Producing Believable Robot Gaze When Comprehending Visually Situated Dialogue

Geert-Jan M. Kruijff[1] and Maria Staudte[2]

[1]Language Technology Lab, German Research Center for Artificial Intelligence (DFKI GmbH), Saarbrücken Germany

[2]Dept. of Computational Linguistics, Saarland University, Saarbrücken Germany

gj@dfki.de

Abstract

The paper presents an implemented approach to producing robot gaze during comprehending visually situated dialogue. The approach is based on an incremental model for processing situated dialogue. In this model, utterance interpretations are build step-by-step, in a "left-to-right" fashion. At each step, grammatical and dialogue-level information is combined with information about the visually situated context. As a consequence, utterance processing can be guided so as to construct only situationally appropriate interpretations. Furthermore, at each step a set of visual referents is determined, to which the unfolding utterance meaning is currently making reference. In the approach, this information is used to drive robot gaze, letting the robot change its fixation onto the most recent visual referent. The underlying assumption is that gaze behavior helps to establish joint attention ("common ground") in a dialogue, if there is congruency between where the robot is looking, and what the (intended) visual referent is. The paper reports on a pilot study in which this assumption is studied. The results show statistically significant interactions between congruence, believability, and appropriateness of referring expression.

Introduction

In situated dialogue, people not only talk – they also look. They look at the visual objects they believe are being referred to. This serves a fundamental function in dialogue. By aligning what they attend to in the visual context, the resulting *joint attention* indicates that they share the same understanding of what is being talked about (Garrod and Pickering, 2004; Pickering and Garrod, 2004).

In this paper, we discuss an implemented approach that makes a robot produce similar behavior when it is trying to understand an utterance. Several empirical studies have confirmed that such robot behavior would make human-robot interaction more natural (Breazeal et al., 2004a; Miyauchi et al., 2004; Sidner et al., 2004; Sidner et al., 2005; Yoshikawa et al., 2006). It is still an open question though how to produce such behavior in a way that it really takes the situation into account. Current approaches primarily rely on scripted behaviors which are not grounded in the visual context.

The approach we present relies on explicitly grounding dialogue in the situated context. The main idea is to use an *incremental* model for dialogue analysis, and step-by-step connect the linguistic representations with information about the visually situated context. From this interconnection we can then derive what the visual objects are that are being talked about, and so drive the robot's *gaze* – i.e. what objects it should fixate on, and when it should move from looking at one object to the next. 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). We have performed a pilot study to empirically evaluate our approach.

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). Like (Brick and Scheutz, 2007) 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 advance on (Brick and Scheutz, 2007) by analyzing utterance meaning incrementally also relative to the structure of the dialogue context, allowing different levels of linguistic description to constrain possible interpretations (Stone and Doran, 1997). We adopt a "packed" representation of the linguistic analyses (Oepen and Carroll, 2000; Carroll and Oepen, 2005) to efficiently handle alternative (i.e. ambiguous) meanings. 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, it can be discarded. Whenever a step in the incremental utterance analysis introduces a new object in the utterance meaning, we thus get a set of possible visual referents for that object description. The basic idea in producing "gaze" is to let the robot look (i.e. fixate) at the visual referent(s) for the most recently added object(s).

An overview of the paper is as follows. We first provide further background to our approach. We discuss relevant psycholinguistic insights in what factors tend to influence understanding situated language, and position our approach in more detail to the current state-of-the-art. We then present our approach in detail. We discuss the cognitive architecture schema we employ (Hawes et al., 2007a; Hawes et al., 2007b), our incremental approach to multi-

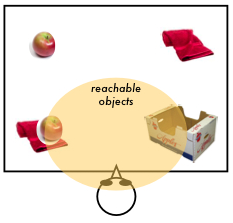
level dialogue analysis, how utterance meaning gets interconnected with the broader context – and how gaze behavior is driven by the resolution of the visual objects that we find the utterance refers to, as we incrementally analyse its possible meanings. Finally, we discuss the results of a pilot experiment we have performed. The pilot investigates the extent to which a user believes the robot has understood what was said on the basis of where the robot looks. Results show statistically significant interactions between believability, and congruency between gaze and intended visual referent. The paper ends with conclusions.

