The Roles of Haptic-Ostensive Referring Expressions in Cooperative, Task-based Human-Robot Dialogue*

Mary Ellen Foster† Ellen Gurman Bard‡ Markus Guhe‡
Robin L. Hill‡ Jon Oberlander‡ Alois Knoll†
† Informatik VI: Robotics and Embedded Systems,Technische Universität München
‡ Human Communication Research Centre, University of Edinburgh

ABSTRACT

Generating referring expressions is a task that has received a great deal of attention in the natural-language generation community, with an increasing amount of recent effort targeted at the generation of multimodal referring expressions. However, most implemented systems tend to assume very little shared knowledge between the speaker and the hearer, and therefore must generate fully-elaborated linguistic references. Some systems do include a representation of the physical context or the dialogue context; however, other sources of contextual information are not normally used. Also, the generated references normally consist only of language and, possibly, deictic pointing gestures.

When referring to objects in the context of a task-based interaction involving jointly manipulating objects, a much richer notion of context is available, which permits a wider range of referring options. In particular, when conversational partners cooperate on a mutual task in a shared environment, objects can be made accessible simply by manipulating them as part of the task. We demonstrate that such expressions are common in a corpus of human-human dialogues based on constructing virtual objects, and then describe how this type of reference can be incorporated into the output of a humanoid robot that engages in similar joint construction dialogues with a human partner.

Categories and Subject Descriptors

H.5.2 [**User Interfaces**]: Natural language; I.2.9 [**Robotics**]: Operator interfaces

General Terms

Design, Human Factors

Keywords

Multimodal dialogue, referring expressions

*This research was supported by the EU project JAST (FP6-003747-IP), `http://www.euprojects-jast.net/`.

HRI'08, March 12–15, 2008, Amsterdam, The Netherlands.

1. INTRODUCTION

When agents—human or artificial—work together on a task involving manipulating objects, an important communicative function is indicating to a conversational partner which of a set of available domain entities should be used. In the natural-language generation (NLG) community, this is a core task called *generation of referring expressions* (GRE); that is, selecting an expression to identify an entity from a set of entities that can be referred to, in a context available to both the speaker and the hearer. Many approaches to this task have been implemented in a number of systems. In fact, the task of attribute selection for the generation of referring expressions was recently the subject of a pilot shared-task NLG evaluation challenge [3].

In terms of referring expressions, *accessibility models* [1] suppose that full referring expressions with elaborate structure and lexical content are needed to make an initial reference to an entity. Such expressions are used for *low accessibility* entities: those which are new to a discourse. Elaborate referring expressions can bring an unattended or novel entities into focus, making new objects more accessible targets for succeeding expressions of lesser elaboration (and accordingly higher accessibility). In practical terms, this concept has set the task for implemented models: they generally aim to generate first mentions of inaccessible objects, which normally consist of noun phrases with articles and modifiers.

However, not all first mentions are made in circumstances where the target is initially inaccessible. In a shared workspace, for example, more effective methods of drawing a partner's attention to a task-critical object might be to point at it, to grasp it and hold it out, or even simply to use it as part of carrying out the task. Kranstedt and Wachsmuth [13] claim that, in such cases, pointing—what we will call *gestural-deictic reference*—is preferable to any linguistic expression. In this paper we describe a another kind of reference, which we observe in common use between human collaborators in a joint task: *haptic-ostensive reference*, that is, reference which involves manipulating an object. We claim that this action, typically accompanied by deictic verbal references, brings the target into the focus of the collaborators' attention. By performing such manipulation actions in haptic-ostensive references, a speaker can use accompanying linguistic expressions that are reduced and, therefore, less costly to plan and communicate than typical NLG initial mentions, by performing actions that are required in any case to support the mutual task. In other words, performing the action makes the entity referred to more accessible.

We begin by surveying approaches to referring-expression generation and discussing the range of linguistic and multimodal behaviour that are supported by current systems. We first give a detailed description of the classic incremental algorithm, which selects attributes to include in purely linguistic references, assuming zero shared knowledge between the speaker and the hearer. We then discuss a range of extensions to this algorithm that add various notions of physical and discourse context to the generation process. These additional contextual factors increase the accessibility of the target objects and hence decrease the elaborateness required in the linguistic reference, which both simplifies the generation task and tends to produce references more similar to those used by humans in practice.

We then turn our attention to haptic-ostensive references, a form of reference that is particularly afforded by task-based interactions in which the participants work together on a common task in a shared workspace. First, we look at the referring phenomena found in a corpus of human-human dialogues where the partners work together to assemble objects in a virtual world. In particular, we concentrate on those initial references to world objects where a linguistic description is combined with manipulating an object in a shared workspace; such references are very common in this task-based domain. We then introduce a human-robot dialogue system that allows the user to work together with a robot on a similar cooperative assembly task in the physical world and show how such haptic-ostensive references can be usefully incorporated into this scenario.

