



Foster, M. E., Giuliani, M., and Isard, A. (2014) Task-based evaluation of context-sensitive referring expressions in human-robot dialogue. *Language, Cognition and Neuroscience*, 29(8), pp. 1018-1034.

There may be differences between this version and the published version. You are advised to consult the publisher's version if you wish to cite from it.

<http://eprints.gla.ac.uk/135661/>

Deposited on: 30 January 2017

Enlighten – Research publications by members of the University of Glasgow
<http://eprints.gla.ac.uk>

Task-based evaluation of context-sensitive referring expressions in human-robot dialogue

Mary Ellen Foster* and Manuel Giuliani

Technische Universität München

Amy Isard

University of Edinburgh

Mary Ellen Foster is now at the School of Mathematical and Computer Sciences, Heriot-Watt University. Manuel Giuliani is now at fortiss GmbH.

*Corresponding author. Email: M.E.Foster@hw.ac.uk.

Author Note

This article integrates and extends the work described in the following conference papers: (Foster et al., 2008; Foster, Giuliani, Isard, et al., 2009; Foster, Giuliani, & Knoll, 2009; Giuliani et al., 2010). The research leading to these results has received funding from the European Union's Sixth Framework Programme (FP6/2003-2006) under grant agreements no. 003747, JAST: Joint Action Science and Technology and no. 045388, INDIGO: Interaction with Personality and Dialogue Enabled Robots, and from the Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 270435, JAMES: Joint Action for Multimodal Embodied Social Systems.

Abstract

The standard referring-expression generation task involves creating stand-alone descriptions intended solely to distinguish a target object from its context. However, when an artificial system refers to objects in the course of interactive, embodied dialogue with a human partner, this is a very different setting: the references found in situated dialogue are able to take into account aspects of the physical, interactive, and task-level context, and are therefore unlike those found in corpora of stand-alone references. Also, the dominant method of evaluating generated references involves measuring corpus similarity. In an interactive context, though, other extrinsic measures such as task success and user preference are more relevant—and numerous studies have repeatedly found little or no correlation between such extrinsic metrics and the predictions of commonly used corpus-similarity metrics.

To explore these issues, we introduce a humanoid robot designed to co-operate with a human partner on a joint construction task. We then describe the context-sensitive reference-generation algorithm that was implemented for use on this robot, which was inspired by the referring phenomena found in the Joint Communication Task corpus of human-human joint construction dialogues. The context-sensitive algorithm was evaluated through two user studies comparing it to a baseline algorithm, using a combination of objective performance measures and subjective user satisfaction scores. In both studies, the objective task performance and dialogue quality were found to be the same for both versions of the system; however, in both cases, the context-sensitive system scored more highly on subjective measures of interaction quality.

Keywords: referring expressions in interactive settings; task-based evaluation; human-robot dialogue

Task-based evaluation of context-sensitive referring expressions in human-robot dialogue

The generation of referring expressions (GRE) is one of the most clearly defined sub-tasks in natural language generation (NLG), and is therefore one of the tasks that has received the most attention. The classic GRE task involves creating an initial, stand-alone description intended solely to distinguish the target from any “distractors” in the area, and the dominant method of evaluating such systems involves measuring the similarity of the generated references to those drawn from a suitable corpus of human-generated descriptions.

In this paper, we consider referring expressions in the context of joint action in a shared workspace, where the dialogue takes place between a human and a humanoid robot. The robot was designed to co-operate with a human partner on a joint construction task, so the referring expressions which it generates take account of the discourse and physical context in which it operates, drawing inspiration from the referring phenomena found in a corpus of human-human dialogues in a similar joint construction scenario.

In this interactive, embodied context, the most important measures of success are extrinsic measures such as task success and user subjective opinions—and it is known that the predictions of the typical intrinsic corpus-based evaluation strategies do not tend to correlate with those of any extrinsic measures. We therefore carried out a task-based evaluation of our context-sensitive algorithm, comparing it with a baseline algorithm in two user studies using a combination of objective performance measures and subjective user satisfaction measures. In both studies, the objective task performance and dialogue quality were found to be the same for both versions of the system; however, in both cases, the context-sensitive system tended to score more highly on subjective measures of the robot’s quality as a conversational partner.

Background

The work described in this paper draws on research and techniques from three main areas: the use of different reference types in human-human dialogue, the automatic generation of referring expressions, and the evaluation of automatically generated output.

Reference types in human-human dialogue

The question of how speakers select appropriate referring expressions in human-human dialogue has been extensively studied in linguistics, and several models have been proposed. Accessibility models (Ariel, 1991) assume that mental representations have varying levels of accessibility to addressees, and that speakers choose among referring expressions to mark these differences. For example, fully specified indefinite descriptions correspond to low accessibility entities—that is, entities that are deemed to be completely unfamiliar to the audience—while definite descriptions, deictic expressions, and pronouns correspond to increasing levels of assumed accessibility.

A similar model is the Givenness Hierarchy of Gundel, Hedberg, and Zacharski (1993), which also assumes that different determiners and pronominal forms conventionally signal different cognitive statuses in the mind of the addressee. The hierarchy ranges from *In Focus* (the highest status, for which a pronoun such as “it” is appropriate) through *Type Identifiable* (the lowest status, for which the only possible reference type is an indefinite noun phrase such as “a *N*”). Every status in the hierarchy entails all lower statuses, meaning that a referent with a particular status can be referred to with the form appropriate to that status or with any lower form. However, pragmatic concerns such as the Maxim of Quantity (Grice, 1975) constrain the reference types, since using a lower-status reference for a high-status object may have unwanted conversational implicatures.

Automatically generating referring expressions

The classic algorithm in GRE—and the one on which most subsequent implementations are based—is the well-known *incremental algorithm* of Dale and Reiter (1995), 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, using a predefined, domain-specific preference ordering to ensure that the most relevant attributes are always included. Each selected attribute is then added to the referring expression,

and the objects without the attribute are removed from the distractor set. This process is executed repeatedly until only the target object remains in the distractor set. This algorithm was inspired by psycholinguistic findings on how humans tend to refer to objects.

The basic incremental algorithm makes several assumptions: (a) 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, (b) that the only goal of the interaction is to indicate the target object to the hearer, and (c) that the speaker has a complete model of the listener's knowledge. Under these assumptions, the only way to refer to the target successfully is to create a fully-specified linguistic expression.

However, as noted by Krahmer (2010), among others, there is ample evidence from psycholinguistics that all of these assumptions are unrealistic when it comes to real-world reference (cf. Clark & Bangerter, 2004). It is known, for example, that speakers tend to adapt to their conversational partners (Clark & Wilkes-Gibbs, 1986; Garrod & Pickering, 2009), and in the preceding section we have presented two well-known models of how this adaptation is based on cognitive status. However, the extent to which speakers are able to model the intended recipient fully is not entirely clear (e.g., Bard & Aylett, 2004; Horton & Keysar, 1996). Also, the physical and discourse context in which the reference takes place has a significant effect on its features: for example, when people work together on a joint task, the referring expressions that they use appear to take into account the task context (Bard, Hill, & Foster, 2008). Linguistic and cultural effects also have an impact—e.g., van der Sluis and Luz (2011) found many significant differences among the attributes selected by individuals from different linguistic groups.

Since the initial description of the incremental algorithm, a number of people have therefore 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. Krahmer and Theune (2002) implemented an extension which takes previous context into account, and found that human subjects preferred the extended version when shown written texts. Kelleher and Kruijff (2006) implemented an algorithm to generate linguistic spatial referring

expressions in situated dialogue, and 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.

Other studies have added the ability to include non-verbal behaviours into the specification of the referring expressions. Van der Sluis (2005), for example, presented a graph-based algorithm that creates multimodal referring acts including deictic pointing by assigning costs to the verbal and non-verbal components of referring expressions and then selecting the combination with minimum cost. Similarly, Kranstedt and Wachsmuth (2005) extended the incremental algorithm by specifying two types of pointing, *object-pointing* and *region-pointing*, and gathered data from empirical studies (Kranstedt, Lücking, Pfeiffer, Rieser, & Wachsmuth, 2006) to determine the normal use of pointing in multimodal reference. Staudte and Crocker (2009) investigated the relationship between gaze and referring expressions in human-robot dialogues, and suggest that gaze is a major factor in the salience of objects in a visual scene, and therefore can be integrated into the production of referring expressions in systems where robot gaze is possible.