Background

Empirical studies in psycholinguistics have investigated what information listeners use when comprehending spoken utterances. These studies use eye-trackers to monitor where people look at in a scene, and when. Knoeferle & Crocker (Knoeferle and Crocker, 2006) argue that these findings identify two core dimensions of the interaction between language and situated experience. One is the *temporal dimension*: Eye movements during utterance comprehension reveal that visual attention is closely time-locked with utterance comprehension. The second is the *information dimension*, indicating how for utterance comprehension listeners draw not only upon linguistic information, but also upon scene understanding and "world knowledge." Below we discuss studies investigating the latter two aspects.

Altmann & Kamide (Altmann and Kamide, 1999) show that listeners use "world knowledge" to anticipate what will be mentioned next in an utterance. They focus their attention on objects before these objects are explicitly referred to. For example, 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. 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 fixates on that object.

Also scene understanding influences how we understand an utterance. For example, consider the figure to the right. Tanenhaus et al (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 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. In (Tanenhaus et al., 1994) the listener cannot directly manipulate the objects. If this is possible, Chambers et al (Chambers et al., 2004) show that also reachability plays a role. Because the listener can only grasp the apple that is on the towel, this is taken as the preferred referent.



These the studies thus show that gaze fixations are derived from how we can resolve a visual referent for an object reference. In establishing referents, listeners use visual and spatial properties of objects, combined with visual salience and "topokinetic" salience derived from object reachability.

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). They 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 that analysis from the set of analyses maintained by the parser.

We use this approach to incremental language processing as the basis for producing gaze fixations and -movements. The basic idea is simple. Whenever a new object description is introduced in the unfolding utterance meaning, we determine the set of possible visual referents. We then let the robot fixate at the visual referent(s)

for the most recently added object(s). Although simple, this approach sets us apart from several other approaches to producing gaze in human-robot interaction. Most approaches adopt fixed scripted behaviors to drive gaze (Sidner et al., 2004), or make the robot look at an "area of change" to signal understanding (Breazeal et al., 2004b). Alternatively, the robot is made to exactly mimic its human partner (Yoshikawa et al., 2006). The problem with these systems is that gaze is not produced on the basis of a deeper understanding of the situation, and how dialogue refers to that situation. This results in a rigid and merely reactive behavior that is not flexible enough to adapt to novel situations.

The approach to producing robot gaze we propose here is a natural extension of an incremental model of situated dialogue processing. Incrementally construed linguistic meaning gradually becomes grounded in the social and physical context in which the dialogue takes place, in ways that reflect the unique and dynamic nature of situations.

Approach

We have implemented our approach in an cognitive architecture based on 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 gaze and dialogue processing.

Cognitive architecture

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

Typically, a subsystem establishes 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; 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.

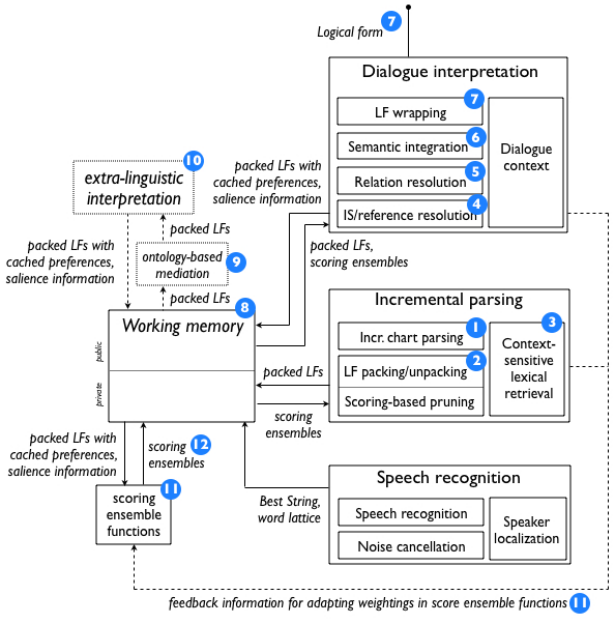


Figure 1: Dialogue processing (comprehension part)

Dialogue analysis

Figure 1 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 is 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 lexicon entry specifies for a word all its possible syntactic and semantic uses. During parsing, this information is used to integrate

[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.

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

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.