2. GENERATING MULTIMODAL REFERRING EXPRESSIONS

Generating referring expressions, linguistic or multimodal, is one of the classic tasks in natural-language generation, and a number of algorithms have been proposed to address this task. This section provides a summary of the main approaches, beginning with the classic incremental algorithm and then describing a number of recent extensions that add richer notions of context to the basic algorithm.

2.1 The Incremental Algorithm

The classic algorithm in referring-expression generation—and the one on which most subsequent implementations are based—is the well-known *incremental algorithm* by Dale and Reiter [7], which selects a set of attributes of a target object to single it out from a set of distractor objects. The algorithm incrementally selects attributes of the object that at least one object from the distractor set does not share. The selected attribute is then used in the referring expression, and the objects without the attribute are removed from the distractor set. This process is executed repeatedly until only the intended object (the target object) remains in the distractor set.

As a concrete example, consider the following object set:

1. big, red, striped fish
2. small, green, striped fish
3. big, red, striped bug
4. tiny, red, spotted bug
5. small, green, spotted fish

In this case, the initial distractor set is $\{1, 2, 3, 4, 5\}$. To generate a reference to the last object in the list, the algorithm goes through a list of preferred attributes in a fixed order, where this order is selected ahead of time; for this example, we will use the list $<$ *type*, *colour*, *size*, *pattern* $>$. In the first iteration, the algorithm takes the *type* attribute and removes all objects from the distractor set that do not match the type of the intended object (fish). This means that objects 3 and 4 are removed from the distractor set, which therefore becomes $\{1, 2, 5\}$. Because using the *type* attribute removes at least one object from the distractor set, it is included in the referring expression.[1]

Because there are still elements in the distractor set other than the intended referent, the algorithm executes a second iteration. This time, it takes the *colour* attribute from the preferred attributes list and removes all elements from the distractor set that do not match the colour of the intended referent (green). This means that object 1 is taken out, so the distractor set is now $\{2, 5\}$. Since the size of the distractor set has again decreased, the *colour* attribute is added to the set of attributes to be used in the referring expression, which becomes $\{$*type*, *colour*$\}$.

In the third iteration, the *size* attribute (small) does not remove a further element from the set, as both remaining entities have the same size, so it is not selected for use in the referring expression. Finally, the *pattern* attribute is considered. Because it rules out object 2, the final distractor set is now $\{5\}$; that is, it contains just the intended object. This means that the algorithm terminates, and the set of attributes to be used in the referring expression is $\{$*type*, *colour*, *pattern*$\}$, resulting in a referring expression like *the spotted green fish*.

To produce expressions that are easy for hearers to interpret, it is important that the (fixed) attribute list be ordered so that the initial attributes are easy to detect. In the example above, *type* and *colour* are placed at the start of the list for exactly this reason. Note that the *colour* attribute is not actually needed to single out the object; referring to it as *the spotted fish* would suffice. However, due to the incremental nature of the algorithm, once an attribute is selected it cannot be "unselected" again. This greedy aspect of the algorithm means that the referring expressions it selects are not necessarily the shortest; however, the algorithm is made computationally tractable, while finding the shortest description is in general an *NP*-hard problem. Also, Dale and Reiter argue that the expressions produced by their algorithm are similar to those produced by humans, who also do not generally choose the shortest possible description.

2.2 Extensions to the Incremental Algorithm

The incremental algorithm outlined in the preceding section assumes that the only common ground between the producer of the referring expression and the audience is a shared knowledge of the features of all of the objects in the world. Since no other knowledge is shared between the interlocutors, it is necessary to create a fully-elaborated linguistic expression in order to refer to the target successfully. Since

[1]Because the *type* attribute is usually realised as a noun, and a noun is an essential part of a noun phrase (which is how a referring expression is normally realised), the *type* attribute is usually selected regardless of whether it also serves to remove elements from the distractor set. However, this point is not relevant for the discussion in this paper.

the initial description of the incremental algorithm, a number of people have proposed extensions to take into account various notions of salience and context to deal with the fact that, in practice, the speaker and the hearer quite often have more context in common.

Kelleher and Kruijff [11], for example, implemented an algorithm to generate linguistic spatial referring expressions in situated dialogue. They extended the incremental algorithm in two ways: by adding a notion of visual and discourse salience, and by constructing a context model based on a set of reduced scene models rather than on a single, complex, exhaustive model. Their algorithm makes use of possible landmarks to generate descriptions like *the man next to the ball*, and bases its selection of attributes on a cognitively-motivated hierarchy of relations.