Evaluating generated output

Evaluating the quality of a generation system is known to be a difficult task: as pointed out by Mellish and Dale (1998), the issues include defining the input and output, choosing what to measure, selecting a control or baseline for comparison, obtaining adequate training or test data, and dealing with disagreement of human judges. All of these problems are more serious than the corresponding problems in evaluating, for example, a natural-language understanding system: generation is a more open-ended task, so the criteria for success are therefore more difficult to define.

Taking into account these difficulties, a range of techniques have been used to examine the quality of generated output. The most technically demanding form of evaluation—but also in many ways the most convincing—is a task-based, comparative study: that is, demonstrating that a

complete system achieves its goals significantly better when the advanced generation components are enabled than when they are not (Reiter, 2011). This sort of study has been used to show, for example, that tailoring descriptions to a user's preferences affected their selection behaviour (Carenini & Moore, 2006), and that users learned more from object descriptions that employed aggregation (Karasimos & Isard, 2004).

An alternative, somewhat less demanding form of human evaluation is to elicit subjective opinions: asking human judges either to assess the quality of the generated output directly, or to respond to a questionnaire asking for their subjective experience of the whole system. This technique may be employed alongside a task-based evaluation, as in the two studies mentioned above; it may also be used on its own, as was done for example by Binsted, Pain, and Ritchie (1997), who asked children to evaluate generated jokes, and Belz and Reiter (2006), who had experts and laypeople judge the quality of generated weather forecasts. Note that, while there is often a correlation, users' subjective judgements of an interaction do not always agree with task performance (e.g., Nielsen & Levy, 1994; Oviatt, 1999).

In addition to the human-based techniques listed above, another popular form of evaluation is to assess the quality of generated output directly using various automated metrics. The most common form of automated evaluation makes use of a corpus of target outputs: the generated output is then compared against the corpus using techniques such as cross-validation (e.g., White, 2004). This style of evaluation can be used to test whether a proposed model behaves like the human speakers it is intended to imitate; it can also be used to predict the interactive performance of a generation system without needing to recruit participants for a task-based or subjective-judgement study. In the former case, it suffices to compare the generated output against the corpus; however, in the latter case, it is necessary to find metrics that not only can be computed automatically, but that also correlate with human judgements of quality.

The danger in relying purely on corpus-based measures in an interactive setting is that they are known to penalise output that differs in any way from the exact corpus examples; this tends to favour "average" outputs that do not make use of the full range of generation possibilities, and

that are often not preferred by actual users in practice. For example, Belz and Reiter (2006) compared a set of NLG systems using a wide range of intrinsic and extrinsic metrics: they found that the correlation was high within each of the two classes of metrics, but that there was almost no relationship between the two classes. Similarly, when Foster (2008) compared methods for selecting the non-verbal behaviour for an automated talking head, the majority-selection method (which always chose the most frequent option) was strongly disliked by the human judges, but scored the highest on all of the corpus-similarity metrics.

Evaluating reference generation. When it comes to the evaluation of reference-generation systems, most studies have concentrated on maximising the human-likeness of the generated references (see Krahmer & van Deemter, 2012), most often in a simplified domain where the shared information between speaker and hearer is small, and successful reference is the primary—or only—goal; this mirrors the underlying assumptions of the incremental algorithm mentioned above. For example, corpus-based similarity was an important criterion for two recent shared-task evaluation challenges in the area of reference generation: the TUNA challenges (Gatt & Belz, 2010), which required systems to generate stand-alone descriptions of individual furniture items or people from a scene containing multiple such entities, and the GREC challenge (Belz, Kow, Viethen, & Gatt, 2010), which addressed the task of generating appropriate references in the context of Wikipedia articles. Several recent evaluations of GRE systems have employed similar corpus-based measures of human-likeness (e.g., van der Sluis & Luz, 2011; Viethen, Dale, & Guhe, 2011). The classic Dale and Reiter (1995) incremental algorithm has itself also been recently examined in a corpus-based study (van Deemter, Gatt, van der Sluis, & Power, 2012) that compared its outputs with those produced by the humans who produced the references in the TUNA corpus (van Deemter, van der Sluis, & Gatt, 2006). This study found that the human-likeness of the output of the incremental algorithm depends crucially on the selection of the domain-dependent preference order, which is difficult to specify for any particular domain.

Just as in the general case of NLG evaluation, the findings of corpus-based studies of GRE

do not tend to bear any relation to the predictions of measures such as task performance or subjective judgements. For example, in addition to measuring corpus similarity, both the TUNA and the GREC challenges also asked humans directly to judge the quality of a subset of the generated outputs; and in both cases, no significant correlation was found between the corpus-based results and the human judgements.

Some GRE evaluation studies have employed other, extrinsic metrics. For example, another series of shared-task evaluation challenges—the GIVE challenges (Koller et al., 2010)—have provided what amounts to an indirect, task-based evaluation of reference generation. In those challenges, users navigate through and interact with a virtual environment, following instructions generated by one of several competing NLG systems. An important aspect of the GIVE task is to press a particular sequence of buttons on the walls of various rooms, and to avoid pressing other buttons which might either do nothing, or else reset the lock on the safe to be opened and require the user to start again. Reference generation is therefore a core task in the GIVE domain, but one which is only indirectly evaluated by the evaluation measures used. In a related study, Koller, Staudte, Garoufi, and Crocker (2012) evaluated a GRE system in the GIVE domain that used an eye tracker to monitor and respond to the user’s focus of attention, and found that the eye-tracking system significantly outperformed two baselines on several measures of task success.

Campana, Tanenhaus, Allen, and Remington (2011) have recently carried out a fully extrinsic evaluation of GRE in which they compared two algorithms: a NATURAL version that took discourse context into account, and a STANDARDIZED version that generated referring expressions that were consistent across all conditions. While both systems would use fully-specified references for all initial mentions (e.g., “the big red triangle”), the NATURAL version would use reduced references such as “the triangle” or “it” for subsequent mentions, while the STANDARDIZED version would continue with the full references. The two algorithms were compared using the dual-task paradigm, which is a general method for investigating how much of a cognitive resource is consumed by a given task. The primary task in the study was to move objects around on screen following the spoken directions of the system, while the secondary

task was to detect flickering lights. While the primary task performance was the same across the conditions, the subjects' accuracy and reaction time on the secondary task were better in the NATURAL condition, indicating that the context-sensitive references reduced cognitive load.

References in the context of joint action in a shared workspace

The goals of the JAST project (“**J**oint **A**ction **S**cience and **T**echnology”) were to investigate the cognitive, neural, and communicative aspects of jointly-acting agents, and to build jointly-acting autonomous systems that communicate and work intelligently on mutual tasks. As part of this project, a corpus of human-human dialogues was gathered in the domain of joint construction, and a robot system was also developed that supported similar joint construction tasks with a human partner.

In this section, we first explore the referring expressions that were found in the corpus of task-based human-human dialogues where the partners work together 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, and—in particular—the participants are able to bring entities into focus by manipulating them as part of the joint task. The distribution of initial mentions in this corpus differ from what would be theoretically expected, indicating that humans do take this broader notion of context into account when referring in this domain.

We then introduce the JAST human-robot dialogue system, which was designed to support similar joint construction tasks together with a human partner. We first give an overview of the system and the scenarios that it supports; we then give a detailed description of the context-sensitive reference-generation algorithm that was implemented for use on this robot system, including examples of its output.

Referring expressions in human-human joint construction

The Joint Construction Task (JCT) corpus (Bard et al., 2008), which was gathered as part of the JAST project, is based on the recording and analysis of humans cooperating with one another, utilising a novel experimental paradigm based around a two-person shared virtual environment (Carletta et al., 2010). The objective was to collaboratively build tangram models from a set of geometrical components, doing so as efficiently and as accurately as possible. To stimulate a range of referring expressions, duplicates were included of most components. The two subjects were present in the same room, using separate computers. A subject could not see their partner's face, but could hear their speech and see their actions in the virtual world. Depending on the experimental condition, the partner's mouse and/or gaze location were sometimes also visible on the screen.

Each linguistic referring expression in the JCT corpus was annotated with its referent in the world and its degree of accessibility (Ariel, 1991), using a similar scheme to that employed by Bard and Aylett (2004). Table 1 (adapted from Bard et al. (2008)) shows the distribution of initial mentions in the JCT corpus across the accessibility levels, ranging from indefinite noun phrases (the most elaborate expressions, indicating the lowest accessibility) through other forms of noun phrases, ending with various types of pronouns.

[Table 1 about here.]

