Incremental, multi-level processing for comprehending situated dialogue in human-robot interaction

Geert-Jan M. Kruijff[1], Pierre Lison[1,2], Trevor Benjamin[1,2],
Henrik Jacobsson[1], and Nick Hawes[3]

[1]Language Technology Lab, German Research Center for Artificial Intelligence (DFKI GmbH), Saarbrücken Germany
[2]Dept. of Computational Linguistics, Saarland University, Saarbrücken Germany
[3]Dept. of Computer Science, University of Birmingham, Birmingham United Kingdom
gj@dfki.de

Abstract

The paper presents work in progress on an implemented model of situated dialogue processing. The underlying assumption is that to understand situated dialogue, communicated meaning needs to be related to situation(s) it refers to. The model couples incremental processing to a notion of bidirectional connectivity, inspired by how humans process visually situated language. Analyzing an utterance in a "word-by-word" fashion, a representation of possible utterance interpretations is gradually built up. In a top-down fashion, the model tries to ground these interpretations in situation awareness, through which they can prime what is focused on in a situation. In a bottom-up fashion, the (im)possibility to ground certain interpretations primes how the analysis of the utterance further unfolds. The paper discusses the implementation of the model in a distributed, cognitive architecture for human-robot interaction, and presents an evaluation on a test suite. The evaluation shows (and quantifies) the effects linguistic interpretation have on priming incremental utterance processing, and discusses how such evaluation can be extended to include situation-relative interpretation.

Introduction

The environments in which we deploy our robots provide them with rich, perceptual experiences. And language provides a combinatoric system that enables us to talk about those environments in a rich variety of ways. The problem is of course then how we can figure out, what an utterance really is supposed to mean in a given context. From psycholinguistics and cognitive science we know that humans use context information to do this. Humans do not wait with processing an utterance until they have heard the end of it. On the contrary. While processing an utterance, they link unfolding interpretations to the dialogue- and situated-context to filter out unlikely interpretations. They use their understanding of the situational context to disambiguate and refine how they comprehend an utterance, and at the same time use what is being talked about to selectively refine their situation awareness. Using context, they pick those meanings out of the myriad of possible meanings, to focus on just those that seem most appropriate in the given context (Altmann and Steedman, 1988; Altmann and Kamide, 2004; Knoeferle and Crocker, 2006).

In this paper, we discuss an implemented model that enables a robot to understand situated dialogue in a similar way. The model relies on explicitly grounding dialogue in the situated context. The main idea is to use an *incremental* model for dialogue analysis, and connect step-by-step the unfolding possible *linguistic* interpretations of an utterance to information about the visually situated context. From this interconnection we can then derive what visual objects are being talked about, and whether the way these referents are referred to, and put into relation, can be grounded in the situated context. We use insights from psycholinguistics in postulating what factors in the visually situated context might play a role (Altmann and Steedman, 1988; Altmann and Kamide, 2004; Knoeferle and Crocker, 2006), and how they affect priming of utterance processing.

Our approach is related to other recent work on incremental language processing for dialogue systems (Allen et al., 1996; Mori et al., 2001; Rosé et al., 2002), and for human-robot interaction (Brick and Scheutz, 2007) (B&S). Like B&S we analyze an utterance for its meaning, not just for syntactic structure (Allen et al., 1996; Mori et al., 2001; Rosé et al., 2002). We make several advances, though. The model incrementally analyzes utterance meaning not only at the grammatical level, but also at dialogue level. B&S only consider the former (parsing). By interpreting an utterance also relative to the dialogue context, the model allows different levels of linguistic description to constrain possible interpretations (Altmann and Steedman, 1988; Stone and Doran, 1997). This presents several advantages. We can (linguistically) resolve contextual references such as deictic pronouns and anaphora. This resolution relates references made in the current utterance to ones made already earlier in the dialogue – i.e., ultimately to visual objects that have already been identified. Furthermore, we can use the dialogue "move" of the utterance to determine what *needs* to be bound. For example, in a greeting like "Hi there!" the model does not need to try and bind "there" to a location.

A further advance is that we adopt a "packed" representation of the linguistic interpretations (Oepen and Carroll, 2000; Carroll and Oepen, 2005) to efficiently handle alterna-

tive (i.e. ambiguous) meanings. Any grammar of a reasonable size will generate multiple syntactic-semantic analyses of an utterance. This can easily result in hundreds of alternative analyses that would need to be checked. A packed representation represents all the information shared across alternative analyses only *once*, which greatly reduces the amount of linguistic content we need to ground. These packed representations are subsequently related to information about the situation and ongoing tasks (Allen et al., 2001; DeVault and Stone, 2003; Gorniak and Roy, 2007). This essentially comes down to trying to resolve how a meaning refers to the current context (Stone and Doran, 1997; Brick and Scheutz, 2007). Intuitively, if a meaning presents an unresolvable reference, or an unresolvable assertion about spatial organization, then it can be discarded.

An overview of the paper is as follows. We start by providing a brief overview of insights of how humans process situated utterances, and position our approach to other work in AI and HRI. We then present our approach. We discuss its implementation using the CoSy Architecture Schema toolkit (Hawes et al., 2007a; Hawes et al., 2007b). Using a test suite with a variety of visual scenes, we evaluate our approach in a systematic way on different types of potential linguistic ambiguity. We measure the effects of linguistic understanding on priming utterance processing. The paper closes with conclusions.

