjunct ... $\wedge\ \langle Tense\rangle\mathbf{past}$ specifies a *Tense* feature for $\mathbf{move}$, indicating it is an action that happened in the past. Finally, the conjunct ... $\wedge\ \langle Patient\rangle(b1 : thing\ \wedge\ \mathbf{block})$ specifies a relation between the $\mathbf{move}$-action, and an affected object (*Patient*) to which this action is applied, namely a $\mathbf{block}$ $b1$.

Basic model We adopt a Reichenbachian models of tense and aspect (Steedman 2005), separating the event time E from the reference time R. (There is also the speaker time S but we leave that out of the equation right now.) E provides the temporal extended context to which we relate R. For connecting R and E we can use the temporal relations from (Allen & Ferguson 1994). We start from the verbal proposition, characterize its aspectual category (which can be changed through e.g. adjuncts), and provide its basic tense, mood, and modality information. The proposition has an individual handle (unique identifier), which we can *temporally* interpret as R – modality and tense are predicated over this handle, and as such also temporally place whatever R is related to: $@_{\{handle:aspcat\}}(\mathbf{prop}\ \wedge \langle Tense\rangle\mathbf{tensevalue}\ \wedge \langle Mood\rangle\mathbf{moodvalue}\ \wedge\ \langle Modality\rangle\mathbf{modalityvalue})$. The mood and modality features provide the basis for representing intentions, and how they are conveyed (Steedman 2005). We model E as an event nucleus, conceived of as an ontologically sorted relational structure. The nucleus has its own handle and aspectual category (by default *accomplishment*). We model aspectual category as an ontological sort, and not as an attribute/value pair, to indicate the different implications for model-theoretic interpretation of the proposition a verb expresses. We follow (Blackburn et al. 1993) and (Steedman 2005) and consider the preparation and the consequent of the event nucleus to be modal relations, which we represent here as $\langle Prep\rangle$ and $\langle Cons\rangle$ respectively. These relations are between individually sorted handles to represent the preparatory activity, the event, and the consequent state. To connect the nucleus with its (temporal) components preparation, event and consequent state, we use the $\langle Contains\rangle$ temporal relation (Allen & Ferguson 1994). For presentational purposes, we use the abbreviations *acc* (accomplishment), *act* (activity), *ach* (achievement), and *cnsq* (consequent state).

Simple tense For simple tenses, R and E refer to the same 'point' in time. We model this by equating R with the aspect of an event nucleus that E corresponds to, through coindexation.[2] The temporal information of R then places R, E, and the event nucleus in the past, present, or future.

(2) a. I moved the block.

b.

$$
\begin{aligned}
&@_{\{m1:acc\}}(\mathbf{move} \\
&\quad \wedge\ \langle Tense\rangle\mathbf{past} \\
&\quad \wedge\ \langle Mood\rangle\mathbf{ind} \\
&\quad \wedge\ \langle Actor\rangle(i1 : person\ \wedge\ \mathbf{i}) \\
&\quad \wedge\ \langle Patient\rangle(b1 : thing\ \wedge\ \mathbf{block})) \\
&\wedge\ @_{\{m1:acc\}}(\mathbf{move} \\
&\quad \wedge\ \langle Contains\rangle(h1 : act\ \wedge\ \langle Prep\rangle(h2 : ach)) \\
&\quad \wedge\ \langle Contains\rangle(h2 : ach\ \wedge\ \langle Cons\rangle(h3 : cnsq)) \\
&\quad \wedge\ \langle Contains\rangle(h3 : cnsq))
\end{aligned}
$$

Example 2 illustrates how we can model simple tenses this way, here for the *simple past*. We both represent R and the event nucleus. Even when the nucleus plays little role in these example, we want to prepare the ground for §2 where we combine tense with aspectual variation. Because a $\mathbf{move}$ action has a duration over time, leading up to a particular culmination point (e.g. reaching a destination), we specify the proposition as an accomplishment. The *Tense* attribute has as value $\mathbf{past}$, to indicate past tense. By using $m1$ as handle for both the proposition and the event nucleus, we equate R and the event nucleus underlying E. The other simple tenses (present, future) are handled similarly.