While Kelleher and Kruijff used properties of the visual scene, they still generated purely linguistic referring expressions. Others have added the ability to include deictic pointing into the specification of the referring expressions. Van der Sluis [20], for example, presented a graph-based algorithm that creates multimodal referring acts including pointing by assigning costs to the verbal and non-verbal components of referring expressions and then selecting the combination with minimum cost. This algorithm distinguishes between different degrees of precision in a pointing gesture, ranging from precise to very imprecise, where the cost of a pointing gesture increases with its precision.

Kranstedt and Wachsmuth [13] also proposed an algorithm for generating multimodal deixis which has a similar flavour to that described by van der Sluis. They extended the incremental algorithm by specifying two types of pointing, *object-pointing* and *region-pointing*, and gathered data from empirical studies [12] to determine the normal use of pointing in multimodal reference. They found that definite descriptions were shorter when pointing was used, and that the length and complexity of the linguistic description depended on the distance between the speaker and the target of the reference. They added *location* as an additional factor whose discriminative power is tested in the incremental algorithm like the other factors and used the modified algorithm to specify the referring behaviour of an embodied agent in a virtual world.

Piwek [15] also considered the generation of multimodal referring expressions including linguistic content and deictic gestures. He studied the referring behaviours in a corpus gathered by Beun and Cremens [4] consisting of dialogues between pairs of Dutch speakers, where one subject instructed the other in building a Lego model. He found, like others, that over half of the initial references included a pointing gesture and that the references that included a pointing gesture were significantly shorter. However, he also found that the use of pointing also depended on the speaker: in fact, some speakers never used pointing at all, while others used it very frequently.

3. HAPTIC-OSTENSIVE REFERENCE IN A SHARED WORKSPACE

The reference-generation algorithms described in the preceding section all extend the basic incremental algorithm to include additional aspects of visual context, and therefore allow the system to assume more common knowledge on the part of the hearer. This increases the accessibility of the target object and therefore in turn allows potentially simpler linguistic expressions to be used. However, all of the systems described above still generally include only the visual arrangement of objects and, possibly, the history of the interaction in their model of the common knowledge of the conversational partners. In the context of a task-based interaction where the partners share a workspace and work together towards a common goal, a much richer notion of context and a fuller set of referring acts are available.

In addition, in those cases where designers of previous systems have consulted corpora of human-generated referring expressions to help make their decisions, the corpora have generally been produced in contexts where the subjects were presented with an array of objects and asked to refer to one of them—e.g., [12, 19]. In the corpus described in [4] and analysed again in [15], the subjects were working together on a construction task; however, only one of the subjects was able to touch the pieces, so again the range of referring possibilities was somewhat limited.

In this section, we explore the referring expressions that are found in a corpus of task-based human-human dialogues where the partners work together on a common task in a shared workspace. This scenario allows for both a richer notion of context and an extended set of referring possibilities. The referring expressions can make use of the task context and the state of the workspace in addition to the history of the discourse and the current visual state. As well, since both partners are able to—and, indeed, must—manipulate objects in the world as part of the task, another possible type of reference becomes possible: *haptic-ostensive reference* [14], that is, referring to an object by manipulating it in the world.

3.1 The Joint Construction Task

The goals of the JAST project ("**J**oint **A**ction **S**cience and **T**echnology") are to investigate the cognitive, neural, and communicative aspects of jointly-acting agents, both human and artificial, and to build jointly-acting autonomous systems that communicate and work intelligently on mutual tasks. One aspect of this project is the recording and analysis the behaviour of humans cooperating with one another in the Joint Construction Task (JCT) [6], which utilises a novel experimental paradigm based around a two-person shared virtual environment. The two subjects operate separate computers but are present in the same room in order to facilitate direct communication. They cannot see each other's faces, but are able to hear each other's speech and to see each other's actions in the virtual world. Depending on the experimental condition, the partner's mouse and gaze location may also be visible on the screen. The objective is to collaboratively build "tangram"-type models from a set of geometrical components, doing so as efficiently and as accurately as possible. The different parts need to be moved, rotated and joined together. Since joining two pieces requires that each partner hold one of the pieces, it is not possible for any individual to complete the assembly process on their own, so joint action is required.