Background

The combinatorial nature of language provides us with 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 in various branches of psycholinguistics and cognitive neuroscience have investigated what information listeners use when comprehending spoken utterances. An important observation across these studies is that interpretation *in context* plays a crucial role in the comprehension of utterance as it unfolds. Following (Knoeferle and Crocker, 2006) we can identify two core dimensions of the interaction between linguistic context and situated context. One is the *temporal dimension*. Attentional processes in situational perception appear to be closely time-locked with utterance comprehension. This can be witnessed by for example eye movements. The second is the *information dimension*. This indicates that listeners not only use linguistic information during utterance comprehension, but also scene understanding and "world knowledge." Below we discuss aspects of these dimensions in more detail.

Multi-level integration in language processing

Until the early 1990s, the dominant model of language comprehension was that of a modular, stage-like process; see for example (Fodor, 1983). On this model, a language user would sequentially construct each level of linguistic comprehension – from auditory recognition all the way to pragmatic, discourse-level interpretation. As (Van Berkum et al., 2003) observe, two hypotheses followed from this view. Firstly, people first construct a local, context-*independent* representation of the communicated meaning, before this meaning is interpreted against the preceding discourse context. Secondly, and related, is the hypothesis that discourse context-related processing only enters the process of language comprehension at a relatively late stage.

Opposing these hypotheses is the view that language comprehension is an incremental process, in which each level of linguistic analysis is performed in parallel. Every new word is immediately related to representations of the preceding input, across several levels – with the possibility for using the interpretation of a word at one level to co-constrain its interpretation at other levels. A natural prediction that follows from this view is that interpretation against dialogue context can in principle affect utterance comprehension *as the utterance is incrementally analyzed*, assisting in restricting the potential for grammatical forms of ambiguity. (Crain and Steedman, 1985; Altmann and Steedman, 1988) phrased this as a *principle of parsimony*: those grammatical analyses are selected that for their reference resolution impose the least presuppositional requirements on a dialogue context.

Since then, various studies have investigated further possible effects of dialogue context during utterance comprehension. Methodologically, psycholinguistic studies have primarily investigated the effects of dialogue context by measuring *saccadic eye movements* in a visual scene, based on the hypothesis that eye movements can be used as indications of underlying cognitive processes (Tanenhaus et al., 1994; Liversedge and Findlay, 2000). Alternatively, cognitive neuroscience-based studies use event-related brain potentials (ERPs) to measure the nature and time course of the effects of discourse context on human sentence comprehension (Van Berkum, 2004).

Both lines of study have found that lexical, semantic and discourse-level integrative effects occur in a closely time-locked fashion, starting already at the phoneme or sub-word level; (Allopenna et al., 1998), and (van Berkum et al., 1999b; Van Berkum et al., 2003; Van Petten et al., 1999). Particularly, a range of discourse-level integrative effects have observed. Referential binding has been shown to play a role in the constraining various types of local syntactic ambiguities, like garden path-constructions (Crain and Steedman, 1985; Altmann and Steedman, 1988; Altmann, 1988), and relative clauses (Spivey et al., 1993; Spivey and Tanenhaus, 1998); (van Berkum et al., 1999a; van Berkum et al., 1999b; Van Berkum et al., 2003). These effects primarily concern a *disambiguation* of already built structures. Integrating semantic and discourse-level information during utterance comprehension also has important *anticipatory* effects. (Tanenhaus et al., 2000; Dahan and Tanenhaus, 2004); (Van Berkum et al., 2005) observe how contextual

information influences what lexical meanings can be anticipated, priming phonological understanding and lexical access. (Contextual information can even override disprefered lexical meaning (Nieuwland and Van Berkum, 2006).)

Anticipatory effects indicate that utterance comprehension is thus not only an incremental process of constructing and then disambiguating. Anticipation enables context-dependent phonological recognition, lexical retrieval, and syntactic construction - without there being a need to generate and test all combinatory possible constructions. Incrementality and anticipation based on multi-level integration appears to give rise to a process in which comprehension arises through a convergence based on constraining and co-activation. Discourse context and the interpretative contexts which are delineated during utterance comprehension converge to become functionally identical (Van Berkum et al., 2003). As a result, ambiguity need not even arise, or is at least being much more limited a priori through context.

An important issue in all of the above remains of course the degree to which integrative effects indeed should commit to a certain understanding. Garden path sentences are a good example. They show that overcommitment risks the need for re-interpretation – an issue for *cognitive control* (Botvinick et al., 2001; Hommel et al., 2002; Novick et al., 2005).

Language processing and situational experience

We already noted before that human language processing integrates *linguistic* and *non-linguistic* information. Below we discuss studies which investigate how categorical and contextual information from situated experience can effect utterance comprehension. These studies use eye-trackers to monitor where people look at in a scene, and when.

(Altmann and Kamide, 1999) present a study revealing that listeners focus their attention on objects before these objects are referred to in the utterance. Figure 1 illustrates the setup of the study. When someone hears "The cat chases the mouse", her gaze already moves to the mouse in the scene before she has actually heard that word;

Figure 1: Mouse, cheese, cat

similarly for "The mouse eats the cheese." Knowing that cats typically chase mice (not cheese), and that the argument structure of *chase* reflects this, the listener *expects* that the next object to be mentioned will be the mouse, and directs gaze to that object. We thus see an anticipatory effect arising from the online integration of lexico-semantic information (verbal argument structure), situational context (the present objects, and the intended action), and categorical knowledge (prototypical object-action relations).