Temporal sequencing of events In communication, we have several ways for indicating temporal sequence of events. Consider the variations in Example 3. Example 3(a) shows how we can use discourse structure to indicate temporal sequence implicitly - events follow up on one another in the order of the utterances in which they are mentioned. In Example 3(b) makes the temporal connection explicit through a connective "and then." That order of mention

[1] http://openccg.sf.net.

[2] Alternatively, R and E could be connected through a temporal relation; currently, an open issue is though which temporal relation from (Allen & Ferguson 1994) this ought to be.

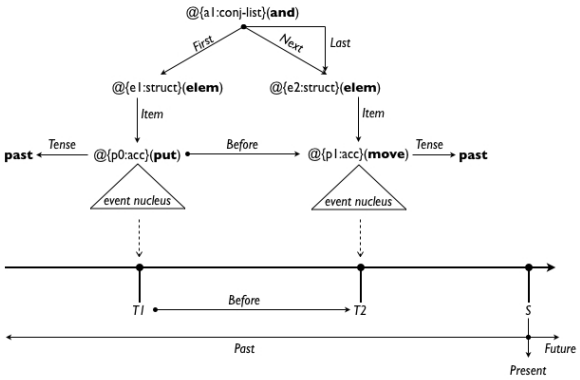

Figure 1: Timeline for *Before* sequence, past tense

need not mimic temporal order is shown in Example 3(c) – discourse structure is related to, but need not be identical with, the temporal order of events.

(3) a. Open the door. Now go into the corridor.
b. You must open the door first, and then you can go into the corridor.
c. You can go into the corridor, after you have opened the door.

To indicate temporal sequencing we use temporal relations between reference times – i.e. we temporally relate handles for the verbal propositions.

(4) I put red block near the green block, and then I moved the yellow block.

Example 4 illustrates temporal sequence in past tense. Figure 1 maps the temporal sequence of Example 4 onto a timeline, using the sequencing, past time references and the effect of *Before*. (S is the speaker time.)

Finally, sequencing may also indicate a *causal* relation, as Example 5 illustrates. In each of the examples, the event in the first conjunct provides a precondition for the event in the second conjunct.

(5) a. Take the pyramid and put it onto the block.
b. You must open the door first, and then you can go into the corridor.

Figure 2 shows the kind of relational structures we build.

Combining tense and aspect The aspectual opposition perfective/progressive can be combined with tense to indicate different types of *state* that currently obtains (Steedman 2005). The perfective leads to R being placed in a consequent state, derived from E which in and by itself would have realized an achievement. The progressive places R in a progressive state, derived from an activity. Tense information then places R in time. We capture the combination of tense and aspect at the level of linguistic meaning through *features* and *relations*. We represent aspect as a feature *Aspect* of the verbal proposition for R, and then relate R to a state in the event nucleus of the event. The event node is based on the verb being modified by the marker or auxiliary verb indicating tense/aspect.

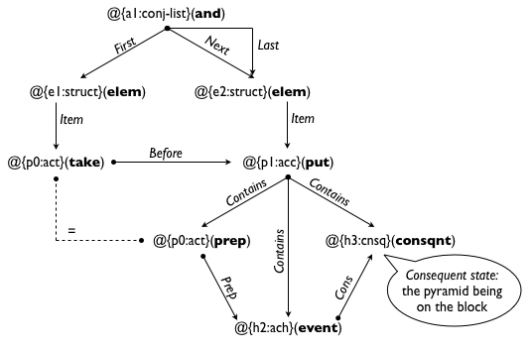

Figure 2: Timeline for *Before* temporal-causal sequence

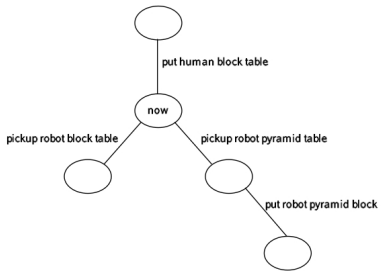

Figure 3: Example planning memory

3 Between dialogue and planning

In this section we present a discussion of how we can combine incremental utterance comprehension with reasoning about intentions and plans. We focus on how we can relate spatiotemporal-causal aspects of linguistically conveyed meaning, and the plan constructs these are intended to reflect. §3.1 presents the planning constructs we use to describe intentions and plans. We then discuss in §3.2 how we can associate linguistic meaning with these constructs. Associating requires we can set up a proper *temporal-causal context* within which we can subsequently interpret the actions conveyed by an utterance. It consists in building the planning constructs to reflect the (possibly complex) event nucleus structure in linguistic meaning. This presents an further interesting possibility to explore the interplay between planning and language comprehension: How we could use this reflection to let planning constructs *contribute to* the completion of event structures in the discourse model?