Semantic structure is represented as an ontologically richly sorted, relational structure – a logical form (Baldridge and Kruijff, 2002). 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 2 illustrates a packed representation of intermediate logical forms for ”put the ball to the left of the box”, packing together 30 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 §). 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.

Resolving referents

In the architecture discussed here we rely for visual referent resolution on a grounding process called *binding*. The basic idea is illustrated in Figure 3. 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 that can 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.

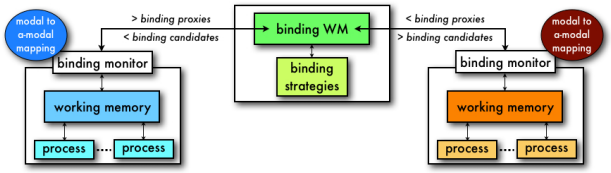


Figure 3: Cross-modal interconnectivity

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.

Producing gaze

The result of binding is that we obtain, after each incremental interpretation step, a set of one or more visual referents for the objects represented by the packed logical forms. Depending on whether binding is able to resolve any syntactic ambiguities (as in e.g. ”put the apple on the towel ... ”), the set of referents may present referential ambiguity, or not.

The gaze subsystem monitors the binding working memory for unions of recently added proxies coming from the dialogue subsystem, bound to visual entities. Based on the (un)ambiguity of these unions, and the completeness of the linguistic analyses, the gaze subsystem will then produce one of the following behaviors:

Saccade from speaker to unambiguous visual referent, fixation: At the start of hearing a new utterance, the robot is looking at the user. As soon as the first visual referent is established, the robot moves to look at the visual referent,