Not only world knowledge can influence online utterance comprehension, also scene understanding can. For example, consider the situation in Figure 2. (Tanenhaus et al., 1994) show that, once the listener has heard "Put the apple on the towel ..." she faces the ambiguity of whether to put the (lone) apple onto the (empty) towel, or to take the apple that is on the towel and
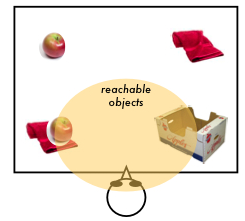

Figure 2: Put, apple, towel, box

put it somewhere else. The ambiguity is revealed as visual search in the scene. Only once she has heard the continuation "... into the box" this ambiguity can be resolved. Interestingly, in (Tanenhaus et al., 1994) the listener cannot directly manipulate the objects. If this is possible (cf. Figure 2), (Chambers et al., 2004) show that also reachability plays a role in comprehending the utterance. Because only one apple is reachable, this is taken as the preferred referent, and as such receives the attention. This underlines the effect *physical embodiment* may have on language comprehension.

Scene understanding also concerns the *temporal projection* towards possible future events (Endsley, 2000). (Altmann and Kamide, 2004; Kamide et al., 2003) show how such projection can also affect utterance comprehension. These studies used a scene with a table, and besides it a glass and a bottle of wine, as illustrated in Figure 3 (left). Investigated was where listeners look when they hear "The woman will put the glass on the table. Then, she will pick up the wine, and pour it carefully into the glass." It turns out that after hearing the "pouring" phrase, listeners look at the table, not the glass. Listeners thus explicitly project the result of the picking action into the scene, imagining the scene Figure 3 (right).

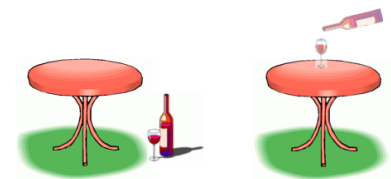These studies reveal that the interaction between vision and language is not *direct*, but *mediated* (Altmann and Kamide, 2004). Categorical understanding

Figure 3: Pouring, wine, glass, table

plays an important role in the sensorimotoric grounding of language. This is further underlined by studies like (Glenberg and Kaschak, 2002; De Vega et al., 2004), following up on the idea of category systems as mediating between perceptual modalities and language (Glenberg, 1997; Barsalou, 1999). These studies show how categorical understanding gives rise to expectations based on affordances, influencing comprehension of spatial or temporal

aspects of action verbs.

In conversational dialogue (Hadelich and Crocker, 2006; Pickering and Garrod, 2004) gaze has been shown to be automatically aligned in simple collaborative interaction. The time intervals between eye-fixations during production and comprehension of a referring expression are shorter than in monologue. This is further evidence for the relevance of visual common ground of interlocutors and how that accelerates the activation of jointly relevant concepts.

Situated language processing in AI/HRI

Studies on how humans process visually situated dialogue show an important aspect of "grounding" is based on how we can resolve a visual referent for an object reference. In establishing referents, listeners use visual and spatio-temporal properties of objects, and combine these properties with various forms of salience.

Several approaches have been proposed for visual referent resolution in human-robot interaction, in relation to language processing. Gorniak & Roy (Gorniak and Roy, 2004; Gorniak and Roy, 2005) present an approach in which utterance meaning is probabilistically mapped to visual and spatial aspects of objects in the current scene. Recently, they have extended their approach to include action-affordances (Gorniak and Roy, 2007). Their focus has primarily been on the grounding aspect, though. Although they use an incremental approach to constructing utterance meaning, grounding meanings in the social and physical context as they are construed, the (im)possibility to ground alternative meanings does not feed back into the incremental process to prune inviable analyses. This is where they differ from e.g. Scheutz et al (Scheutz et al., 2004; Brick and Scheutz, 2007). Scheutz et al present a model for incremental utterance processing in which the analyses are pruned if it is impossible to find visual referents for them.

Our approach to incremental language analysis is closely related to that of Scheutz et al. We incrementally build up a representation of utterance meanings, in parallel to syntactic analyses (Steedman, 2000). In this we (jointly) differ from other approaches such as (Allen et al., 1996; Mori et al., 2001; Rosé et al., 2002), who only build syntactic analyses. We advance on Scheutz et al in several ways, though. We analyze utterance meaning incrementally not only at the level of grammar, but also relative to the structure of the dialogue context. This allows different levels of linguistic description to constrain possible interpretations (Stone and Doran, 1997). Furthermore, we do not deal with individual analyses, but with a "packed" representation (Oepen and Carroll, 2000; Carroll and Oepen, 2005) to handle linguistic ambiguity. Ambiguity is inherent in natural language. Often, parts of an utterance may be understood in different ways. Packing provides an efficient way to represent ambiguity. Parts shared across different analyses are represented only once, and ambiguities are reflected by different ways

in which such parts can be connected. These packed representations are subsequently related to information about the (possibly dynamic) situation (Kruijff et al., 2006) and ongoing tasks (Allen et al., 2001; DeVault and Stone, 2003; Brenner et al., 2007; Gorniak and Roy, 2007). Should a possible meaning turn out to present an unresolvable reference, we discard it from the set of analyses the parser maintains.

Approach

Our approach has been implemented as part of an artificial cognitive architecture, built using the CoSy Architecture Schema Toolkit (CAST) (Hawes et al., 2007a; Hawes et al., 2007b). For the purpose of this paper, we focus on an architecture consisting of subsystems for visual and spatial processing of the situation, for interconnecting ("grounding") content across subsystems, and for dialogue processing.