3.1 Planning constructs

Roughly speaking, AI Planning models the semantics of events as database updates: logical preconditions describe under which circumstances an event may occur, its effects (or postconditions) describe the changes to the current world state after its occurance. Plans represent logical and temporal relations between sets of such events. Usually, AI Planning systems are used to construct plans for achieving particular user-defined goals. In this work, however, we use a Planning system to verify whether dialogue moves by a human can be interpreted as *meaningful* intentions or planning goals. An intention is considered meaningful if the planning system can find a plan that achieves it and which is causally

linked to prior plans and plan executions.

The key ideas of our approach are independent of a particular planning formalism. For the reference implementation we currently work on, we use the multagent planning language MAPL, a formalism specifically tailored to continual multiagent planning (Brenner 2005; Brenner & Nebel 2006). It uses a slight variant of STRIPS-like propositional planning, the SAS^+ formalism (Bäckström & Nebel 1995) that allows multi-valued state variables instead of propositions. The main reason for doing this in the context of multiagent planning is that we want to represent *ignorance* about facts in addition to propositional truth and falsity. For dialogue interpretation and planning, the use of multi-valued state variables has an important side effect: it enables the representation of wh-questions (content questions) and answers to them as planning operators. A wh-question in our formalism thus corresponds to an intention to know the value of a particular state variable. In basic cases, we can directly match an interrogative *(where, who)* to a state variable *(location, agent)*; more complex questions correspond to quantified formulae specifying referential constraints for the unknown value of the state variable.

For this initial presentation of our approach to situated comprehension, we restrict our attention to *sequential* plans, i.e. plans defining a total execution order on the set of its actions. In order to describe more complex temporal event relations (e.g. "while") the plan model will be extended to allow for partially ordered sets of events and concurrency constraints, as described in (Brenner 2005).

An embodied cognitive agent is continually reassessing its opportunities for intention-driven behavior. If it is acting in an environment together with other agents, it has to take their actions into account, too. To enable reasoning about both future and past plans (by the system itself and by others) we have developed a new data structure, the *planning memory* (PM). In a PM, an agent can store the history of its state and plan changes as well as plans describing possible future behaviour (Figure 3). Planning memories are tries (retrieval trees) that can efficiently store large numbers of plans and states. Since plans resulting from continual updates often still share a common prefix, the trie data structure is particularly well suited to store such histories.
There is always one *now* node in the PM that corresponds to the current state of the world as known to the robot and that is used to evaluate present tense utterances against. The position of the *now* node is updated according to the actions executed or perceived by the system. PMs also store other kinds of information that will be important for our future work. In particular, a PM provides access to its *update history* and can also store the *reasons* why a plan was changed or abandoned for another one. In Figure 3 plan branches further to the right indicate more recent plan changes, i.e. the robot has decided not to pick up the block, but the pyramid next. The *now* node will be updated accordingly when the action is performed. Since information about previous plans, goals and knowledge is kept in the PM the robot can still answer introspective questions of the form "Why did you...?" and even "Why didn't you...?".

3.2 Associating planning constructs and (partial) linguistic meaning representations

During incremental utterance comprehension, we try to associate a –possibly partial– linguistic meaning representation with planning constructs. This association is done against the PM. Cross-modal binding over referents connects these action plans with a dialogue model talking about those plans. This makes it possible to use grounding on the PM to state *preferences* over possible interpretations, and to formulate *expectations* about how an utterance is likely to be continued. Preferences over interpretations are based on the (in)availability of a plan for an interpretation, and the likelihood of that plan, whereas expectations arise from possible plan continuations.

In incremental utterance comprehension, logical forms are mostly *partial*. They cover only that part of the utterance which has already been analyzed, starting from the first word. Planning is thus usually acting on incomplete information. This information may identify relative or absolute time points on which actions are said to happen, and provide information about the nature of the action-eventuality (i.e. state or event). Furthermore, besides tense and aspect information, a logical form may set up a modal context. The examples below illustrate these points.