The use of a range of referring expressions was stimulated by including doubles of every part but one in order to prevent the use of simple, unique descriptions using only shape or colour (e.g., expressions such as *the purple triangle* were always ambiguous). Figure 1 gives an example screen image 10 seconds into a construction trial, showing the target

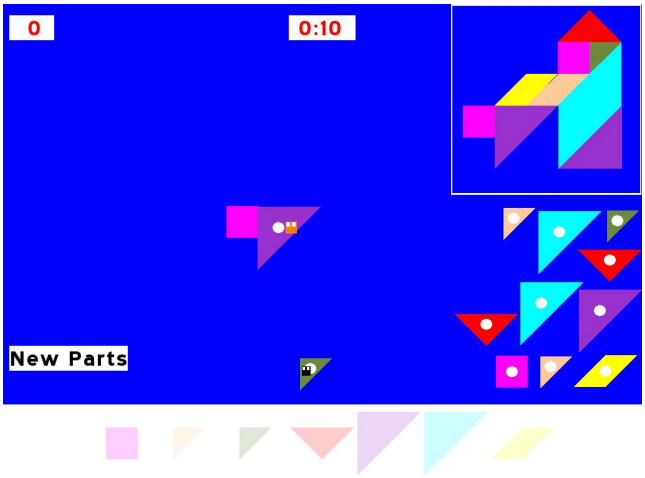

Figure 1: Interface for the Joint Construction Task

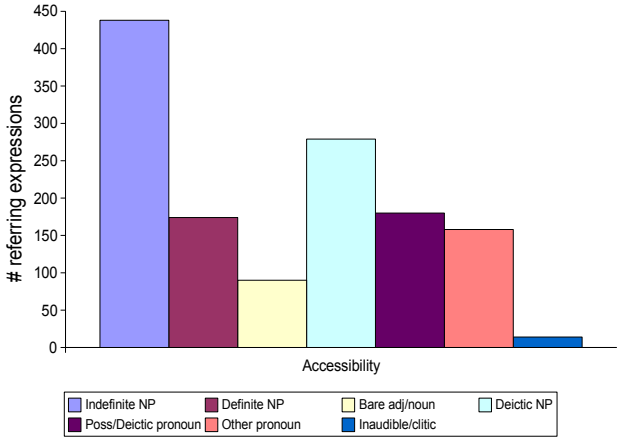

Figure 2: Number of initial mentions of different types in the JCT corpus

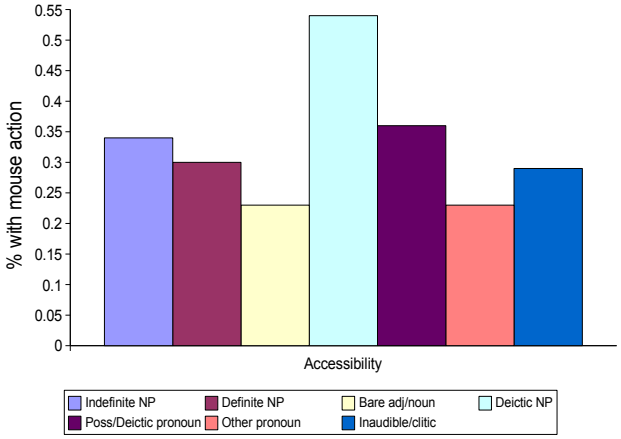

Figure 3: Percentage of initial mentions in the JCT corpus with a concurrent mouse manipulation

model at the top right and a (non-exhaustive) set of parts underneath. A subassembly consisting of a pink square and purple triangle is about to be manipulated by one person while the other person is working with a green triangle, as indicated by the position of the mouse cursors.

The multimodal JCT corpus comprises 32 dyads, each of whom completed 16 models in this environment. For each dyad, speech was permitted for half of the trials, and it is the speech during these trials that we consider here. The speech of both partners during an interaction was transcribed, and the transcribed speech was precisely time-aligned with all the visual and action components of the construction process. As well, each linguistic referring expression was annotated with its referent in the world and its degree of accessibility [1], using a similar scheme to that employed by Bard and Aylett [2]. In the following section, we describe some of the features of these referring expressions, concentrating on the *initial mentions*: that is, the first time in a given trial that a particular object is mentioned.

3.2 Referring Expressions in the JCT Corpus

Figure 2 shows the distribution of initial mentions in the corpus across the accessibility levels, ranging from indefinite noun phrases (the most elaborate expressions, indicating the lowest accessibility) through other forms of noun phrases, various types of pronouns, and finally inaudible or cliticised mentions (indicating the highest level of accessibility of the referent). As would be expected for initial mentions, the highest number of expressions are the definite and indefinite noun phrases; however, there are also a large number of deictic noun phrases (e.g., *this green triangle*), deictic pronouns (*this*, *that*), and other pronouns (*it*, *they*) among the initial mentions, indicating that objects in this domain are often highly accessible even before they are mentioned.