Since nearly all of the tangram parts come in identical pairs, an initial reference to any object would theoretically be expected to be an indefinite expression such as “a purple triangle” or “one of the pink squares” (Bard et al., 2008). However, as shown in Table 1, the actual referring behaviour of the corpus speakers differs markedly from this theoretical prediction: only 17% of the first mentions were indefinite, with the remaining mentions distributed across other categories including definite NPs (“the red bit”), deictic expressions (“this green triangle”), and pronouns such as “this” and “it.” This suggests that the speakers often considered objects to be highly accessible even before they are mentioned.

A particular feature of reference in such a collaborative workspace is that an entity can be brought into focus (in the centering theory sense described by Grosz, Weinstein, and Joshi (1995)) by the speaker manipulating it as part of the joint task. Indeed, an analysis of the JCT mentions found that 36% of the initial mentions in the JCT were accompanied by a concurrent mouse manipulation, with the percentage rising to 54% for initial deictic references (Foster et al., 2008). Note that there was no relationship between referring behaviour and task performance: very accurate tangrams were built in all conditions, and the main factor affecting performance was the presence or absence of the cross-projected mouse cursor.

Human-robot joint construction

The results described above demonstrate that, in situated dialogue in a joint workspace, the standard assumptions about GRE do not hold: speakers make use of much more of the task, physical, and interaction contexts when deciding how to refer to an entity, even when making an initial reference. In other words, in terms of accessibility models (Ariel, 1991), the speakers are clearly assuming a higher degree of accessibility on the part of the addressee than would be expected in theory.

[Figure 1 about here.]

Based on those findings, we have implemented a context-sensitive GRE algorithm for use in an artificial agent designed to work together with a human partner on a similar joint construction task. The artificial agent we used was the JAST humanoid robot (Figure 1), consisting of a torso, a pair of manipulator arms with grippers, mounted in a position to resemble human arms, along with an animatronic talking head (van Breemen, Yan, & Meerbeek, 2005) capable of producing facial expressions, rigid head motion, and lip-synchronised synthesised speech. The user and the robot work together to assemble wooden construction toys on a common workspace, coordinating their actions through speech, gestures, and facial expressions.

The robot is able to pick up and move objects in the workspace and perform simple assembly tasks, but relies on the user to do more complex manipulations. To make joint action

necessary for success in the assembly task, the workspace is divided into two areas—one belonging to the robot and one to the user—so that the robot must hand over some pieces to the user to complete the construction task.

[Figure 2 about here.]

[Figure 3 about here.]

The robot supports two different interaction scenarios. In Scenario 1, the task knowledge is asymmetric: only the robot knows the assembly plan for a particular compound object. It instructs the user on carrying out the plan, explaining the necessary assembly steps and retrieving pieces as required, with the user performing the actual assembly actions. A sample dialogue from Scenario 1 is shown in Figure 2. In Scenario 2, on the other hand, the knowledge of the two participants is symmetric: both the robot and the user know the assembly plan and jointly execute it. As an additional feature in Scenario 2, the user's knowledge is not assumed to be correct: rather, the user may have been given an incorrect plan for building the target objects. In that case, the robot must detect and correct any incorrect actions. When interacting in this scenario, the robot therefore monitors the user's actions, offering pieces as required, and detects and responds to any errors in carrying out the plan; this scenario is illustrated in Figure 3.

The robot system incorporates components which use both sub-symbolic and symbolic processing. It includes a goal inference module based on dynamic neural fields (Bicho, Erlhagen, Louro, & Costa e Silva, 2011; Bicho, Louro, & Erlhagen, 2010), which is able to select the robot's next actions based on the human user's actions and utterances. Given a particular assembly plan and the knowledge of which objects the user has picked up, this module can also determine when the user has made an error. The system also incorporates a dialogue manager based on the TrindiKit dialogue management toolkit (Larsson & Traum, 2000), which implements the information-state based approach to dialogue management. Messages from all of the system's input channels—speech, object recognition, and gesture recognition—are processed and combined by a multimodal fusion component (Giuliani & Knoll, 2008), which is the link

between the symbolic and the sub-symbolic parts of the system. The fusion component then communicates with the goal inference module, which calculates the next action instructions for the robot and also determines if the user made an error. From there, fusion combines the information from goal inference with the input data and sends unified hypotheses to the dialogue manager.

When it receives input hypotheses from the fusion system, the dialogue manager uses the dialogue history along with the physical and task context to choose an appropriate system response, and sends a high-level specification of the desired response to the presentation planner. The presentation planner then develops this 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 necessary object references. Finally, 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 also 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 take into account the concurrent robot actions (such as picking up objects) when selecting a reference type. Also, since the target of all generated references is known at plan time, the presentation planner can add multimodal behaviours such as looking at a target object at the same time as a spoken reference is made, which further enhances the multimodal reference process.

Context-sensitive reference generation in JAST

Two strategies were implemented in the JAST human-robot system for generating references to objects in the world: a **basic** version that uses only the standard incremental algorithm (Dale & Reiter, 1995) to select properties, and a **context-sensitive** version that uses more of the physical, dialogue and task context to help select the references, resulting in a wider

range of references that resemble the phenomena found in the human-human JCT dialogues. The basic algorithm can produce a definite or indefinite reference, using the most appropriate combination of attributes according to the incremental algorithm. The context-sensitive algorithm also generates pronominal and deictic references, in some contexts where the basic algorithm would always produce a definite reference.

The details of the context-sensitive algorithm were inspired by the findings from the JCT data in that a wide range of reference types are permitted, and also in that the physical context is incorporated into the reference-selection process. The particular details of the algorithm were not directly based on the JCT data, as the context of the JCT corpus—while similar to the human-robot context at a high level—is sufficiently different that the results could not be directly applied. In particular, the JCT dialogues are much more symmetrical, both at the task level and at the linguistic level, meaning that the task contexts were more complex and the possible references more varied.

The context-sensitive algorithm is similar to the other extensions to the Incremental Algorithm mentioned above, as well as to the “NATURAL” algorithm studied by Campana et al. (2011). One novel aspect is in its use of multiple distractor sets depending on the circumstances under which the reference was being made. We also draw inspiration from centering theory (Grosz et al., 1995) when determining the appropriate referring expression type, with the goal of maintaining the coherence of the discourse.

In the sample interaction in Figure 3, the user picks up an incorrect piece, and the robot detects the error and describes the correct assembly procedure. The underlined references show the range of output produced by the context-sensitive reference generation module; for the basic system, the references would all have been “the red cube.”

Context-sensitive reference algorithm. Our reference generation algorithm uses the dialogue reference history and distractor sets to choose the most appropriate reference. The algorithm works at a language-independent level: only the final surface-realisation step makes use of a language-specific grammar to generate text in either German or English. The grammar is

defined in OpenCCG (White, 2006), an open-source implementation of Combinatory Categorical Grammar (Steedman, 2000)—a unification-based categorial framework which is both linguistically and computationally attractive.

We introduce the idea of using different distractor sets depending on the circumstances in which the current object is being referred to. We use a two-step process to create a referential expression:

- First, we process the distractor set and select which properties (if any) of the current object to include in the reference.
- We then use the dialogue history and position of the current object to choose the reference type.

We will describe each of these steps in turn, and then work through two examples from Figure 3 to illustrate the algorithms.

Choosing properties to include. In this domain there are two types of objects which we need to refer to: concrete objects in the world (everything which is on the table, or in the robot’s or user’s hand), and objects which do not yet exist, but are in the process of being created as part of the collaborative task. Non-existent objects have an empty distractor set. For concrete objects, we have defined three different types of distractor set:

1. All the concrete objects in the world. This is used if the current object has not been mentioned before.
2. All the objects referred to since the last mention of the current object. This is used if the current object has previously been mentioned during this dialogue.
3. All the pieces needed to build a target object. This is a special case which is used if the current object is a construction piece being used to create a target object, and is being referred to in a negative statement, e.g. “You don’t need a large slat to build a railway signal.”

Choosing a reference type. When choosing a referring expression, we first process the distractor set, comparing the properties of the current object with the properties of all distractors. We start with the distractor object type; if a distractor has a different type from the current object,

it is removed from the distractor set. We then process each remaining distractor in turn; for each of its properties, if the distractor has a different value from the current object, the current object's property value is added to the list of properties to use, and if any properties differ, it is removed from the distractor set.

We then choose the type of referring expression. We first check whether the current object exists, and if it does not, or is part of a negative reference, we use an **indefinite reference**.

We then look for the previous reference to the current object. If such a previous reference exists, we also determine whether—in the terminology of centering theory (Grosz et al., 1995)—that reference was the *focus* of the discourse. We then use the following algorithm, illustrated in Table 2, to choose the reference type.