(6) a. *Tense*
Put ...
b. *Tense, temporal modifier*
After you moved the ...
c. *Tense, aspect, temporal modifier*
Now that you have put the pyramid on the block,
d. *Tense, temporal modifier, modality*
Next, you will need to ...

The partial utterances in Example 6 illustrate "in words" the partial content planning may need to act on. Example 6(a) provides the simplest case. The utterance starts with the main verb. The partial logical forms express the verbal proposition with its tense and mood, its possible semantic argument structure, and the aspectual category. One possible logical form is shown in Example 7 below.

(7) a. Put ...
b. $@_{\{p0:acc\}}(\mathbf{put}$
$\wedge\ \langle Tense\rangle\mathbf{present}$
$\wedge\ \langle Mood\rangle\mathbf{imp}$
$\wedge\ \langle Actor\rangle(h1 : hearer\ \wedge\ \mathbf{robot})$
$\wedge\ \langle Patient\rangle(p1 : thing)$
$\wedge\ \langle Destination : WhereTo\rangle(o1 : region)$
$\wedge\ @_{\{p0:acc\}}(\mathbf{put}$
$\wedge\ \langle Contains\rangle(h1 : act\ \wedge\ \langle Prep\rangle(h2 : ach))$
$\wedge\ \langle Contains\rangle(h2 : ach\ \wedge\ \langle Cons\rangle(h3 : cnsq))$
$\wedge\ \langle Contains\rangle(h3 : cnsq))$

Example 6(b) illustrates a case where we have a temporal modifier, and the main verb in past tense. The temporal modifier "After" identifies a point in time in the future relative to the "move" event time (cf. the *reltimepoint* sort).

(8) a. After you moved the ...

b. $@_{\{m0:acc\}}(\mathbf{move}$
$\wedge\ \langle Tense\rangle\mathbf{past}$
$\wedge\ \langle Mood\rangle\mathbf{ind}$
$\wedge\ \langle Actor\rangle(h1 : hearer\ \wedge\ \mathbf{you})$
$\wedge\ \langle Patient\rangle(p1 : thing)$
$\wedge\ \langle Time : When\rangle(a1 : reltimepoint\ \wedge\ \mathbf{after})$
$\wedge\ @_{\{m0:acc\}}(\mathbf{move}$
$\wedge\ \langle Contains\rangle(h1 : act\ \wedge\ \langle Prep\rangle(h2 : ach))$
$\wedge\ \langle Contains\rangle(h2 : ach\ \wedge\ \langle Cons\rangle(h3 : cnsq))$
$\wedge\ \langle Contains\rangle(h3 : cnsq))$

Example 6(c) shows a more involved case. We have a main verb with tense and aspect information (perfective), a temporal modifier identifying an absolute point in time, and a construction which overall identifies a state.

(9) a. Now that you have put the pyramid on the block,
b. $@_{\{b0:cnsqstate\}}(\mathbf{be}$
$\wedge\ \langle Tense\rangle\mathbf{past}$
$\wedge\ \langle Mood\rangle\mathbf{ind}$
$\wedge\ \langle Restr\rangle(h1 : hearer\ \wedge\ \mathbf{you})$
$\wedge\ \langle Scope\rangle(p0 : acc\ \wedge\ \mathbf{put}$
$\wedge\ \langle Actor\rangle(h1 : hearer)$
$\wedge\ \langle Patient\rangle(p1 : thing\ \wedge\ \mathbf{pyramid})$
$\wedge\ \langle Dir : WhereTo\rangle(p1 : thing\ \wedge\ \mathbf{block}))$
$\wedge\ \langle Time : When\rangle(a1 : abstimepoint\ \wedge\ \mathbf{now})$
$\wedge\ @_{\{p0:acc\}}(\mathbf{put}$
$\wedge\ \langle Contains\rangle(h1 : act\ \wedge\ \langle Prep\rangle(h2 : ach))$
$\wedge\ \langle Contains\rangle(h2 : ach\ \wedge\ \langle Cons\rangle(b0 : cnsq))$
$\wedge\ \langle Contains\rangle(b0 : cnsq))$

Example 9 shows a logical form for Example 6(c). The logical form represents the *consequent state* as the root node, scoping over the event. The state is connected to the event nucleus of the main verb through co-indexation. The temporal modifier is attached to the root, to indicate the set the time point of the state to "now". The temporal relations between the components of the event nucleus thereby place the put-event in the past of "now."