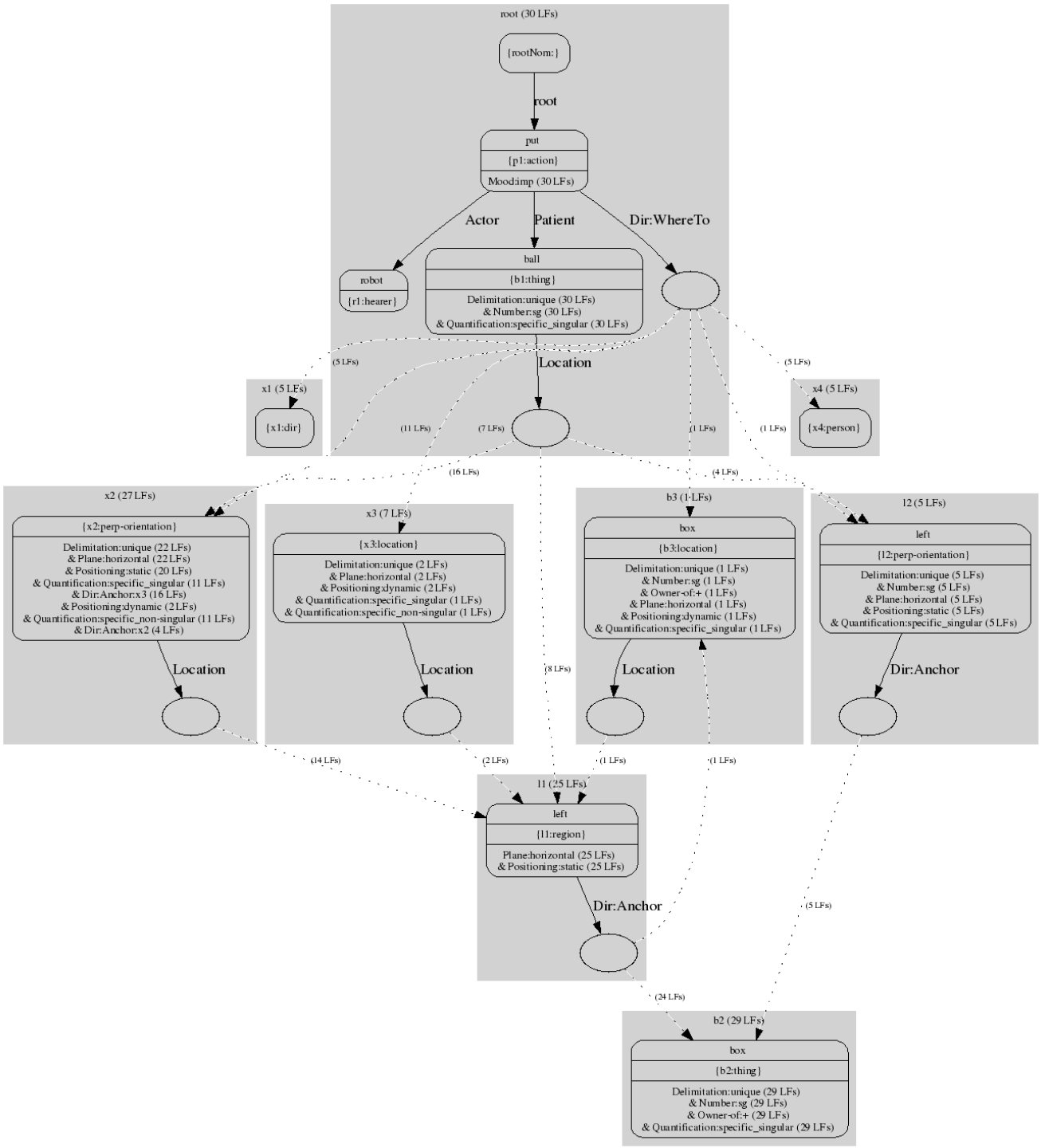


Figure 2: Example packed logical form - ”put the ball to the left of the box”

and fixates on it.

Saccade from unambiguous visual referent to next unambiguous referent, fixation: If the next expression unambiguously refers to a new visual object, the robot moves to this new object and fixates.

Saccade between ambiguous referents: If the current set of visual referents is ambiguous, the robot will move between these objects.

Saccade back to listener: Once the utterance has been fully analyzed, the robot returns to looking at the speaker.

These gaze behaviors rely in an essential way on the incremental approach to dialogue processing, as discussed above. Furthermore, a fundamental assumption is that the robot's gaze is only "natural" or believable (i.e. contributing to joint attention (Garrod and Pickering, 2004)) if it there is congruency between the intended referent and what the robot fixates on. In the next section we present the results of a pilot study which investigated this assumption.

Pilot experiment

The approach we present here is based on a fundamental assumption about the relation between gaze (notably, fixation), and visual referents. Namely, we assume that a fixation is *congruent* when the robot looks at the "right" visual referent. This has an important *hypothesized* effect on establishing joint attention in visually situated dialogue. If a robot would produce an incongruent fixation, the speaker would presumably believe that the robot did not understand her correctly. We have performed a pilot study to investigate this potential effect of incongruent fixations, in comparison to congruent fixations. This follows up on earlier studies, e.g. the one by Yoshikawa et al (Yoshikawa et al., 2006) who show that congruent gaze is perceived more natural than staring, or gaze behavior that appears independent of what the speaker communicates.

The main hypothesis for the pilot study was that "congruency between gaze fixation and intended referent leads to higher degree of belief that the robot understands which visual referent is referred to in the utterance." Or, vice versa, that incongruent gaze (i.e. fixation on the wrong visual referent) yields lower believability in the robot having understood. As baseline, we used fixation on the visually most salient item. This does not require the robot to resolve the referent. All we need to do is just trigger a behavior to look at an object. The baseline reveals how much a robot does, or does not, need to be able to relate situation awareness with dialogue processing to yield convincing interactive behavior.

The pilot study is set up as a web-experiment, in which people are shown 35 videos. Figure 4 shows a screenshot from the browser. We uploaded the videos to GoogleVideo, to make sure anyone could view them independent of platform or browser. Each video shows a visual scene of a robot with an arm, standing at a table-top scene including two or more colored objects. Some of these objects the robot is capable of manipulating, some not. Then, the robot is told an utterance, in which one visual object is referred to. While comprehending the utterance, the robot subsequently fixates at one of the objects in the visual scene. Each video takes approximately 7 to 8 seconds. After the video, the subject is asked two questions:

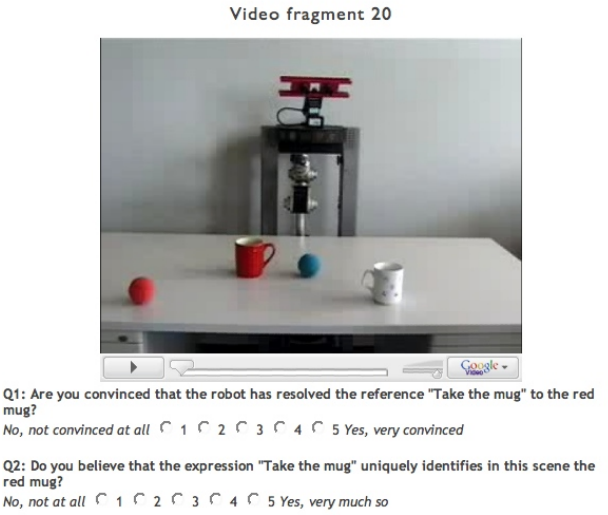


Figure 4: Browser screenshot

Q1 "Are you convinced that the robot has managed to resolve the reference XYZ to the right object (namely, Q)?" (Answer on a 5-point Lykert scale, "(1) not convinced at all ... (5) yes, totally convinced.")

Q2 "Do you believe that the expression XYZ is appropriate to uniquely identify the object Q in the scene?" (Answer on a 5-point Lykert scale, "(1) no, not appropriate at all ... (5) yes, very appropriate.")

We performed the pilot study with 15 subjects, 5 female and 10 male. Some of these subjects were familiar with robots, though none with our system. We solicited subjects by email. Subjects were not offered any financial compensation. Each subject was given the following information.

Nature of the workspace The subject is told where the robot can reach.

Nature of the objects The subject is told that all objects can be referred to as "things", which objects the robot can grasp, and that the robot can push all objects (within reach).

Nature of the instructions told to the robot The subject is told the robot may be given a command to manipulate an object, or just a description of an object in the scene

Below we discuss in more detail the principled design approach we took for generating the visual scenes for the videos, and present the results and their discussion.

Design

The point of the pilot study was to investigate congruency between gaze fixation, and visual referents. We therefore

needed to design the visual scenes for the videos in such a way that we would control the factors that influence the (potential) ambiguity of a referring expression.

For each scene, we wanted to consider a number of scenarios. Given the baseline of fixating on the visually most salient object, we wanted to systematically vary the objects and scene structure relative to two fundamental conditions: (1) the intended referent has the same visual salience as a distractor, or (2), the intended referent has a lower visual salience than the visually most salient object. This implies that for a referring expression, if the visually most salient item is not in the distractor set for the expression, incongruency arises automatically (i.e. baseline gaze, versus congruent gaze).

To bring about potential (in)congruency in these conditions, we thus needed to consider the contrast between the visually most salient item, and the intended visual referent. Following (Dale and Reiter, 1995; Kelleher and Kruijff, 2006) we set up a basic template for a visual object, consisting of its material and contrastive properties, spatial relations, and visual and topokinetic salience.

Then, given a visual salience condition, we selected a set of two or more objects that would enable us to systematically vary distractor factors (relative to the intended referent) based on type, and material and contrastive properties. Subsequently, using a 4 by 5 matrix grid, we positioned objects in the scene such that we obtained the desired visual salience condition, topokinetic salience, and spatial relations that could be used to uniquely identify a referent. Figure 5 shows the salience measures (relative to the robot's viewpoint) we used in determining how these measures acted as distractor factors.

Figure 6 gives an example of a visual scene. For the condition of distinct visual salience, the utterance "Look at the ball" would have as intended referent $b2$, but would yield an incongruent gaze fixation at $m1$ under a baseline behavior ($m1$ being the visually most salient object).