No previous reference.

- If the robot is holding the current object, we use a **deictic reference**.
- If the current object is concrete and there are no distractors, we use a **definite reference**.
- If the current object is concrete and there are distractors we use an **indefinite reference**.

Previous reference was the focus.

- If the previous reference was within the same turn, we use a **pronoun**.
- If the previous reference was in an earlier turn and the robot is holding the current object we use a **deictic reference**

- If the previous reference was in an earlier turn and the robot is not holding the current object, we use a **pronoun**.

Previous reference was not the focus.

- If the robot is holding the current object, we make a **deictic reference**
- if the previous reference was a pronoun, definite, or deictic, we use a **definite reference**.
- If the previous reference was indefinite and there are no distractors, we use a **definite reference**
- if there are distractors, we use an **indefinite reference**.

If there are any properties in the list, and the reference which has been chosen is not a

pronoun, we add them.

[Table 2 about here.]

Examples of the reference algorithm. We will illustrate the reference-selection strategy with two cases from the dialogue in Figure 3.

Utterance 4 “a yellow cube”. This object is going to be referred to in a negative context as part of a windmill under construction, so the distractor set is the set of objects needed to make a windmill: {red cube, blue cube, small slat, small slat, green bolt, red bolt}.

We select the properties to use in describing the object under consideration, processing the distractor set. We first remove all objects which do not share the same type as our object under consideration, which leaves {red cube, blue cube}. We then compare the other attributes of our new object with the remaining distractors—in this case “colour.” Since neither cube shares the colour “yellow” with the target object, both are removed from the distractor set, and “yellow” is added to the list of properties to use.

There is no previous reference to this object, and since we are making a negative reference, we automatically choose an indefinite article. We therefore select the reference “a yellow cube.”

Utterance 6 “it” (a green bolt). This object has been referred to before, earlier in the same utterance, so the distractor set is all the references between the earlier one and this one—{red cube}. Since this object has a different type from the bolt we want to describe, the distractor set is now empty, and nothing is added to the list of properties to use.

There is a previous definite reference to the object in the same utterance: “the green bolt.” This reference was focal, so we are free to use a pronoun if appropriate. Since the previous reference was definite, and the object being referred to does exist, we choose to use a pronoun. We therefore select the reference “it.”

Experiments: Evaluating reference strategies in context

In the context of the scenarios supported by the JAST robot, a basic reference strategy—which always chooses the same reference regardless of context—is sufficient in that it

always makes it possible for the robot's partner to know which item is needed. On the other hand, the varied forms produced by a more complex, context-sensitive mechanism may increase the naturalness of the system output. It is known that people respond well to reduced expressions like "this cube" or "it" when they are used by another person (Bard et al., 2008) or by a speech system (Campana et al., 2011); we need to see if these benefits also apply to a physically embodied robot system.

To address this question, the human-robot dialogue system was evaluated through a pair of user studies in which participants interacted with the complete system. Using a between-participants design, the studies both compared the two reference strategies, measuring the participants' subjective reactions to the system along with a range of objective measures of dialogue efficiency, dialogue quality, and task success. These studies therefore provide a task-based evaluation of the success of the two reference strategies, along with an indication of the participants' opinions of both of the strategies. The first study addressed Scenario 1, where the robot instructs its partner as illustrated in Figure 2; the second study concentrated on Scenario 2, which adds the error-monitoring and correction features as in Figure 3.

Participants

[Table 3 about here.]

The details of the participants in the two studies are shown in Table 3. In addition to gathering basic demographic information, we also asked the participants to rate their knowledge of computers, of speech-recognition systems, and of human-robot systems, all on a scale of 1–5. They were also asked to indicate their major area of study (if any); for both experiments, the two most frequent responses were Informatics and Mathematics. None of this demographic information had a significant effect on the study results presented below. Participants were compensated for their participation in both experiments.

Scenarios

[Figure 4 about here.]

The participants in both studies stood in front of the table facing the robot (as in Figure 1), equipped with a headset microphone for speech recognition.

Experiment 1 scenario. Each participant built the same three objects in collaboration with our human-robot interaction system, always in the same order; the interactions were conducted in German. The first target object was a “windmill” (German: *Windmühle*) (Figure 4a), which has a sub-component called a “tower” (*Turm*) (Figure 4b). After the windmill had been completed, the system then described how to build an “L shape” (*Buchstabe L*) (Figure 4c). Finally, the robot instructed the participant on building a “railway signal” (*Bahnsignal*) (Figure 4d), which combines an L shape with a tower. The participant was not given any instructions in advance on how to build any of the target objects, and had to rely entirely on the instructions given by the robot.

Before the system explained each target object, the experimenter configured the workspace with exactly the pieces required to build it. The pieces were always distributed across the two work areas in the same way to ensure that the robot would always hand over the same pieces to each participant. For the windmill, the robot handed over one of the cubes and one of the slats; for the L shape, it handed over both of the required slats; while for the railway signal, it handed over both cubes and both slats.

For objects requiring more than one assembly operation (i.e., all but the L shape), the system gave names to all of the intermediate components as they were built. For example, the windmill was always built by first making a tower and then attaching the slats to the front. When the railway signal was being built, the system always asked the participant if they remembered how to build a tower and an L shape. If they did not remember, the robot explained again; if they did remember, the robot simply asked them to build another one using the pieces on the table.

Experiment 2 scenario. Each participant built two objects in collaboration with the system, always in the same order. The first target object was the windmill (Figure 4a); after the

windmill was completed, the robot and human then jointly built a railway signal (Figure 4d). As in Experiment 1, the interactions took place in German.

For both target objects, the user was given a building plan on paper. To induce an error, both of the plans given to the participants instructed them to use an incorrect piece: a yellow cube instead of a red cube for the windmill, and a long (seven-hole) slat instead of a medium (five-hole) slat for the railway signal. The participants were told in advance that their plan contained an error and that the robot would correct them when necessary, but did not know the nature of the error. When the human picked up or requested an incorrect piece during the interaction, the system would detect the error and explained to the human what to do in order to assemble the target object correctly.

The pieces required for the target object—plus a set of additional pieces in order to make the reference task more complex—were placed on the table, using the same layout for every participant. The layout was chosen to ensure that there would be points in the interaction where the participants had to ask the robot for building pieces from the robot’s workspace, as well as situations in which the robot automatically handed over the pieces.

Independent variables

In both of these studies, we manipulated one independent variable—the reference strategy—which had two possible levels, **basic** and **context-sensitive**. The basic algorithm was an implementation of the Dale and Reiter (1995) incremental algorithm, and in Experiment 1 it also included pronominal references. The context-sensitive algorithm was as described in the preceding section. Participants were assigned to conditions using a between-participants design, so that each participant interacted with the system using a single reference strategy throughout. In Experiment 1, 22 participants encountered the basic strategy and 21 the context-sensitive one; Experiment 2 had 19 participants for basic and 22 for context-sensitive.

Dependent variables

We gathered a wide range of dependent measures: objective measures derived from the system logs and annotated video recordings, as well as subjective measures based on the participants' own ratings of their experience interacting with the system. We also recorded the distribution of references generated by the system under different conditions to allow the output to be compared to the referring behaviour found in the JCT corpus.

Objective measures. We collected a range of objective measures from the log files and videos of the interactions. Like Litman and Pan (2002), we divided our objective measures into three categories based on those used in the PARADISE framework (Walker, Litman, Kamm, & Abella, 1997): dialogue efficiency, dialogue quality, and task success.

For Experiment 1, we collected the following measures.

- Three **dialogue efficiency** measures: the mean duration of the interaction in seconds and in system turns, and the mean time taken by the system to respond to the participant's requests.
- Four **dialogue quality** measures: the number of times that the user asked for instructions to be repeated, the number of times that the participant failed to take an object that the robot attempted to hand over, the number of times that the participant looked at the robot, and the percentage of the total interaction that they spent looking at the robot. We considered the gaze-based measures to be measures of dialogue quality since participants tend to look at their partner more often when there is a problem in a physical task-based interaction (Argyle & Graham, 1976).
- Two **task success** measures: how many of the (two) target objects were constructed as intended, judged by the video recordings, and whether the participants learned how to construct the tower and L-shape sub-components, judged by whether they said *yes* or *no* when they were asked if they remembered these during the construction of the railway signal.