Using information about tense, aspect, and aspectual category, we can set up plan constructs. We do not map utterances mentioning actions to plan operators directly, but interpret them as statements about future or past *intentions*. The planner then uses its own action ontology to reconstruct the corresponding plans. This is important because humans talk using completely different action descriptions, abstractions, a mixture of commands and goal descriptions, etc. E.g., when the parser produces a logical form for "Now put the pyramid on the block" the *planner-based intention recognition system* (PIRS) maps this first to a to a "raw" intention. This intention reflects the planning goal.

(10) a. "Now put the pyramid on the block"
b. Planning goal: *(exists (?o1 - pyramid ?o2 - block) (on ?o1 ?o2))*

The use of the planning ontology to interpret utterances and translate to to a goal formula is described in (Brenner *et al.* 2007). Additionally, the PIRS uses the logical form then to determine other pieces of information for the intention. The imperative mood of the utterance is interpreted as describing an intended future state, which the temporal modifier indicates is to be reached from the "now" state. Furthermore, the imperative makes the robot the *Actor* of the verb, and PIRS uses this to establish that the robot is to achieve the goal. This then yields the following plan.

(11) plan 1: ⟨ pickup robot pyramid table, put robot pyramid block ⟩

As Example 11 shows, the plan involves more than just a "put" action: the human has not fully specified what the robot should do, but expressed his intentions which the system "recoined" in terms of its own ontology. Even more importantly, the robot uses its own action *granularity* and generates a multi-action plan as an interpretation of the utterance. This is in contrast, for example, to Lochbaum's approach (Lochbaum 1998) where each utterance must correspond to a single method in the dialogue plan.

One particularly important aspect of this approach is that referential constraints used in the utterance need not be resolved by the linguistic analysis, but can be passed through to the planner. The goal formula in Example 10 does not refer to specific objects; the reference constraints will instead be resolved "lazily" only when the planner is called. Whether or not the planner is able to satisfy a goal formula provides given the *now* state of the PM provides a strong indication for the validity of an interpretation. For example, the command

(12) Put the big red block to the left of the green pyramid to the right of the small pyramid.

allows two possible interpretations, depending on whether the block is already to the left of the green pyramid or not. Fig. 4 shows the translation of the logical form for one of the interpretations into a planning goal. This translation is generated automatically by the system described in (Brenner *et al.* 2007). Depending on the *now* state of the PM the planner can disambiguate the command by either finding a plan resolving the constraints or not.

```
(exists (?o0 - block ?p0 ?ip0 - waypoint) (and
  (and (initially (size ?o0 big)) (initially
    (colour ?o0 red)) (exists
    (?o11 - pyramid ?ip11 - waypoint) (and (initially
      (colour ?o11 green)) (initially (pos ?o11 ?ip11))
     (left_of ?ip0 ?ip11) )))
  (exists (?o1 - pyramid ?p1 ?ip1 - waypoint) (and
    (initially (size ?o1 small)) (initially
     (pos ?o0 ?ip0)) (pos ?o0 ?p0) (initially
     (pos ?o1 ?ip1)) (pos ?o1 ?p1)
    (right_of ?p0 ?p1) ))))
```

Figure 4: Auto-generated planning goal for one of two possible interpretations of Example 12

The same is true when the human talks about past events. "Now that you have put the pyramid on the block" is mapped to the same intention expressed in terms of the planning ontology. However, the temporal relations between the components of the event nucleus described in the logical form

indicate to the PIRS that the intention must be satisfied in the "now" state, and that actions by the robot should have led up to it. Both assumptions can be validated by queries to the planning memory. The result of the queries can lead to changing the probability for this interpretation of the utterance, and the generation of appropriate responses. For example, if both queries confirm the interpretation, then the robot can produce a simple acknowledgment. On the other hand, if "now" does not satisfy the intention a response like "Oh, but the pyramid is no longer/not yet on the block", or if the plan in the planning memory leading up to "now" was not executed by the robot, a response like "No, someone else put the pyramid on the block" could be produced.