Of particular interest for the current study are the referring expressions that combine a linguistic reference with manipulation (moving or rotation) of the same on-screen object—that is, the *haptic-ostensive* references. Overall, about 36% of the initial linguistic references in the JCT corpus (476 of 1333) were accompanied by such a mouse manipulation. Figure 3 shows the proportion of the expressions of each accessibility class that were accompanied by an object manipulation. As can be seen, the proportion was approximately the same for all accessibility levels of the linguistic reference; the only exception was the deictic noun phrases, which had a significantly higher rate of mouse actions. In fact, 150 of the 279 deictic-NP initial mentions (54%) had a corresponding mouse action. A χ^2 test on these data indicates that that there was a strong association between the accessibility of the referring expression and the likelihood of a matching mouse action ($\chi^2 \approx 59.3$, df $= 6$, $p < 0.0001$).

Figure 4 contains excerpts from the JCT corpus illustrating typical haptic-ostensive references. In each excerpt, all verbal referring expressions are indicated, along with any mouse manipulations. For example, in excerpt 1, speaker A moved a purple triangle at the same time as saying *a purple one*; this was the first time that the purple triangle was mentioned on that trial. Similarly, in excerpt 4, *this bottom thing* referred to the square that speaker B was moving at the same time as they spoke. Many of the references from the JCT are "thinking out loud", where a speaker narrates actions as they are performed; there are also cases like excerpt 2 where a haptic-ostensive reference is clearly designed to communicate information to the partner.

1. Accessibility 0 (Indefinite NP)

A Mm I no I'm grabbing [a purple one]$_{\text{move}}$.

B Okay .

2. Accessibility 1 (Definite NP)

B Uh okay , well , we could do the top bit . See like this , do you wannna get [that red triangle] .

A Okay , okay . Um

B There the one um [the one that I'm just moving now]$_{\text{move}}$, there .

A Okay , so you're not grabbing [it] now , are you ? Okay , okay .

3. Accessibility 1 (Definite NP) and 2.5 (Deictic pronoun)

B And I'll get [this]$_{\text{move}}$

B And then [the red one]$_{\text{move}}$

A 'Kay I've got [the yellow]$_{\text{move}}$

B Cool

4. Accessibility 2 (Deictic NP)

B Here let's put [this bottom thing]$_{\text{move}}$ like that

Figure 4: Haptic-ostensive referring expressions from the JCT corpus

4. HAPTIC-OSTENSIVE REFERENCE IN HUMAN-ROBOT DIALOGUE

The previous section described referring expressions that were found in a corpus of human-human dialogues, recorded as part of the JAST project, in which the participants performed a mutual task in a shared workspace. This corpus contains referring phenomena that go beyond those covered by existing models of multimodal referring expressions: since participants could—and, indeed, had to—manipulate objects in the world as part of the task, those manipulations (moving and rotating the tangram pieces) were themselves used as methods of increasing the accessibility and licensing the use of less elaborate linguistic expressions like deictic phrases and pronouns. Following the terminology of [14], we have called such references *haptic-ostensive*.

In addition to the JCT recordings, another sub-project of JAST is the construction of a human-robot dialogue system designed to support similar task-based collaborative assembly dialogues. However, in this case, the dialogue is situated in the physical world, with a humanoid robot co-operating with a human partner. This scenario also affords similar referring phenomena. In this section, we first describe the human-robot dialogue system in detail, and then outline the roles that haptic-ostensive reference can play in dialogues between this system and a user. At the end of the section, we give some details of the implementation of the JAST system and show how referring expressions are generated.

4.1 The JAST Human-Robot Dialogue System

The JAST human-robot dialogue system [8, 17] is designed to be a platform to integrate the project's empirical findings on cognition and dialogue with research on autonomous robots, by supporting symmetrical, multimodal collaboration between a human and a robot on a joint assembly task similar to that used in the human-human JCT experiments described in the preceding section.

Figure 5: The JAST dialogue robot

The robot (Figure 5) consists of a pair of mechanical arms with grippers, mounted in a position to resemble human arms, and an animatronic talking head [18] capable of producing facial expressions, rigid head motion, and lip-synchronised synthesised speech. The input channels consist of speech recognition, object recognition, robot sensors, and face tracking; the outputs include synthesised speech, head and face motions, and actions of the robot arms. The user and the robot work jointly to assemble a wooden construction toy on a common work area, coordinating their actions through speech, gestures, and facial displays. Joint action can take several forms: for example, the robot may ask the user to provide assistance by holding one part of a larger assembly, or by assembling or disassembling components. In the current version of the system, the robot is able to manipulate objects in the workspace and to perform simple assembly tasks, and the primary form of interaction is one in which the robot instructs the user on how to build a particular compound object. Figure 6 shows a sample interaction between the user and the JAST system in which the system instructs the user on performing a particular assembly task.