In CAST, we conceive of a cognitive architecture as a distributed collection of subsystems for information processing (Hawes et al., 2007a; Hawes et al., 2007b). Each subsystem consists of one or more processes, and a working memory. The processes can access sensors, effectors, and the working memory to share information within the subsystem. We divide processes into unmanaged, data-driven and managed, goal-driven processes. A data-driven process writes information onto the working memory in an "unmanaged" fashion, typically whenever that information becomes available (e.g. from a sensor). A goal-driven process performs a specific type of interpretation of information available in working memory. This is a "managed" process controlled by the subarchitecture's task manager. The task manager decides when a goal process may, or may not, carry out its processing. This enables the subarchitecture to synchronize various forms of information processing.
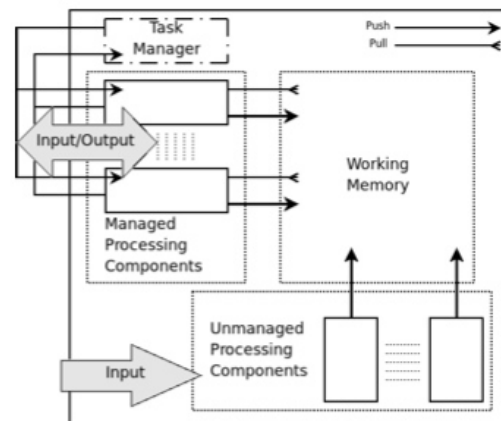

Figure 4: Abstract organization of a subarchitecture

Subsystems can also share information with other subsystems. We do this by monitoring a working memory of another subsystem, and reading/writing content to it.

Typically, a subsystem uses its own representation formats to deal most efficiently with the data it needs to handle. For example, the visual working memory contains regions of interest generated by a segmentor and proto-objects generated by interpreting these regions, whereas the dialogue subsystem contains logical forms generated from parsing utterances, and spatial reasoning maintains abstractions of physical objects with qualitative spatial relationships between them.

In our overall system, we have subsystems for vision, dialogue processing, manipulation, spatial reasoning (local scenes as well as multi-level maps), planning, coordination, and binding (used for symbol grounding). Several instantiations of this system have been described elsewhere (Hawes et al., 2007a; Brenner et al., 2007; Kruijff et al., 2007). Together, these subsystems create a system that can learn and communicate about objects and spatial locations with a user, and perform manipulation and navigation tasks.
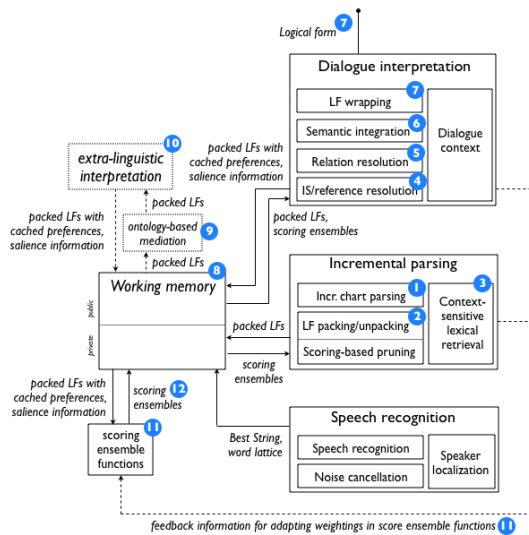

Figure 5: Dialogue processing (comprehension part)

Figure 5 illustrates the comprehension side of our dialogue processing subsystem.[1] (The numbers in the text refer to the round, blue labels in the figure.)

For speech recognition we use Nuance v8.5, to which the subsystem connects over a SIP connection. This enables us to use any number of microphones to "speak" to the robot – enabling both face-to-face and remote dialogue. Using an 8-microphone array on the robot we can do basic forms of noise cancellation and speaker localization. Speech recognition stores a recognition result on working memory in the form of a best string. Once this information becomes available, an incremental parsing process is triggered.

We have factorized (incremental) parsing into several, interconnected functions: the incremental parsing process itself (1), packing/unpacking and pruning of incrementally construed analyses of utterance meaning (2), and context-sensitive lexical retrieval (3). Parsing is based on a bottom-up Early chart parser (Sikkel, 1999) built for incrementally parsing Combinatory Categorial Grammar (Steedman, 2000; Baldridge and Kruijff, 2003). Its implementation relies on basic functionality provided by OpenCCG[2].

Incremental chart parsing creates partial, and integrated analyses for a string in a left-to-right fashion. As each word in the utterance is being scanned, the parser retrieves from the lexicon (3) a set of lexical entries. A lexical entry specifies for a word all its possible syntactic and semantic uses. During parsing, this information is used to integrate the word into possible analyses. By factorizing out lexical retrieval we have made it possible to use information about the situated- and task-context to restrict what lexical meanings are retrieved ("activated") for a word. After each word, the parser's chart maintains one or more possible analyses in parallel. These analyses represent the syntactic and semantic structure built for the utterance so far, and indicate possible ways in which these analyses can be continued by means of open arguments.
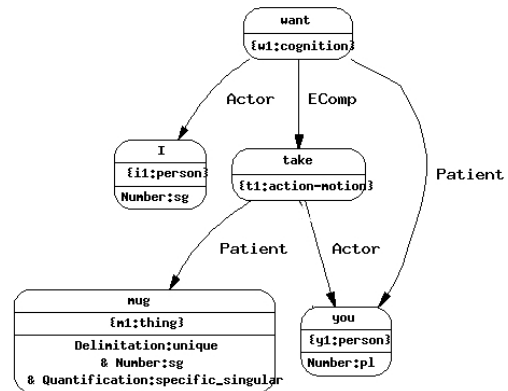

Figure 6: Logical form "I want you to take the mug"

Semantic structure is represented as an ontologically richly sorted, relational structure – a logical form (Baldridge and Kruijff, 2002). Figure 6 gives an example of a logical form (system output). Each node has a unique identifier with an associated ontological sort (e.g. $t1$ of sort $action-motion$), and a proposition (e.g. **take**). Nodes are connected through named relations. These indicate how the content of a single node contributes to the meaning of the whole expression. For example, "you" ($y1$) both indicates the one whom something is wanted of (*Patient*-relation from

---

[1]Most of the indicated processes have been implemented at the time of writing. Under construction are still *semantic integration* and *IS* i.e. information structure resolution.