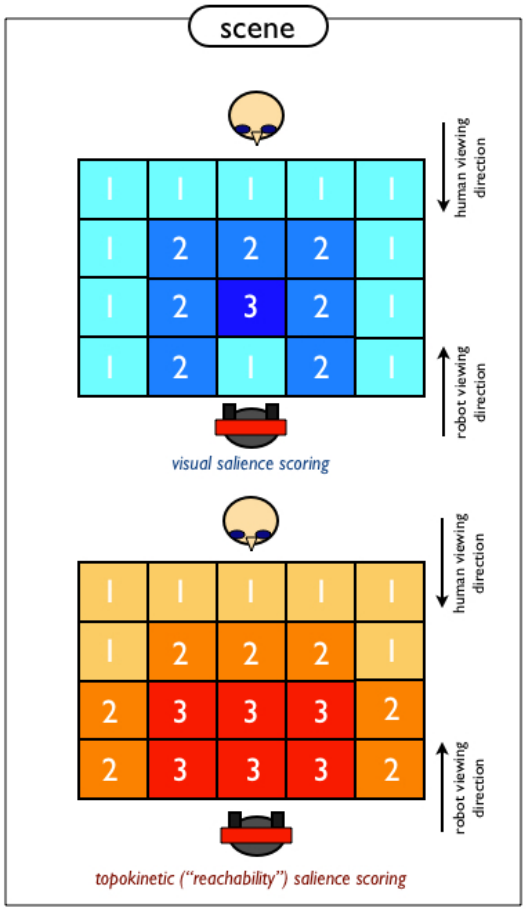


Figure 5: Visual and topokinetic salience measures

Results

We analyzed the results from the pilot study relative to the visual salience conditions. Within these conditions, we looked at two types of variance: (1) Variance in the relation between congruence/incongruence and believability (question Q1), and (2) variance in the relation between congruence/incongruence, believability (question Q1), and appropriateness of the referring expression (question Q2).

The first type of variance reveals the basic impact of (in)congruence on believability (one-way ANOVA). The second type reveals more about the relation between congruence, believability, and how much information we need in a referring expression to rule out distractors. By computing the variance in the latter type we can investigate the role visual salience plays as distractor factor.

Table 1 gives the results for the variation between congruency and believability. Results are statistically significant (p=0.001) across both conditions. Figure 7 shows the boxplot for the condition different visual salience; the

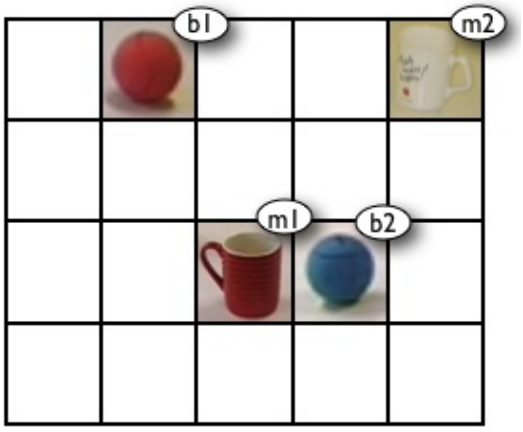


Figure 6: Sample visual scene

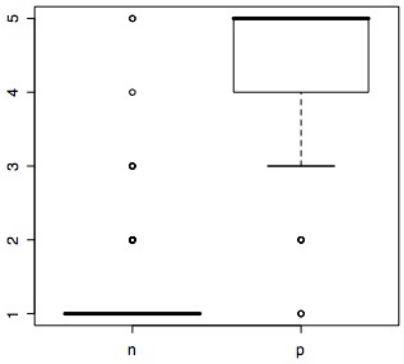


Figure 7: Boxplot: Congruency ∼ believability (diff. vis. salience)

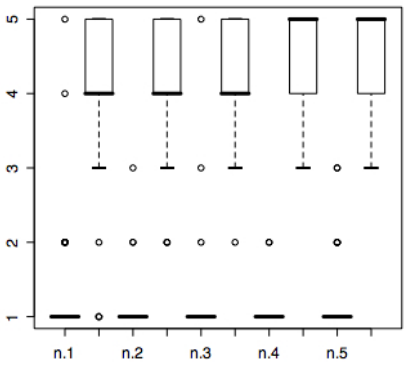


Figure 8: Boxplot: Believability ∼ congruency * appropriateness (diff. vis. salience)

boxplot for the other condition is similar.

Condition	F-value	Significance
Distractors with eq. vis. salience	961.1	0.001
Distractors with diff. vis. salience	301.08	0.001

Table 1: Variance: Congruency ∼ Believability (Q1) (one-way ANOVA)

Table 2 gives the results for the variation between congruency, believability, and appropriateness of the referring expression to uniquely identify the intended visual referent (two-way ANOVA). Again, results are statistically significant (p=0.001) across both conditions.Figure 8 and Figure 9 show the boxplots for the conditions different respectively equal visual salience.