Like the JCT scenario, this human-robot scenario is also structured to ensure that cooperation between the partners is essential to complete the assembly task. In this case, the common workspace is divided into three areas: one belonging to the robot, one belonging to the human, and an area in the middle to which both have access. The pieces necessary for building the desired assembly are distributed across these areas so that no single agent is able to reach all of the required components. The system tracks the location of all objects in the world and is always able to determine which are in each area of the table.

- JAST: *[picks up and holds out green cube]* First, fasten this green cube to the end of a short slat using the red bolt.
- USER: Should I use this slat? *[pointing to a five-hole slat]*
- JAST: No, use a three-hole slat. *[picking up and holding out a three-hole slat]*
- USER: *[takes slat; picks up bolt]*
- JAST: *[wiggles green cube]* When you are done, screw this cube onto the bolt.
- USER: *[takes the cube]* Okay.
- JAST: *[picks up yellow cube and orange bolt]* Now, fasten this yellow cube to the end of another short slat with this bolt.

Figure 6: Sample human-robot interaction

4.2 Roles of Haptic-Ostensive Reference

Haptic-ostensive referring actions are applicable to a variety of situations in the JAST human-robot scenario. All of these cases take advantage of the fact that the joint construction setting with a shared workspace provides rich sets both of possible actions and of contextual information that affects the accessibility of objects in the world. Since the robot's assembly abilities are limited in the current system, most of its actions consist of picking up objects and handing them over to the user for assembly; however, as the robot's own assembly abilities increase, it can also use similar actions to refer to objects as it picks them up and uses them itself, as the humans did in the JCT dialogues.

Haptic-ostensive action can be seen as an alternative to a pointing gesture. In fact, because picking up an object requires more effort, it can be seen as a more "intense" pointing gesture, and may also be a more accurate one. The advantages of picking up the object instead of simply referring by using a gestural-deictic action are that in the joint construction task, the object referred will almost certainly have to be manipulated as part of the interaction; that is, it is likely to be used as in an upcoming step of the assembly process, and would need to be picked up by one of the partners in any case. Thus, the advantage is to take this action that is necessary for task completion and to use it at the same time for drawing attention to the object. As the discussion in Section 3 showed, this combination of referring and manipulation is common in this type of joint construction interaction. Haptic-ostensive reference also allows the linguistic content of references to be less elaborate—and thus more like the expressions used by humans.

In the remainder of this section, we discuss several scenarios where haptic-ostensive reference is particularly appropriate in the JAST human-robot dialogue scenario.

Thinking out loud.

Many of the haptic-ostensive references in the JCT dialogues (e.g., Figure 4) took the form of "thinking out loud", in which one participant described the action they were performing as it took place. This type of response is especially important in a human-robot dialogue, in which it may not always be immediately obvious from the robot's motor actions what its intention is. If the system says, for example, *We need this now* while picking up a cube that is needed for the next step of the assembly plan, the user is more likely to understand the intended motor action, and the interaction is likely to be smoother.

Correcting the user.

Another, more specific situation where haptic-ostensive reference is useful is as part of making a correction to an action of the human partner. In the current system configuration, it is the robot that knows the plan of assembling the toy and that instructs the user. Thus, one of the tasks the robot performs is to instruct the human on the next step in executing the plan [9]. This means that the robot always knows which one of the objects is required at any given point in time; it also means that the robot can detect when the human is about to use an object different from the one required in that step, and that it is the responsibility of the robot to correct the human in such situations. Using the haptic-ostensive action (as in the second system turn in Figure 6), the robot can produce behaviours in which it not only verbally tells the human that he or she is using the wrong object, but in which it also picks up the correct object and hands it to the human. Correcting the user in this way helps to ensure that they actually use the correct component, as the robot is actually putting it in their hands.

Referring to an object in the robot's hand.

Finally, there is the special case in which an object is already in one of the robot's hands. This happens, for example, after a complex object, say the wing of a plane, has just been assembled. If the object is required in the next step of the plan, the easiest way for the robot to refer to the object is to keep it in its hand (instead of putting it down) and refer to it by stating

- Now attach $[\text{this}]_{\text{hold}}$ to the fuselage.

In this case the accompanying action is not a picking-up action but a short movement of the object in the robot's hand, as in the "wiggle" action in the third system turn in Figure 6. In this case, other forms of multimodal reference such as pointing or eye gaze are not available, so the robot must use a form of haptic-ostensive reference to direct the user's attention.