*w*1), and the one who is to perform the taking action (*Actor*-relation from *t*1). Nodes carry additional features, e.g. *i*1 identifies a singular person.

After each step in incremental parsing, the current set of logical forms is packed to create a more efficient representation for computing with logical forms (Oepen and Carroll, 2000; Carroll and Oepen, 2005). Figure 7 illustrates the development of the packed packed representation for "take the mug". At the first step ("take"), 6 logical forms are packed together, showing we essentially have two alternative interpretations: "take" as an action, and as part of the expression "take a look." The second step ("take the") makes it clear we only need to look at the action-interpretation. The possible meanings for the determiner is expressed at the node for the Patient. At this point we have an *overspecified* meaning. Although the delimination is unique, we cannot tell at this point whether we are dealing with a singular object, or a non-singular (i.e. plural) object – all we know it has to be one or the other. This becomes determined in the third step ("take the mug").
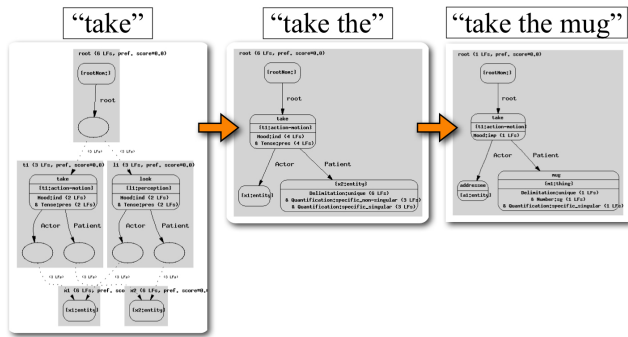

Figure 7: Example packed logical forms

Once the parser has created a packed representation, this is provided to the working memory. At this point, several processes for dialogue interpretation further interpret the representation, by providing discourse referents for the objects and events in the logical forms (4) and trying to connect the utterance to the preceding dialogue context in terms of rhetorical relations and dialogue moves (Asher and Lascarides, 2003). The resulting interpretations are related to the packed logical forms through "caches". A cache is a representation in which content is associated with other content, maintaining a mapping between unique keys in the two content representations. By using caches on top of the packed logical forms, we achieve a scalable approach for multi-level dialogue interpretation.

The packed logical forms, together with any dialogue-level interpretation of the content, is then provided to subsystems for extra-linguistic interpretation (8–10) (see below). The result of such interpretation is one or more preference orders over the interpretations representation by the packed logical forms. Technically, a scoring function is a partial order over substructures in packed logical forms. We can define ensembles over these functions to integrate their preferences, as e.g. suggested in (Kelleher, 2005) for salience functions. Before each next parsing step, packed logical forms are then pruned based on scoring ensembles, and the parse chart is updated.

In the architecture discussed here we rely for visual referent resolution on a grounding process called *binding*. The basic idea is illustrated in Figure 8. Each subsystem can have a binding monitor, which is a process that monitors the subsystem'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 an *amodal* format used in the binding subsystem. This is based on the idea of ontology-mediated information fusion, cf. (Kruijff et al., 2006).

The resulting representation is then written to the working memory in the binding subsystem. There it acts as a *proxy* – namely, as a proxy for content in the originating subsystem. The binding subsystem now applies strategies to combine proxies with similar content, but coming from different subsystems. Proxies can then be combined form unions. The power of the binding mechanism is that we can use a mixture of early- and late-fusion, and represent content at any level of abstraction.
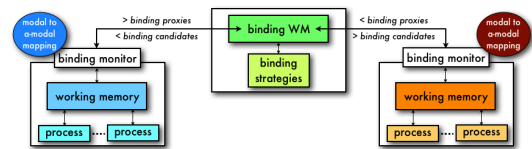

Figure 8: Cross-modal interconnectivity

Particularly, proxies from an individual subsystem can form relational structures. We thus can represent "the blue mug" as a single proxy, as well as "the blue mug next to the red box" as a relational structure connecting two proxies. Like individual proxies, the binder will try to connect relational structures – and either succeeding in doing so, e.g. if there is a blue mug next to the red box, or failing. This is crucial for situated dialogue processing (cf. also (Scheutz et al., 2004; Brick and Scheutz, 2007)).

Once we have a packed representation of logical forms, alternative relational structures are presented as proxies to the binding subsystem. By monitoring which relational structures can be bound into unions, and which ones cannot, we can prune the set of logical forms we maintain for the next step(s) in incremental parsing. We thus handle examples such as those discussed in (Brick and Scheutz, 2007) through an interaction between our binding subsystem, and the subsystem for dialogue processing.

Evaluation

Below we present preliminary results of an evaluation of the system. At the time of writing, we can only present statistical results evaluating the linguistic aspects of our processing model – not for the impact cross-modal binding has on linguistic processing. We do describe below how we will be able to statistically evaluate the impact of binding.

Design & measures

We have designed a set of eleven visual scenes, in which we can systematically vary the potential ambiguity of a visual object relative to specific types of referring expressions. Figure 9 gives an example of such a scene. Assuming we are looking at the scene from the robot's viewpoint, expressions such as "the blue thing" or "the blue ball" uniquely refer to the blue ball (with identifier $b2$). If we furthermore take e.g. visual and topokinetic salience into account, the referring expression "the mug" in "take the mug" has a strong preference for the red mug ($m1$) as being the visual referent (the white mug ($m2$) being less visually salient, and unreachable).
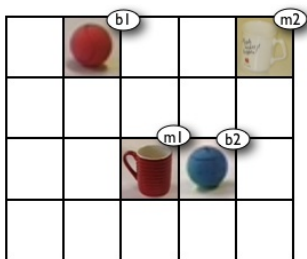

Figure 9: Sample visual scene

For these scenes, we have formulated a total of 58 utterances. These utterances express either commands ("put the mug to the left of the ball") or assertions ("the mug is red"). The utterances vary in length, with a distribution as given in Table 1. The (weighted) average length of the utterances in the evaluation is 6.07 words.