For Experiment 2, we collected a reduced set of measures, leaving out the learning of sub-components, which was not part of the second scenario, and some of the measures which did not produce significant results in Experiment 1 and which required significant video analysis

effort. The measures were as follows:

- Two **dialogue efficiency** measures: the mean duration of the interaction as measured both in seconds and in system turns;
- Two **dialogue quality** measures: the number of times that the robot gave explanations, and the number of times that the user asked for instructions to be repeated; and
- One **task success** measure: how many of the (two) target objects were constructed as intended (i.e., as shown in Figure 4).

Subjective measures. In addition to the above objective measures, we gathered a range of subjective measures through a questionnaire which the participants filled out after the interaction.

The questionnaire for Experiment 1 was based on the one used in the user evaluation of the COMIC dialogue system (White, Foster, Oberlander, & Brown, 2005), with modifications to address specific aspects of the human-robot dialogue system and the experimental manipulation in this study. There were 47 items in total, each of which requested that the participant choose their level of agreement with a given statement on a five-point Likert scale.

The items on the second questionnaire were based on those used in Experiment 1, but were adapted for the features of Scenario 2: the number of questions was reduced, and specific questions were added addressing the robot's behaviour when the user made an error. Note that Experiment 2 formed part of a larger study (Bard et al., 2009) where the emphasis was on error detection and recovery, and the same usability questionnaire was also used in an evaluation of another human-robot system performing a similar task (Bicho et al., 2010), so a number of items specific to the Experiment 1 scenario were removed. The questionnaire was presented using software that let the participants choose values between 1 and 100 with a slider.

The items were divided into the following categories; in each case, the first number of items applies to Experiment 1 and the second to Experiment 2:

Perceived intelligence of the robot Fifteen (twelve) items measuring how intelligent the participant felt the robot was during the interaction;

Quality of the interaction Twelve (nineteen) items measuring how smoothly the participant felt

the overall interaction went;

Task ease and success Eleven (six) items asking the participant how easy they found the various assembly tasks and how well they thought they performed; and

User feelings Nine (nine) items asking participants to rate their feelings while using the system.

The full sets of questionnaire items are available on request.

Results

[Table 4 about here.]

[Table 5 about here.]

Objective results. The findings for the objective measures are summarised in Tables 4 and 5. For each measure, we give the following information: the mean and the sample standard deviation for each of the two groups of participants, along with the significance level obtained on a two-tailed Mann-Whitney U test comparing the two sets of results. In the Experiment 1 results, the measures marked with a * were derived from the video recordings, and were therefore computed on the data from 40 participants, as three did not give permission for recordings; the remaining Experiment 1 measures were computed on the data from all 43 participants. As can be seen from the table, the choice of reference strategy had no significant effect on any of the objective measures considered in this study. A detailed analysis of the Experiment 1 participants' gaze and object-manipulation behaviour immediately after various forms of generated references from the robot also failed to find any significant differences between the various reference types.

[Table 6 about here.]

[Table 7 about here.]

Subjective results. Tables 6 and 7 show the mean response to each group of items from the user-satisfaction questionnaires, with the responses for negatively-posed questions inverted. For each group of items, we first computed Cronbach’s alpha to test the internal consistency. As shown in the first column of the table, the consistency was generally found to be acceptable ($\alpha \geq 0.7$) in Experiment 1, and somewhat lower (α between 0.555 and 0.779) in Experiment 2. The remainder of the tables show the mean and sample standard deviation for each measure, grouped by the reference strategy; as with the objective measures, the significance level from a two-tailed Mann-Whitney U test is shown in the final column. In summary, these tables demonstrates that the reference strategy had no significant overall effect on any of the classes of subjective measures.

However, on closer examination, there were some indications of differences between the reference-generation strategies. On Experiment 1, a sub-set of the “interaction quality” items referred specifically to the quality of the robot’s instructions. These items are shown in Table 8 (in the original German and in English translation), along with the mean responses to each item from the two groups on a 5-point Likert scale. The better score is highlighted for each question; for the negatively-posed questions (i.e., the third and fourth ones in the list), where a lower score would actually be better, the scores are inverted in the table. For five of the six questions in this sub-category, the participants who heard the context-sensitive references scored the system more highly. As shown in the bottom line of the table, the mean score on these items was 3.89 for the participants who heard the context-sensitive references, compared to 3.51 for the participants who heard the basic references. This overall trend resulted in a marginally significant difference between the two groups on these items: $z = -1.88, p = 0.06$.¹

[Table 8 about here.]

For Experiment 2, a marginal overall effect ($p = 0.082$) was found on the “Interaction quality” items of the questionnaire. In this case, a closer examination of the items in this

¹With a Bonferroni correction, the required significance level for this test would actually be $p = \frac{0.05}{4} = 0.0125$.

class—which proved to be quite a heterogeneous set ($\alpha = 0.555$)—showed that the reference strategy did have a particularly noticeable effect on responses to two of the items in this class. The two items are shown in Table 9: both of these items address the user’s feeling of knowing what they could do at any point in the interaction, and the responses on the two items are reasonably correlated with each other ($\alpha = 0.61$). As shown in the bottom line of the table, the mean score on these items for participants who heard the basic references was 36.2, while the context-sensitive participants gave a mean response of 56.0; a Mann-Whitney U test found that this effect was highly significant even with a Bonferroni correction: $z = 2.63, p < 0.001$.

[Table 9 about here.]

We also carried out a detailed examination of the relationship among the various subjective and objective measures, using a PARADISE analysis (Walker et al., 1997). For Experiment 1, this analysis found that the primary predictors of subjective user satisfaction were the dialogue length, the number of repetition requests, and the participants’ recall of the system instructions (Foster, Giuliani, & Knoll, 2009); for Experiment 2, the main predictors were the participant’s performance at the assembly task and the number of times they had to ask for instructions to be repeated (Giuliani et al., 2010).

[Table 10 about here.]

[Table 11 about here.]

Distribution of generated references. We analysed the logs to determine the distribution of generated references over the course of the study: we computed the total number of references of each type, as well as a separate count of initial references only. The results are shown in Tables 10 and 11. The main difference between the context-sensitive algorithm and the basic algorithm used in Experiment 1 was that only the context-sensitive version generated deictic references; it did so for 36% of the initial mentions and 22% of the total references. By design, it was not possible for either algorithm to generate a pronominal initial reference. In Experiment 2,

the basic algorithm was further modified to remove the ability to generate pronouns, while the context-sensitive algorithm generated a similar overall distribution to Experiment 1. We used a χ^2 test to compare these counts to the distribution of initial references used by the humans in the JCT (Table 1). In all cases, the generated frequencies were significantly different to those from the JCT; in Experiment 1: $65.2 < \chi^2 < 771$, $df = 3$, $p \approx 0$ and Experiment 2: $90.6 < \chi^2 < 497$, $df = 3$, $p \approx 0$.

Discussion

The choice of reference strategy had no significant effect on any of the objective measures on either study: the results on all measures of dialogue efficiency, dialogue quality, and task success were indistinguishable. On the other hand, the responses on selected items from the subjective questionnaires suggest that the choice of reference strategy had a positive effect. On Experiment 1, the participants tended to rate the robot as a better instruction giver if it used contextually varied, situationally-appropriate referring expressions; on Experiment 2, they felt much more confident about the overall flow of the dialogue with context-sensitive references. This agrees with the findings from the JCT dialogues (where partners used references that would appear inadequate on the surface) and the recent study of Campana et al. (2011): in both of those studies, the choice of referring expressions had no impact on primary task performance. In the simpler task addressed in the current human-robot study, where the task baseline is if anything even higher than in the JCT, it is therefore not surprising that the primary effect of the reference strategy was subtle and subjective, rather than large and objective.

Both the basic and the context-sensitive algorithms generated patterns of references that differed significantly from the initial references found in the JCT corpus. However, despite the superficial similarity between the JCT task and the JAST robot scenario evaluated, the necessary technical limitations on the human-robot system mean that the dialogues resulting from this experiment are quite different than those in the JCT. In particular, the human-robot dialogues are much shorter and less elaborate, as the objects being constructed were much less complex. Also,

the situation of the two participants is more asymmetrical than in the human-human condition, and the possible interactions constrained by the interactive capabilities of the system: only the robot knows the assembly plan, while only the user carries out the actions, and the user's possible spoken contributions are very limited. The difference in overall reference patterns is therefore not surprising. It is worth noting that, despite the differences, only the context-sensitive system made use of the full range of reference types that were found in the JCT data.