(Brenner *et al.* 2007) describe how action commands (but also other intentions in general) often also express referential constraints for objects in the current state. We can extend this approach to modelling several temporally and causally related events. The logical form describes the temporal relations between events. Each such event corresponds to an intention expressed as a planning goal. When constraints in the goal formula refer to the "current" state, this is no longer automatically interpreted as the "now" state of the planning memory, but as the state reached by achieving preceding intentions (according to the logical form).

(13) Take the pyramid and put it onto the block.

Consider Example 13. Here, a complete plan has been described by the human. Since, however, the system may have another ontology or plan on a different granularity level, the two intentions mentionend are matched by two subsequent plans (that happen to have only one action each):

(14) a. plan 2: $\langle$ pickup robot pyramid table$\rangle$
b. plan 3: $\langle$ put robot pyramid block $\rangle$

The goal state of plan 2 was used as the "now" state for finding plan 3. Since this interpretation does lead to two valid plans, that can be combined into another valid plan (plan 1 from above), this interpretation of the command is reasonable and thus is added to the planning memory. The robot thus not only knows what to do next, but also has a context to interpret future utterances against.

To recapitulate, we map the utterance to an intention, i.e. planning goal, and then try to construct the underlying plans. In the simplest case, we try to match "current" constraints with the *now* state, and check whether a plan exists for *then* states. To handle more complex cases, we enable the system to reason about several, temporally and causally related commands (or reports about actions performed by the human). These are then translated into the following kinds of queries to the planning memory: (1) Does a fact hold *now* (or in some other state)? (2) Is plan P executable *now* (or in some other state)? (3) Could *now* be the result of executing of plan P? (4) Which actions are enabled *now*? and, (5) Which actions will be enabled by executing plan P? Based on the results of the queries, we can state a preference order over the logical forms, based on how planning memory can support their interpretations, and form predictions about actions (on objects) that are likely to follow up on the plans supporting the interpretations.

3.3 Using planning for information completion in situated dialogue comprehension

There is a close relation between discourse structure and temporal structure. When we interpret a dialogue, we try to determine how the different utterances are related. Particularly if these utterances describe events, purely linguistic knowledge will not suffice to determine these relations – we also need to involve "world knowledge" to reason about the spatio-temporal and causal relations between the described events (Asher & Lascarides 2003). The continuous planning approach we have described above provides us with situated context within which we can perform these inferences, in a continuous and defeasible fashion. (Asher & Lascarides 2003) describe an approach for determining rhetorical relations between utterances, on the basis of defeasible inferences about temporal-causal structure between events. This is one fashion in which these inferences, can contribute to discourse interpretation. Another possibility is how the structures established in planning memory can help to *complement* the information we can derive linguistically. In a dialogue, relations between objects and events need not always be explicitly mentioned. Yet, these implicit relations may be essential for interpreting the dialogue as a coherent structure. *Bridging* provides a nice illustration. In bridging, we need to resolve an implication arising from a preceding utterance, to set up a context within which we can interpret the next utterance.

(15) a. I picked up a block. One side was yellow.
b. I went into the room. The door was already open.

Example 15 shows some simple examples of bridging. In Example 15(a) the mention of an identifiable "side" as belonging to the block is based on the simple *bridging inference* that a block has identifiable sides. To connect the two utterances in Example 15(b), we infer that to get into a room, there needs to be a gateway to access it. A door is a type of gateway, which can be either open or closed. In this context, we can make sense of the statement that there is a door which is in a state of being open.

We would like to explore how planning inferences can contribute information, to make those relations explicit which the planner *needs* to establish when creating planning constructs for interpreting utterance meaning. Planning does not resolve all such bridging inferences – Example 15(a) relies on *categorical* inference rather than temporal inference. On the other hand, the temporal-causal interpretation for Example 15(b) relies on making explicit that there is a known gateway, through which we can pass to get into the room. Given that planning deals with temporal-causal structure, we are interested in how planning could contribute to specifying event nuclei. Against the background of cross-modal binding, we could use this information in discourse-level comprehension to bind events in the discourse context. The upshot is that we do not need to model more than selectional preferences in linguistic meaning where it concerns the event nucleus. We can elide large parts of it in the linguistic specification, until planning completes information

based on already construed plans.

(16) a. When I picked up the toy longhorn, I took it by its head.

b. When I pushed the ball, it came to a halt after half a meter.