The utterances include referring expressions, which may be ambiguous relative to the scene for which they have been formulated. This enables us to investigate the interplay between different forms of ambiguity. First, we want to explore to what degree we can resolve purely linguistic ambiguity (notably, syntactic PP-attachment ambiguities) against non-ambiguous situations. Second, we want to evaluate to what degree ambiguity in situation awareness can be resolved through non-ambiguous linguistic meaning – or, if both would be ambiguous, to what degree we can still reduce the ambiguity. By systematically varying the ambiguity in the scenes, and in the structure of the utterances, we can properly evaluate these factors.

In the experiment, we have used two incremental parsers. One is the incremental parser which uses grammatical

length	# utterances
16	1
14	2
13	2
12	4
11	7
10	4
9	3
8	5
7	4
6	5
5	11
4	10

Table 1: Distribution of #utterances over lengths

knowledge to prune analyses during parsing ("pruning"). The other parser does not do any pruning, and functions as baseline ("baseline"). Below we show results of the pruning parser relative to the baseline performance.

Results

We present here results that measure the improvements the pruning parser makes over the baseline in terms of number of final analyses, the size of the resulting packed logical form, and time to obtain all complete analyses. The first two aspects measure memory use. Memory use is a factor that has an important impact on situated grounding of language. The fewer analyses, and the smaller the packed logical form, the less *varying* (or ambiguous) information we need to try and bind to information in other modalities.
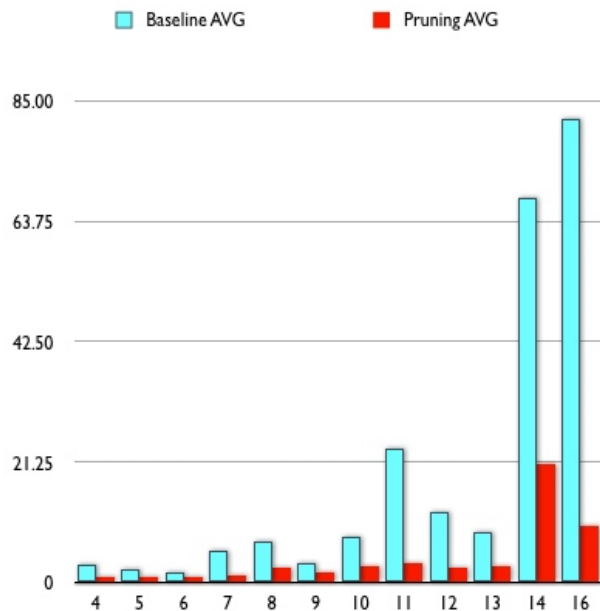

Figure 10: Number of final analyses (baseline, pruning)

Figure 10 shows a bar chart of the number of final analy-

ses produced by the baseline parser (light-blue, left) and the pruning parser (red, right). Using weighted averages, we get a 65.92% improvement of the pruning parser over the baseline. This improvement is statistically significant (one-way analysis of variance, F value = 27.036, Pr > 0.001).
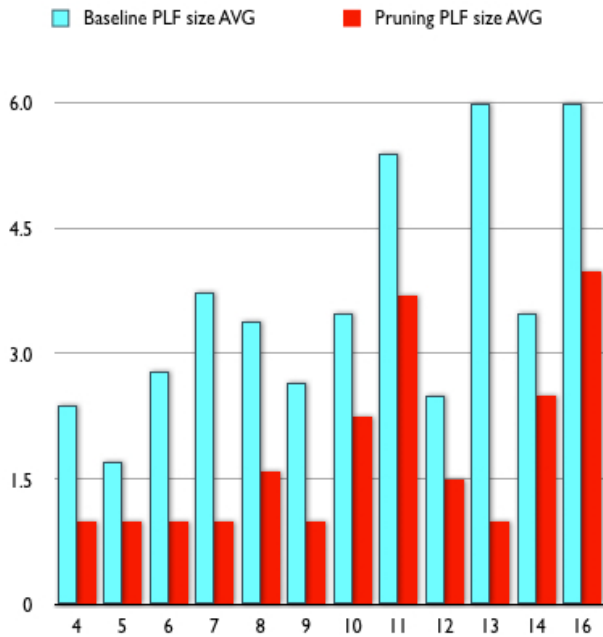

Figure 11: Number of final packed LF size (baseline, pruning)

If we look at the variation in size of the packed logical forms, we see a similar improvement. Figure 11 plots the sizes of the resulting packed logical forms against the utterance length, for the two parsers. This shows a 49.87% improvement of the pruning parser over the baseline (weighted average). Again, this result is statistically significant (one-way analysis of variance, F value=6.5283, Pr >0.01).

Figure 12 gives the results for time to parse completion, for the pruning parser and the baseline. On a weighted average, the pruning parser presents a 6.04% over the baseline (statistically significant, F value = 115.40, Pr > 0.001).

Discussion

The results show improvements of the pruning parser over the baseline in terms of memory use, and in time to completion. We have obtained these improvements on a data set of 58 utterances of varying complexity – not on isolated examples – and shown them to be statistically significant.