The subjective questionnaire was modified from Experiment 1 to Experiment 2 because—as noted above—the second experiment formed part of a larger study (Bard et al., 2009) where the primary emphasis was on error detection and recovery; this unfortunately meant that the instruction-quality items from Table 8 were removed, so no comparison could be made. On the other hand, the questionnaire items that were most affected by the reference strategy in Experiment 2 (Table 9) were also included in the Experiment 1 questionnaire, but there the reference strategy did not have an impact. This is most likely due to a combination of two factors: first, from a user standpoint, the interactions in Scenario 2 are much less predictable, as the robot may intervene at any time with a correction. Second, adding the goal-inference system as an additional input to the dialogue manager made the system responses significantly slower, which would also tend to emphasise any unpredictable aspects of the interaction. Given these factors, it is not surprising that the general decrease in interaction quality produced by the basic references manifested itself on Experiment 2 in lower responses to these two questions.

In summary, these results suggest that—at least in this scenario—the references generated by the context-sensitive strategy were perceived to be of at least as good quality than those generated by the basic strategy, and possibly even of higher quality.

Conclusions

We have described a humanoid robot that is designed to cooperate with a human partner on a joint construction task. Inspired by the referring phenomena found in the JCT corpus of human-human dialogues in a similar joint construction domain, we have implemented a

context-sensitive reference-generation algorithm that takes into account aspects of the physical, discourse, and task context. We have then compared the context-sensitive algorithm to the basic Dale and Reiter (1995) incremental algorithm through a pair of task-based evaluations, considering objective measures of dialogue efficiency, dialogue quality, and task success along with subjective measures gathered from a usability questionnaire. In both cases, we have found no difference between the performance of the two algorithms in terms of the objective measures, but both studies did find a tendency for the context-sensitive references to improve the participants' subjective impressions of interacting with the robot system.

The current studies contrast with the state-of-the-art in the evaluation of referring expressions, where the dominant technique is to generate stand-alone references, and to evaluate them by comparing against a corpus of human-generated references from a similar null context. That form of evaluation is appropriate if the goal is simply to model human performance as accurately as possible; however, if the goal is to generate useful references in the context of an interactive system, the most important criterion of success is instead the effect of different forms of reference on the hearer. Since it has been demonstrated repeatedly that, in the context of natural-language generation, intrinsic corpus-similarity metrics show very little correlation with any extrinsic task-performance or subjective preference metrics, it is necessary to carry out this sort of task-based evaluation to assess the performance of referring expressions in a situated, interactive context. Even though such evaluations are more logistically difficult to run in practice, due to the necessity of recruiting participants and deploying a robust, fully interactive system, there is really no alternative: at present, corpus similarity is an inadequate metric for any evaluation of a reference-generation system intended for use in an interactive context. Note that Spanger, Iida, Tokunaga, Terai, and Kuriyama (2013) recently came to a similar conclusion regarding the evaluation of referring expressions in the context of collaborative situated dialogues.

The findings of these studies also confirm that it is worth the effort of developing more context-sensitive reference-generation algorithms for use in interactive contexts: even if (as here) the overt impact of the more sophisticated reference strategy is subtle, it is still likely to have an

overall positive effect on users of the system—even if that is only at the subjective level rather than at the objective level of task performance or dialogue quality. In future work, it would be useful to measure the impact of any advanced reference-generation strategy on users' opinions of an artificial agent more systematically, using measurement instruments such as the GODSPEED questionnaire series (Bartneck, Kulić, Croft, & Zoghbi, 2009), which is designed to be a general-purpose measurement tool for human-robot interaction.

Although the generation algorithm was inspired by the referring phenomena found in the JCT corpus of human-human joint construction dialogues, the distribution of references generated by the system were significantly different from the distribution found in the JCT. This was largely because, even though the high-level task is similar between the two applications, the human-robot dialogues are much less symmetrical at both the task and the linguistic levels, and—because of the technical constraints on the robot system—the range of possible utterances is much narrower than those that were employed by the JCT corpus speakers. However, comparing the automatically generated referring expressions with those found in a corpus is still worth doing: as noted earlier, automated evaluations are much easier to perform in practice than evaluations involving human subjects, as long as some corpus-derived metrics can be found that agree with the human results. Carrying out such a study would have several requirements. First, it would require a closer mapping between the corpus scenario and the scenario supported by the artificial system than was possible here. Also, a proper comparison would require a corpus that is “semantically transparent” (van Deemter et al., 2006): that is, full contextual information as shown in Table 2 would need to be available for each referring expression in the corpus. As more corpora of human-generated references in interactive contexts become available—the JCT corpus, along with other corpora such as REX (Tokunaga, Iida, Terai, & Kuriyama, 2012) and iMap (Guhe & Bard, 2008)—it may eventually be possible to find automated evaluation metrics that do correlate with the factors that are relevant when automatically generating output in interactive contexts.

Acknowledgements

The authors thank all of our JAST colleagues for productive collaboration, and the journal reviewers for useful comments on previous versions of this article.

Funding

The research leading to these results has received funding from the European Union's Sixth Framework Programme (FP6/2003-2006) under grant agreements no. 003747, JAST: Joint Action Science and Technology and no. 045388, INDIGO: Interaction with Personality and Dialogue Enabled Robots, and from the Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 270435, JAMES: Joint Action for Multimodal Embodied Social Systems.

References

- Argyle, M., & Graham, J. A. (1976). The Central Europe experiment: Looking at persons and looking at objects. *Environmental Psychology and Nonverbal Behavior*, *1*(1), 6–16. doi: 10.1007/BF01115461
- Ariel, M. (1991). The function of accessibility in a theory of grammar. *Journal of Pragmatics*, *16*(5), 443–463. doi: 10.1016/0378-2166(91)90136-L
- Bard, E. G., & Aylett, M. P. (2004). Referential form, word duration, and modeling the listener in spoken dialogue. In J. C. Trueswell & M. K. Tanenhaus (Eds.), *Approaches to studying world-situated language use: Bridging the language-as-product and language-as-action traditions* (pp. 173–192). The MIT Press.
- Bard, E. G., Bekkering, H., Bicho, E., de Bruijn, E., Cuijpers, R., Erlhagen, W., ... Sousa, M. (2009). *Final report on the evaluation of the robot system* (Deliverable No. 5.13). JAST Project.
- Bard, E. G., Hill, R., & Foster, M. E. (2008). What tunes accessibility of referring expressions in task-related dialogue? In *Proceedings of the 30th Annual Meeting of the Cognitive Science Society (CogSci 2008)* (pp. 945–950). Washington, DC, USA. Retrieved from <http://csjarchive.cogsci.rpi.edu/proceedings/2008/pdfs/p945.pdf>
- Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, *1*, 71–81. doi: 10.1007/s12369-008-0001-3
- Belz, A., Kow, E., Viethen, J., & Gatt, A. (2010). Generating referring expressions in context: The GREC task evaluation challenges. In *Empirical methods in natural language generation* (pp. 294–327). Springer Berlin/Heidelberg. doi: 10.1007/978-3-642-15573-4_15
- Belz, A., & Reiter, E. (2006). Comparing automatic and human evaluation of NLG systems. In *Proceedings of the 11th Conference of the European Chapter of the Association for*

- Computational Linguistics (EACL 2006)* (pp. 313–320). Trento, Italy. Retrieved from <http://aclweb.org/anthology/E06-1040.pdf>
- Bicho, E., Erlhagen, W., Louro, L., & Costa e Silva, E. (2011). Neuro-cognitive mechanisms of decision making in joint action: a human-robot interaction study. *Human Movement Science, 30*(5), 846–868. doi: 10.1016/j.humov.2010.08.012
- Bicho, E., Louro, L., & Erlhagen, W. (2010). Integrating verbal and nonverbal communication in a dynamic neural field architecture for human-robot interaction. *Frontiers in Neurorobotics, 4*(5). doi: 10.3389/fnbot.2010.00005
- Binsted, K., Pain, H., & Ritchie, G. (1997). Children’s evaluation of computer-generated punning riddles. *Pragmatics & Cognition, 5*(2), 305–354. doi: 10.1075/pc.5.2.06bin
- Campana, E., Tanenhaus, M. K., Allen, J. F., & Remington, R. (2011). Natural discourse reference generation reduces cognitive load in spoken systems. *Natural Language Engineering, 17*(03), 311–329. doi: 10.1017/S1351324910000227
- Carenini, G., & Moore, J. D. (2006). Generating and evaluating evaluative arguments. *Artificial Intelligence, 170*(11), 925–952. doi: 10.1016/j.artint.2006.05.003
- Carletta, J., Hill, R., Nicol, C., Taylor, T., de Ruiter, J., & Bard, E. (2010). Eyetracking for two-person tasks with manipulation of a virtual world. *Behavior Research Methods, 42*(1), 254–265. doi: 10.3758/BRM.42.1.254
- Clark, H. H., & Bangerter, A. (2004). Changing ideas about reference. In I. A. Novacek & D. Sperber (Eds.), *Experimental pragmatics* (pp. 25–49). Palgrave Macmillan.
- Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition, 22*(1), 1–39. doi: 10.1016/0010-0277(86)90010-7
- Dale, R., & Reiter, E. (1995). Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science, 19*(2), 233–263. doi: 10.1207/s15516709cog1902_3
- Foster, M. E. (2008). Automated metrics that agree with human judgements on generated output for an embodied conversational agent. In *Proceedings of the 5th International Conference*