Condition	F-value	Significance
Distractors with eq. vis. salience	299.16	0.001
Distractors with diff. vis. salience	1005.67	0.001

Table 2: Variance: Believability (Q1)∼ Congruency * Appropriateness (Q2) (two-way ANOVA)

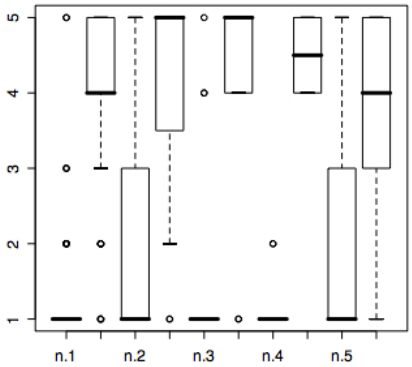


Figure 9: Boxplot: Believability ∼ congruency * appropriateness (eq. vis. salience)

Discussion

The results all show statistically significant interactions between congruency of fixation, and believability. The results thus confirm the main hypothesis of the pilot study.

Across the conditions, we can see an interesting pattern appear. In the condition under which visual objects are distractors because they have equal visual salience, incongruency is particularly negative (Table 1). The importance of proper linguistic reference, i.e. the production and comprehension of contextually appropriate referring expressions, because clear if we combine this result with the variance in relation to appropriateness (Table 2 and Figure 9). We need further experimentation to determine the exact impact of the different distractor factors on resolution of visual referents. Having said that, we see these results as strengthening the argument that natural language processing for human-robot interaction requires taking into account semantic and pragmatic factors – well beyond the level of simple syntactic analysis – if we want robots to produce believable gaze.

This conclusion may be strengthened further if we look at the other condition, in which distractors differ in visual salience. In this condition, incongruence arises automatically if referring expressions are not resolved. Tables 2 and 1 show the sharp contrast that can be observed between congruent and incongruent fixations in this case.

Conclusions

The paper presented an approach to robot gaze production, which we implemented using the CAST framework. The core of the approach is constituted by an incremental model of dialogue analysis, and the possibility to bind utterance meaning to visual referents. Based on what referents are becoming referred to as the utterance analysis unfolds, robot gaze is driven to move its fixation from one visual object to another. The approach is based on the assumption that, for such gaze to contribute to establishing joint attention in situated dialogue, fixations need to be congruent with the (intended) visual referents. We presented a pilot study which showed statistically significant interactions between congruence, believability, and appropriateness of referring expression – thus providing initial

empirical support for the approach.

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.

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. 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.

Breazeal, C., Brooks, A., Gray, J., Hoffman, G., Kidd, C., Lee, H., Lieberman, J., Lockerd, A., and Mulanda, D. (2004a). Humanoid robots as cooperative partners for people. *Int.Jnl. Humanoid Robots*.

Breazeal, C., Hoffman, G., and Lockerd, A. (2004b). Teaching and working with robots as a collaboration. In *Proceedings of Third International Joint Conference on Autonomous Agents and Multi Agent Systems (AAMAS'04)*, pages 1028–1035, New York, NY.

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.

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

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

Garrod, S. and Pickering, M. (2004). Why is conversation so easy? *Trends in Cognitive Sciences*, 8:8–11.

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.

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.

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)*.

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

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.

Miyauchi, D., Sakurai, A., Makamura, A., and Kuno, Y. (2004). Active eye contact for human-robot communication. In *Proceedings of CHI 2004*, pages 1099–1104. ACM Press.

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.

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.

Sidner, C. L., Kidd, C. D., Lee, C. H., and Lesh, N. (2004). Where to look: A study of human-robot engagement. In *ACM International Conference on Intelligent User Interfaces (IUI)*, pages 78–84. ACM.

Sidner, C. L., Lee, C., Kidd, C., Lesh, N., and Rich, C. (2005). Explorations in engagement for humans and robots. *Artificial Intelligence*, 166(1-2):140–164.

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

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

Yoshikawa, Y., Shinozawa, K., Ishiguro, H., Hagita, N., and Miyamoto, T. (2006). Responsive robot gaze to interaction partner. In *Proceedings of Robotics: Science and Systems II (RSS'06)*, pages 287–294.