These results are in and by themselves not surprising – if a parser does pruning, it should do better than a baseline which does not. What is more interesting in the light of situated dialogue processing is that, even when we do use grammatical knowledge to select analyses, this may still not be enough to reduce the final number of analyses to 1. If that
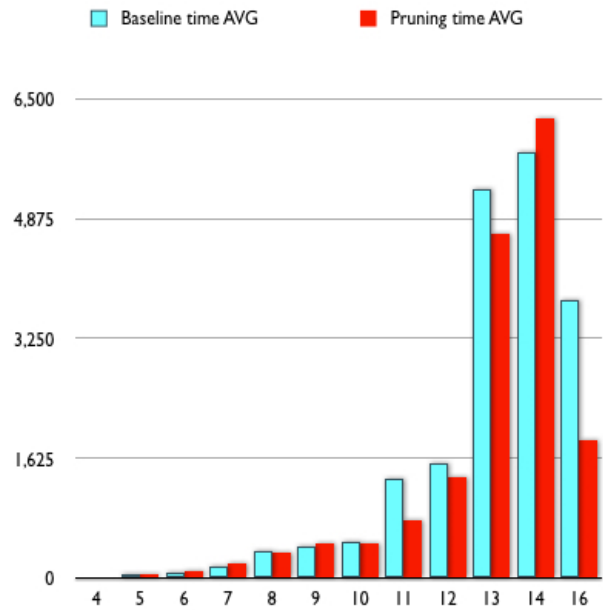

Figure 12: Time (ms) to parse completion (baseline, pruning)

were the case, then there would be no need to use grounding in the situation. On the data set we have used, we have a (weighted) average of 2.71 final analyses for the pruning parser (against a weighted average of 10.77 for the baseline).

Our next step is to evaluate our system, including the visual scenes on which the utterances have been formulated. The system enables us to prune analyses based on what content in a packed logical form can (or cannot) be grounded in situation awareness.

Conclusions

We presented work on an implemented model of situated dialogue processing. The model is based on the idea that to understand situated dialogue, linguistic meaning needs to be coupled to the situated context. Processing dialogue incrementally, information about the dialogue- and situated context can help at each step to focus the linguistic analysis. The implemented has been evaluated on a data set of 58 utterances formulated on 11 different visual scenes. Investigating the effects of using linguistic knowledge, the results show that using such knowledge can greatly improve the performance of an incremental parser, but cannot fully reduce linguistic ambiguity. This confirms the need for including information about the situated context to further reduce that ambiguity. We are currently planning follow-up evaluations that will investigate these effects further.

Acknowledgements

References

Allen, J., Byron, D., Dzikovska, M., Ferguson, G., Galescu, L., and Stent, A. (2001). Towards conversational human-computer interaction. *AI Magazine*, 22(4):27–37.

Allen, J., Miller, B., Ringger, E., and Sikorski, T. (1996). A robust system for natural spoken dialogue. In *Proceedings of the 34th Meeting of the Association for Computational Linguistics (ACL'96)*, pages 62–70.

Allopenna, P., Magnuson, J., and Tanenhaus, M. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, 38(4):419–439.

Altmann, G. and Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73(3):247–264.

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., editors, *The Interface of Language, Vision, and Action: Eye Movements and The Visual World*, pages 347–386. Psychology Press, New York NY.

Altmann, G. M. (1988). Ambiguity in sentence processing. *Trends in Cognitive Sciences*, 2(4).

Altmann, G. T. and Steedman, M. (1988). Interaction with context during human sentence processing. *Cognition*, 30(3):191–238.

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

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

Baldridge, J. and Kruijff, G.-J. M. (2003). Multi-modal combinatory categorial grammmar. In *Proceedings of EACL'03*, Budapest, Hungary.

Barsalou, L. (1999). Perceptual symbol systems. *Behavioral & Brain Sciences*, 22:577–660.

Botvinick, M., Braver, T., Barch, D., Carter, C., and Cohen, J. (2001). Conflict monitoring and cognitive control. *Psychological Review*, 108(3):624–652.

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)*.

Brick, T. and Scheutz, M. (2007). Incremental natural language processing for HRI. In *Proceeding of the ACM/IEEE international conference on Human-Robot Interaction (HRI'07)*, pages 263 – 270.

Carroll, J. and Oepen, S. (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.

Chambers, C., Tanenhaus, M., and Magnuson, J. (2004). Actions and affordances in syntactic ambiguity resolution. *Jnl. Experimental Psychology*, 30(3):687–696.

Crain, S. and Steedman, M. (1985). On not being led up the garden path: The use of context by the psychological syntax processor. In Dowty, D. R., Karttunen, L., and Zwicky, A. M., editors, *Natural language parsing: Psychological, computational, and theoretical perspectives*. Cambridge University Press.

Dahan, D. and Tanenhaus, M. (2004). Continuous mapping from sound to meaning in spoke-language comprehension: Immediate effects of verb-based thematic constraints. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(2):498–513.

De Vega, M., Robertson, D., Glenberg, A., Kaschak, M., and Rinck, M. (2004). On doing two things at once: Temporal constraints on actions in language comprehension. *Memory and Cognition*, 32(7):1033–1043.

DeVault, D. and Stone, M. (2003). Domain inference in incremental interpretation. In *Procedings of the Fourth Workshop on Inference in Computational Semantics (ICOS'04)*.

Endsley, M. (2000). Theoretical underpinnings of situation awareness: A critical review. In Endsley, M. R. and Garland, D. J., editors, *Situation awareness analysis and measurement*. Lawrence Erlbaum.

Fodor, J. (1983). *The Modularity of Mind*. The MIT Press, Cambridge MA.

Glenberg, A. (1997). What memory is for. *Behavioral & Brain Sciences*, 20:1–55.

Glenberg, A. and Kaschak, M. (2002). Grounding language in action. *Psychonomic Bulletin & Review*, 9(3):558–565.

Gorniak, P. and Roy, D. (2004). Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research*, 21:429–470.