We can illustrate these ideas on the above examples. In Example 16(a), the when-clause introduces an event structure, for which planning already *predicts* that there has to be a grabbing-like event as preparatory activity. Because grabbing something is a precondition for picking it up, the grabbing event is already made explicit in the plan on which we interpret the when-clause. The taking-clause can then be interpreted in the context of the temporal-causal structure inferred for the when-clause, binding the taking into the event nucleus as preparatory activity. We can see this as a type of bridging relation at event-level. Example 16(b) provides another example of such event-level bridging. Here, the intermediate event of the ball *rolling* is not mentioned. That the ball needs to be rolling is an inference the planner needs to draw, otherwise it cannot connect the activity of pushing with the goal state of the ball coming to a halt. The result is a structure on which we can ground an event nucleus including pushing-rolling-halting. If the discourse model is completed with the information that there has to be a rolling event, with an identifiable referent grounded in a temporal-causal structure, subsequent utterances can coherently refer to that event, as illustrated in Example 17.

(17) When I pushed the ball, it came to a halt after half a meter. It didn't roll for very long.

4 Conclusions

We have presented the key concepts for integrating incremental parsing and planner-based utterance interpretation. Following (Steedman 2005) we discussed a model for representing tense, aspect, and aspectual categories and coercion at the level of linguistic meaning. These representations are built compositionally, and incrementally, during utterance comprehension. Communicative intentions detected by the linguistic analysis are mapped to planning goals. Using an appropriate planning language and the Planning Memory, a new data structure for storing and reasoning about past and future plans, a planner can be used to validate whether an an interpretation is supported by intentions and corresponding plans from the current spatiotemporal-causal context. The planner can state preferences over interpretations, and make predictions about what actions and objects are likely to be mentioned still as the utterance unfolds. Preferences and predictions are fed back to the language system, to influence selectional attention during utterance comprehension – guiding the parser, and possibly non-verbal aspects such as gaze (Staudte 2006). At the end of the article, we considered how the constructed plans can be used to complement information in the discourse model, to make temporal-causal relations between events explicit – and thus referable.

We have implemented the individual components of the model, and are currently integrating them into a system for collaborative human-robot interaction. This enables us to consider the effects of the *embodied situatedness* of the system on how dialogue and intentions should be understood – effects which are abstracted away from in approaches such as (Lochbaum 1998) or (Blaylock, Allen, & Ferguson 2003).

References

Allen, J., and Ferguson, G. 1994. Actions and events in interval temporal logic. *J. Logic and Computation* 4(5).

Altmann, G., and Kamide, Y. 2004. Now you see it, now you don't: Mediating the mapping between language and the visual world. In Henderson, J., and Ferreira, F., eds., *The Interface of Language, Vision, & Action: Eye Movements & The Visual World*.

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

Bäckström, C., and Nebel, B. 1995. Complexity results for SAS+ planning. *Computational Intelligence*.

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

Blackburn et al., P. 1993. Back and forth through time and events. In *Proceedings of the 9th Amsterdam Colloquium*.

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

Brenner, M., and Nebel, B. 2006. Continual planning and acting in dynamic multiagent environments. In *Proc. Int. Symposium on Practical Cognitive Agents and Robots*.

Brenner, M.; Hawes, N.; Kelleher, J.; and Wyatt, J. 2007. Mediating between qualitative and quantitative representations for task-oriented human-robot interaction. In *Proc. IJCAI 2007*.

Brenner, M. 2005. Planning for multiagent environments: From individual perceptions to coordinated execution. In *ICAPS Worksh. Multiagent Planning and Scheduling*.

Knoeferle, P., and Crocker, M. 2006. The coordinated interplay of scene, utterance, and world knowledge: evidence from eye tracking. *Cognitive Science*.

Kruijff, G.-J. 2001. *A Categorial-Modal Logical Architecture of Informativity: Dependency Grammar Logic & Information Structure*. Ph.D. Dissertation, Charles Univ., Prague, Czech Republic.

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

Moens, M., and Steedman, M. 1988. Temporal ontology and temporal reference. *Computational Linguistics*.

Staudte, M. 2006. Grounding robot gaze production in a cross-model category system. Master's thesis, Department of Computational Linguistics, Saarland University.

Steedman, M. 2000. *The Syntactic Process*. Cambridge, MA: The MIT Press.

Steedman, M. 2005. The productions of time. Draft 5.0.