4.3 Implementation Details

Incorporating such haptic-ostensive references into the output of the JAST human-robot system is currently in the final stages of completion. The implementation task is made easier by the modular architecture of the system, in which all references to world objects are planned by a separate, dedicated *reference generator* module, which takes as input the set of world objects to refer to and returns a specification of the reference type to use, drawing from information about the current state of the dialogue, the locations of all objects of the world, the current stage of the construction task, and the context in which the reference should be made.

The reference-generator module fits into the pipeline-style output-processing system illustrated in Figure 7. The *decision maker* processes input from the user and, using all of

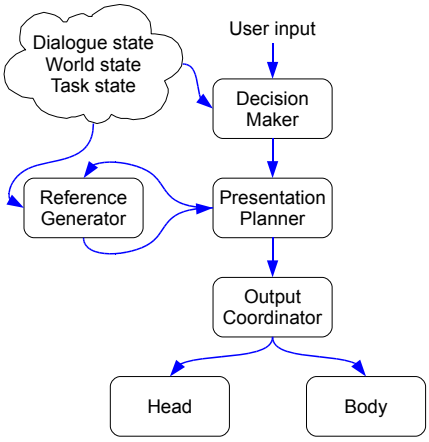

Figure 7: The JAST output-generation system

the available state information, selects an appropriate high-level system response. The *presentation planner* then develops this high-level response specification into a set of commands for each of the output channels (the talking head and the robot arms). It is the presentation planner that calls the reference generator to decide how to realise any object references that are necessary to realise the high-level response specification. The fully-formed output plan is sent to the *output coordinator*, which translates it into concrete plans for the talking head and the robot arms and manages the execution of the plans to ensure that output is coordinated temporally and spatially. Crucially, reference generation takes place as part of multimodal output planning, so the module is able to select coordinated verbal and non-verbal actions to realise a reference; that is, the reference generator can actually select actions to change the state of the world (such as picking up objects) if they are necessary for the selected reference type.

5. CONCLUSIONS AND FUTURE WORK

We have summarised current approaches to the generation of multimodal referring expressions, describing the types of contextual information that are used and the range of multimodal output that is produced. We have then concentrated on particular type of interaction—cooperative task-based dialogue in a shared workspace—and shown that in this context, *haptic-ostensive* references in which an object in the world is manipulated are common, particularly in conjunction with deictic linguistic content. This type of referring expression takes advantage of manipulations that must be performed to support the common task to increase the accessibility of the targets of referring expressions, and therefore to permit less elaborate and more natural linguistic references. In addition to demonstrating that such references are common in human-human dialogues based on tasks in a virtual world, we have also listed a number of ways in which these reference can play a role in the output of a humanoid robot designed to co-operate with a human on mutual assembly tasks, coordinating actions with speech and gesture.

Work in this area will continue along several lines: continued exploration of the corpus of human-human dialogues, evaluation of the generated haptic-ostensive references in the context of the human-robot system, and integrating an understanding of this type of reference into the input-processing components of the system. We discuss each of these lines in more detail.

The analysis presented in Section 3 concentrated on one particular aspect of the corpus data: the use of haptic-ostensive references for initial mentions of world objects in spoken dialogue. The corpus data contains many other features which can be used to explore other aspects of multimodal reference in this domain. It would be interesting to study non-initial references as well as initial references to see whether the behaviour patterns are similar. As well, the corpus also contains also a number of trials in which the participants were forbidden from speaking; it would be informative to study how participants in such a condition managed to coordinate their actions and whether the actions they used in these conditions are different than those used when accompanied by speech. We also have eye-tracking data from all of the dyads, and we are currently studying the relationship between dialogue phenomena and the cross-recurrence of the eye tracks [16].

We intend to run a full-scale user evaluation of the human-robot dialogue system in the near future, using metrics such as user satisfaction, task success, and dialogue efficiency to compare the quality of the system under a range of different configurations. As described in Section 4.3, we are currently integrating an enhanced referring-expression generation component into the system, and we hope that one of the comparisons in the user study will be between a system that uses haptic-ostensive reference at appropriate times as described in Section 4.2 and one that does not.

Finally, we also believe that the system should be able not only to generate this type of task-based embodied reference, but also to understand them when the user produces them. This will require that the visual sensors correctly recognise gestures such as picking up and object and holding it towards the robot, and also that the multimodal input-processing system correctly combine such gestures with spoken content to produce messages. Properly understanding multimodal referring expressions is a task that is at least as complex as correctly generating them—e.g., [5, 10]—but for a system to support true collaborative dialogue, it must be able to deal fully with such expressions.

6. ACKNOWLEDGEMENTS

Thanks to Jan Peter de Ruiter for useful discussions, and to Craig Nicol for invaluable help in querying the JCT corpus.