Gorniak, P. and Roy, D. (2005). Probabilistic grounding of situated speech using plan recognition and reference resolution. In *Proceedings of the Seventh International Conference on Multimodal Interfaces (ICMI 2005)*.

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

Hadelich, K. and Crocker, M. (2006). Gaze alignment of interlocutors in conversational dialogues. In *Proc. 19th CUNY Conference on Human Sentence Processing*, New York, USA.

Hawes, N., Sloman, A., Wyatt, J., Zillich, M., Jacobsson, H., Kruijff, G., Brenner, M., Berginc, G., and Skocaj, D. (2007a). Towards an integrated robot with multiple cognitive functions. In *Proceedings of the Twenty-Second Conference on Artificial Intelligence (AAAI'07)*.

Hawes, N., Zillich, M., and Wyatt, J. (2007b). 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.

Hommel, B., Ridderinkhof, K., and Theeuwes, J. (2002). Cognitive control of attention and action: Issues and trends. *Psychological Research*, 66:215–219.

Kamide, Y., Altmann, G., and Haywood, S. (2003). The timecourse of prediction in incremental sentence processing: Evidence from anticipatory eye-movements. *Jnl. Memory and Language*, 49(1):133–156.

Kelleher, J. (2005). Integrating visual and linguistic salience for reference resolution. In *Proceedings of the 16th Irish conference on Artificial Intelligence and Cognitive Science (AICS-05)*.

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

Kruijff, G., Kelleher, J., and Hawes, N. (2006). Information fusion for visual reference resolution in dynamic situated dialogue. In André, E., Dybkjaer, L., Minker, W., Neumann, H., and Weber, M., editors, *Perception and Interactive Technologies (PIT 2006)*. Spring Verlag.

Kruijff, G., Zender, H., Jensfelt, P., and Christensen, H. (2007). Situated dialogue and spatial organization: What, where... and why? *International Journal of Advanced Robotic Systems, Special Issue on Human and Robot Interactive Communication*, 4(1):125–138.

Liversedge, S. and Findlay, J. (2000). Saccadic eye movements and cognition. *Trends in Cognitive Science*, 4(1):6–14.

Mori, D., Matsubara, S., and Inagaki, Y. (2001). Incremental parsing for interactive natural language interface. In *2001 IEEE International Conference on Systems, Man, and Cybernetics*, volume 5, pages 2880–2885.

Nieuwland, M. and Van Berkum, J. (2006). When peanuts fall in love: N400 evidence for the power of discourse. *Journal of Cognitive Neuroscience*, 18(7):1098–1111.

Novick, J., Trueswell, J., and Thompson-Schill, S. (2005). Cognitive control and parsing: Reexamining the role of Broca's area in sentence comprehension. *Cognitive, Affective, and Behavioral Neuroscience*, 5(3):263–281.

Oepen, S. and Carroll, J. (2000). Ambiguity packing in constraint-based parsing: Practical results. In *Proceedings of the 6th Applied Natural Language Processing Conference (ANLP 2000)*, pages 162–169.

Pickering, M. and Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27:169–225.

Rosé, C. P., Roque, A., Bhembe, D., and VanLehn, K. (2002). An efficient incremental architecture for robust interpretation. In *Proceedings of the Human Languages Technologies Conference*.

Scheutz, M., Eberhard, K., and Andronache, V. (2004). A real-time robotic model of human reference resolution using visual constraints. *Connection Science Journal*, 16(3):145–167.

Sikkel, K. (1999). *Parsing Schemata*. Springer Verlag.

Spivey, M. and Tanenhaus, M. (1998). Syntactic ambiguity resolution in discourse: Modeling the effects of referential context and lexical frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24:15211543.

Spivey, M., Trueswell, J., and Tanenhaus, M. (1993). Context effects in syntactic ambiguity resolution: discourse and semantic influences in parsing reduced relative clauses. *Canadian Journal of Experimental Psychology*, 47(2):276–309.

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

Stone, M. and Doran, C. (1997). Sentence planning as description using tree-adjoining grammar. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL'97)*, pages 198–205.

Tanenhaus, M., Magnuson, J., Dahan, D., and Chambers, G. (2000). Eye movements and lexical access in spoken-language comprehension: Evaluating a linking hypothesis between fixations and linguistic processing. *Journal of Psycholinguistic Research*, 29(6):557–580.

Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., and Sedivy, J. (1994). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268:1632–1634.

Van Berkum, J. (2004). Sentence comprehension in a wider discourse: Can we use erps to keep track of things? In Carreiras, M. and Jr., C. C., editors, *The on-line study of sentence comprehension: Eyetracking, ERPs and beyond*, pages 229–270. Psychology Press, New York NY.

van Berkum, J., Brown, C., and Hagoort, P. (1999a). Early referential context effects in sentence processing: Evidence from event-related brain potentials. *Journal of Memory and Language*, 41:147–182.

Van Berkum, J., Brown, C., Zwitserlood, P., Kooijman, V., and Hagoort, P. (2005). Anticipating upcoming words in discourse: Evidence from erps and reading times. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 31(3):443–467.

van Berkum, J., Hagoort, P., and Brown, C. (1999b). Semantic integration in sentences and discourse: Evidence from the n400. *Journal of Cognitive Neuroscience*, 11(6):657–671.

Van Berkum, J., Zwitserlood, P., Brown, C., and Hagoort, P. (2003). When and how do listeners relate a sentence to the wider discourse? evidence from the n400 effect. *Cognitive Brain Research*, 17:701–718.

Van Petten, C., Coulson, S., Rubin, S., Plante, E., and Parks, M. (1999). Time course of word identification and semantic integration in spoken language. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25(2):394–417.