7. REFERENCES

[1] M. Ariel. The function of accessibility in a theory of grammar. *Journal of Pragmatics*, 16(5):443–463, 1991. doi:10.1016/0378-2166(91)90136-L.

[2] E. G. Bard and M. P. Aylett. Referential form, word duration, and modeling the listener in spoken dialogue. In J. C. Trueswell and M. K. Tanenhaus, editors, *Approaches to Studying World-Situated Language Use: Bridging the Language-as-Product and Language-as-Action Traditions*. The MIT Press, 2004.

[3] A. Belz, A. Gatt, E. Reiter, and J. Viethen, editors. *The Attribute Selection for Generation of Referring Expressions Challenge*, 2007. `http://www.csd.abdn.ac.uk/research/evaluation/`.

[4] R.-J. Beun and A. Cremers. Object reference in a shared domain of conversation. *Pragmatics and Cognition*, 6(1–2):121–152, 1998.

[5] D. K. Byron. Understanding referring expressions in situated language: Some challenges for real-world agents. In *Proceedings of the First International Workshop on Language Understanding and Agents for Real World Interaction*, 2003.

[6] J. Carletta, C. Nicol, T. Taylor, R. L. Hill, J. P. de Ruiter, and E. G. Bard. Eye-tracking for two-person tasks with manipulation of a virtual world. *Behavior Research Methods*, under revision.

[7] R. Dale and E. Reiter. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(2):233–263, 1995. doi:10.1207/s15516709cog1902_3.

[8] M. E. Foster, T. By, M. Rickert, and A. Knoll. Human-robot dialogue for joint construction tasks. In *ICMI '06: Proceedings of the 8th international conference on Multimodal interfaces*, pages 68–71, Banff, Alberta, Canada, November 2006. doi:10.1145/1180995.1181009.

[9] M. E. Foster and C. Matheson. Representing and using assembly plans in cooperative, task-based human-robot dialogue. 2008. In submission.

[10] D. Gergle, C. P. Rosé, and R. E. Kraut. Modeling the impact of shared visual information on collaborative reference. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 1543–1552, 2007. doi:10.1145/1240624.1240858.

[11] J. D. Kelleher and G.-J. M. Kruijff. Incremental generation of spatial referring expressions in situated dialog. In *Proceedings of the 44th annual meeting of the ACL (COLING-ACL 2006)*, pages 1041–1048, 2006. doi:10.3115/1220175.1220306.

[12] A. Kranstedt, A. Lücking, T. Pfeiffer, H. Rieser, and I. Wachsmuth. Deictic object reference in task-oriented dialogue. In G. Rickheit and I. Wachsmuth, editors, *Situated Communication*, pages 155–207. Mouton de Gruyter, 2006.

[13] A. Kranstedt and I. Wachsmuth. Incremental generation of multimodal deixis referring to objects. In *Proceedings of the 10th European Workshop on Natural Language Generation (ENLG-05)*, pages 75–82, Aberdeen, Scotland, August 2005.

[14] F. Landragin, N. Bellalem, and L. Romary. Referring to objects with spoken and haptic modalities. In *Proceedings of the Fourth IEEE International Conference on Multimodal Interfaces*, pages 99–104, 2002. doi:10.1109/ICMI.2002.1166976.

[15] P. Piwek. Modality choice for generation of referring acts: Pointing versus describing. In *Proceedings of the Workshop on Multimodal Output Generation (MOG 2007)*, Aberdeen, Scotland, 2007.

[16] D. C. Richardson, R. Dale, and N. Z. Kirkham. The art of conversation is coordination: common ground and the coupling of eye movements during dialogue. *Psychological Science*, 18(5):407–413, May 2007. doi:10.1111/j.1467-9280.2007.01914.x.

[17] M. Rickert, M. E. Foster, M. Giuliani, T. By, G. Panin, and A. Knoll. Integrating language, vision and action for human robot dialog systems. In *Proceedings of HCI International 2007*, Beijing, China, July 2007. doi:10.1007/978-3-540-73281-5_108.

[18] A. J. N. van Breemen. iCat: Experimenting with animabotics. In *Proceedings of the AISB 2005 Creative Robotics Symposium*, 2005.

[19] K. van Deemter, I. van der Sluis, and A. Gatt. Building a semantically transparent corpus for the generation of referring expressions. In *Proceedings of the 4th International Conference on Natural Language Generation (INLG)*, Sydney, Australia, 2006.

[20] I. F. van der Sluis. *Multimodal Reference: Studies in Automatic Generation of Multimodal Referring Expressions*. PhD thesis, University of Tilburg, 2005.