- on Natural Language Generation (INLG 2008)* (pp. 95–103). Salt Fork, Ohio, USA.
Retrieved from <http://www.aclweb.org/anthology/W08-1113>
- Foster, M. E., Bard, E. G., Hill, R. L., Guhe, M., Oberlander, J., & Knoll, A. (2008). The roles of haptic-ostensive referring expressions in cooperative, task-based human-robot dialogue. In *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction (HRI 2008)* (pp. 295–302). Amsterdam, Netherlands. doi: 10.1145/1349822.1349861
- Foster, M. E., Giuliani, M., Isard, A., Matheson, C., Oberlander, J., & Knoll, A. (2009). Evaluating description and reference strategies in a cooperative human-robot dialogue system. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI-09)* (pp. 1818–1823). Pasadena, California, USA. Retrieved from <http://ijcai.org/papers09/Papers/IJCAI09-302.pdf>
- Foster, M. E., Giuliani, M., & Knoll, A. (2009). Comparing objective and subjective measures of usability in a human-robot dialogue system. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-IJCNLP 2009)* (pp. 879–887). Singapore.
Retrieved from <http://www.aclweb.org/anthology/P09-1099>
- Garrod, S., & Pickering, M. J. (2009). Joint action, interactive alignment, and dialog. *Topics in Cognitive Science*, 1(2), 292–304. doi: 10.1111/j.1756-8765.2009.01020.x
- Gatt, A., & Belz, A. (2010). Introducing shared tasks to NLG: The TUNA shared task evaluation challenges. In *Empirical methods in natural language generation* (pp. 264–293). Springer Berlin/Heidelberg. doi: 10.1007/978-3-642-15573-4_14
- Giuliani, M., Foster, M. E., Isard, A., Matheson, C., Oberlander, J., & Knoll, A. (2010). Situated reference in a hybrid human-robot interaction system. In *Proceedings of the 6th International Natural Language Generation Conference (INLG 2010)* (pp. 67–76). Trim, Ireland. Retrieved from <http://aclweb.org/anthology/W10-4207>
- Giuliani, M., & Knoll, A. (2008). MultiML: a general purpose representation language for multimodal human utterances. In *Proceedings of the 10th International Conference on*

- Multimodal Interfaces (ICMI 2008)* (pp. 165–172). Chania, Crete, Greece. doi: 10.1145/1452392.1452424
- Grice, H. P. (1975). Logic and conversation. In D. Davidson & G. Harman (Eds.), *The logic of grammar* (pp. 64–75). Dickenson.
- Grosz, B. J., Weinstein, S., & Joshi, A. K. (1995). Centering: a framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2), 203–225. Retrieved from <http://aclweb.org/anthology/J95-2003>
- Guhe, M., & Bard, E. G. (2008). Adapting referring expressions to the task environment. In *Proceedings of the 30th Annual Meeting of the Cognitive Science Society (CogSci 2008)* (pp. 2404–2409). Washington, DC, USA. Retrieved from <http://csjarchive.cogsci.rpi.edu/proceedings/2008/pdfs/p2404.pdf>
- Gundel, J. K., Hedberg, N., & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, 69(2), 274–307. Retrieved from <http://www.jstor.org/stable/416535>
- Horton, W. S., & Keysar, B. (1996). When do speakers take into account common ground? *Cognition*, 59(1), 91–117. doi: 10.1016/0010-0277(96)81418-1
- Karasimos, A., & Isard, A. (2004). Multi-lingual evaluation of a natural language generation system. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)* (pp. 829–832). Lisbon, Portugal. Retrieved from <http://www.lrec-conf.org/proceedings/lrec2004/pdf/134.pdf>
- Kelleher, J. D., & Kruijff, G.-J. M. (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 (COLING-ACL 2006)* (pp. 1041–1048). Sydney, Australia. doi: 10.3115/1220175.1220306
- Koller, A., Staudte, M., Garoufi, K., & Crocker, M. (2012). Enhancing referential success by tracking hearer gaze. In *Proceedings of the 13th Annual Meeting of the Special Interest*

- Group on Discourse and Dialogue (SIGdial 2012)* (pp. 30–39). Seoul, South Korea.
Retrieved from <http://www.aclweb.org/anthology/W12-1604>
- Koller, A., Striegnitz, K., Gargett, A., Byron, D., Cassell, J., Dale, R., . . . Oberlander, J. (2010). Report on the second NLG challenge on generating instructions in virtual environments (GIVE-2). In *Proceedings of the 6th International Natural Language Generation Conference (INLG 2010)* (pp. 243–250). Trim, Ireland. Retrieved from <http://aclweb.org/anthology/W10-4233>
- Krahmer, E. (2010). What computational linguists can learn from psychologists (and vice versa). *Computational Linguistics*, 36(2), 285–294. doi: 10.1162/coli.2010.36.2.36201
- Krahmer, E., & Theune, M. (2002). Efficient context-sensitive generation of referring expressions. In K. van Deemter & R. Kibble (Eds.), *Information sharing: Reference and presupposition in language generation and interpretation* (pp. 223–264). CSLI Publications.
- Krahmer, E., & van Deemter, K. (2012). Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1), 173–218. doi: 10.1162/COLI_a_00088
- Kranstedt, A., Lücking, A., Pfeiffer, T., Rieser, H., & Wachsmuth, I. (2006). Deictic object reference in task-oriented dialogue. In G. Rickheit & I. Wachsmuth (Eds.), *Situated communication* (pp. 155–207). Mouton de Gruyter.
- Kranstedt, A., & Wachsmuth, I. (2005). Incremental generation of multimodal deixis referring to objects. In *Proceedings of the 10th European Workshop on Natural Language Generation (ENLG-05)* (pp. 75–82). Aberdeen, Scotland, UK. Retrieved from <http://aclweb.org/anthology/W05-1608>
- Larsson, S., & Traum, D. (2000). Information state and dialogue management in the TRINDI dialogue move engine toolkit. *Natural Language Engineering*, 6(3&4), 323–340. doi: 10.1017/S1351324900002539
- Litman, D. J., & Pan, S. (2002). Designing and evaluating an adaptive spoken dialogue system. *User Modeling and User-Adapted Interaction*, 12(2–3), 111–137. doi:

10.1023/A:1015036910358

- Mellish, C., & Dale, R. (1998). Evaluation in the context of natural language generation. *Computer Speech & Language*, 12(4), 349–373. doi: 10.1006/csla.1998.0106
- Nielsen, J., & Levy, J. (1994). Measuring usability: preference vs. performance. *Communications of the ACM*, 37(4), 66–75. doi: 10.1145/175276.175282
- Oviatt, S. (1999). Ten myths of multimodal interaction. *Communications of the ACM*, 42(11), 74–81. doi: 10.1145/319382.319398
- Reiter, E. (2011). Task-based evaluation of NLG systems: Control vs real-world context. In *Proceedings of the UCNLG+Eval: Language Generation and Evaluation Workshop* (pp. 28–32). Edinburgh, Scotland, UK. Retrieved from <http://www.aclweb.org/anthology/W11-2704>
- Spanger, P., Iida, R., Tokunaga, T., Terai, A., & Kuriyama, N. (2013). A task-performance evaluation of referring expressions in situated collaborative task dialogues. *Language Resources and Evaluation*. (Advance online publication) doi: 10.1007/s10579-013-9240-5
- Staudte, M., & Crocker, M. W. (2009). Producing and resolving multi-modal referring expressions in human-robot interaction. In *Proceedings of the PRE-CogSci Workshop at CogSci 2009*. Amsterdam, Netherlands. Retrieved from <http://pre2009.uvt.nl/pdf/staudte.pdf>
- Steedman, M. (2000). *The syntactic process*. The MIT Press.
- Tokunaga, T., Iida, R., Terai, A., & Kuriyama, N. (2012). The REX corpora: A collection of multimodal corpora of referring expressions in collaborative problem solving dialogues. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*. Istanbul, Turkey. Retrieved from http://www.lrec-conf.org/proceedings/lrec2012/pdf/676_Paper.pdf
- van Breemen, A., Yan, X., & Meerbeek, B. (2005). iCat: an animated user-interface robot with personality. In *Proceedings of the 4th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2005)* (pp. 143–144). Utrecht, Netherlands. doi:

10.1145/1082473.1082823

- van Deemter, K., Gatt, A., van der Sluis, I., & Power, R. (2012). Generation of referring expressions: Assessing the incremental algorithm. *Cognitive Science*, 36(5), 799–836. doi: 10.1111/j.1551-6709.2011.01205.x
- van Deemter, K., van der Sluis, I., & Gatt, A. (2006). Building a semantically transparent corpus for the generation of referring expressions. In *Proceedings of the 4th International Conference on Natural Language Generation (INLG 2006)* (pp. 130–132). Sydney, Australia. Retrieved from <http://www.aclweb.org/anthology/W06-1420>
- van der Sluis, I. (2005). *Multimodal reference: Studies in automatic generation of multimodal referring expressions*. Doctoral dissertation, University of Tilburg.
- van der Sluis, I., & Luz, S. (2011). A cross-linguistic study on the production of multimodal referring expressions in dialogue. In *Proceedings of the 13th European Workshop on Natural Language Generation (ENLG 2011)* (pp. 53–62). Nancy, France. Retrieved from <http://www.aclweb.org/anthology/W11-2807>
- Viethen, H., Dale, R., & Guhe, M. (2011). The impact of visual context on the content of referring expressions. In *Proceedings of the 13th European Workshop on Natural Language Generation (ENLG 2011)* (pp. 44–52). Nancy, France. Retrieved from <http://www.aclweb.org/anthology/W11-2806>
- Walker, M. A., Litman, D. J., Kamm, C. A., & Abella, A. (1997). PARADISE: A framework for evaluating spoken dialogue agents. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL 1997)* (pp. 271–280). Madrid, Spain. doi: 10.3115/976909.979652
- White, M. (2004). Reining in CCG chart realization. In *Proceedings of the 3rd International Conference on Natural Language Generation (INLG 2004)* (pp. 182–191). Brockenhurst, England, UK. doi: 10.1007/978-3-540-27823-8_19
- White, M. (2006). Efficient realization of coordinate structures in Combinatory Categorical Grammar. *Research on Language and Computation*, 4(1), 39–75. doi:

10.1007/s11168-006-9010-2

White, M., Foster, M. E., Oberlander, J., & Brown, A. (2005). Using facial feedback to enhance turn-taking in a multimodal dialogue system. In *Proceedings of HCI International 2005* (Vol. 8). Las Vegas, Nevada, USA.

Table 1

Distribution of initial mentions in the JCT corpus (Bard et al., 2008)

Accessibility	Count	%
Indefinite NP / Bare N	225	17
Definite NP	488	36
Deictic / Possessive Pronoun	464	35
Other Pronoun	160	12

Table 2

Reference generation algorithm

previous ref			robot holding	distractors	reftype
focal	same turn	indef			
N			Y	-	deictic
N			N	N	definite
N			N	Y	indefinite
Y	Y	-	-	-	pronoun
Y	N	-	Y	-	deictic
Y	N	-	N	-	pronoun
N	-	-	Y	-	deictic
N	-	-	N	Y	indef
N	-	-	Y	-	deictic
N	-	N	N	-	definite
N	-	Y	N	N	definite
N	-	Y	N	Y	indefinite

Table 3

Participants' demographic information

	Experiment 1	Experiment 2
Total <i>n</i>	43	41
<i>n</i> males	27	33
Mean age	24.5	24.5
Min age	14	19
Max age	55	42
Computer knowledge	3.4	4.1
ASR knowledge	2.3	2.0
HRI knowledge	2.0	1.7
<i>n</i> Informatics	12	14
<i>n</i> Mathematics	10	14

Table 4

Experiment 1: Objective results

Measure	Basic (Stdev)	Context-Sensitive (Stdev)	Mann-Whitney p
Duration (s)*	300.5 (45.7)	310.3 (62.9)	0.82
Duration (turns)	14.45 (2.09)	14.81 (1.97)	0.46
Response time*	2.51 (0.89)	3.11 (1.30)	0.10
Rep requests	1.95 (2.10)	1.76 (1.45)	0.85
Failed gives	1.27 (1.64)	0.86 (0.96)	0.59
Looks at robot*	21.0 (4.84)	26.4 (10.2)	0.11
Look at robot (%)*	28 (9)	27 (9)	0.72
Correct assembly	0.71 (0.28)	0.73 (0.32)	0.64
Recall	0.80 (0.30)	0.79 (0.30)	0.92

*Computed on $n = 40$ participants; all other measures computed on $n = 43$ participants.

Table 5

Experiment 2: Objective results

Measure	Basic (Stdev)	Context-Sensitive (Stdev)	Mann-Whitney p
Duration (s.)	404.3 (62.8)	410.5 (94.6)	0.90
Duration (turns)	29.8 (5.02)	31.2 (5.57)	0.44
Explanations	2.21 (0.63)	2.41 (0.80)	0.44
Rep requests	0.26 (0.45)	0.32 (0.78)	0.68
Successful trials	1.58 (0.61)	1.55 (0.74)	0.93

Table 6

Experiment 1: Subjective results

Category	Cronbach's α	Basic (stdev)	Context-Sensitive (stdev)	Mann-Whitney p
Perceived intelligence	0.868	3.56 (0.71)	3.54 (0.64)	0.93
Interaction quality	0.848	3.62 (0.71)	3.91 (0.61)	0.22
Task ease/success	0.857	3.98 (0.80)	4.21 (0.50)	0.45
User feelings	0.690	3.57 (0.68)	3.75 (0.52)	0.37

Table 7

Experiment 2: Subjective results

Category	Cronbach's α	Basic (Stdev)	Context-Sensitive (Stdev)	Mann-Whitney p
Perceived intelligence	0.779	75.8 (14.4)	73.4 (11.7)	0.80
Interaction quality	0.555	66.7 (10.2)	71.7 (10.1)	0.082
Task ease/success	0.610	80.7 (14.0)	81.4 (12.4)	0.53
User feelings	0.687	65.9 (15.9)	65.3 (12.2)	0.80

Table 8

Experiment 1: Mean responses to questionnaire items addressing the quality of the robot's instructions

Question	Basic (Stdev)	Context-Sensitive (Stdev)
Der Roboter gab mir nützliche Anweisungen. The robot gave me useful instructions.	3.86 (0.91)	3.68 (0.95)
Es war einfach den Anweisungen des Roboters zu folgen. It was easy to follow the robot's instructions.	3.33 (0.91)	3.68 (0.82)
Der Roboter gab zu viele Anweisungen auf einmal. The robot gave too many instructions at once.	3.19 (1.36)*	4.16 (1.07)*
Die Anweisungen des Roboters waren zu ausführlich. The robot's instructions were too detailed.	3.71 (1.15)*	4.32 (0.75)*
Immer wenn der Roboter über Bauteile gesprochen hat, wusste ich genau, von welchem Bauteil er spricht. When the robot talked about pieces, I always knew exactly which piece it meant.	3.29 (1.27)	3.47 (1.35)
Der Roboter gab mir gute Anweisungen The robot gave me good instructions.	3.67 (0.73)	3.68 (1.16)
	3.51 (0.70)	3.89 (0.67)

*Negatively-posed question; value shown is 6 – mean

Table 9

Experiment 2: User responses to questionnaire items testing the user's confidence in the interaction

Item	Basic (Stdev)	Context-Sensitive (Stdev)
Wenn der Roboter mich nicht verstand, dann war mir klar, wie ich reagieren musste. When the robot did not understand me, it was clear what I had to do.	34.7 (23.7)	50.4 (28.8)
Ich wusste zu jedem Zeitpunkt der Unterhaltung, was ich machen oder sagen konnte. At each point in the conversation, I knew what I could do or say.	37.6 (25.1)	61.6 (29.6)
	36.2 (19.5)	56.0 (24.6)

Table 10

Experiment 1: Distribution of generated references

	Basic				Context-Sensitive			
	<i>Initial</i>	<i>%</i>	<i>All</i>	<i>%</i>	<i>Initial</i>	<i>%</i>	<i>All</i>	<i>%</i>
Indefinite	91	18	560	32	84	18	516	33
Definite	414	82	1044	60	218	46	586	37
Deictic	0	0	0	0	173	36	350	22
Pronoun	0	0	136	8	0	0	121	8

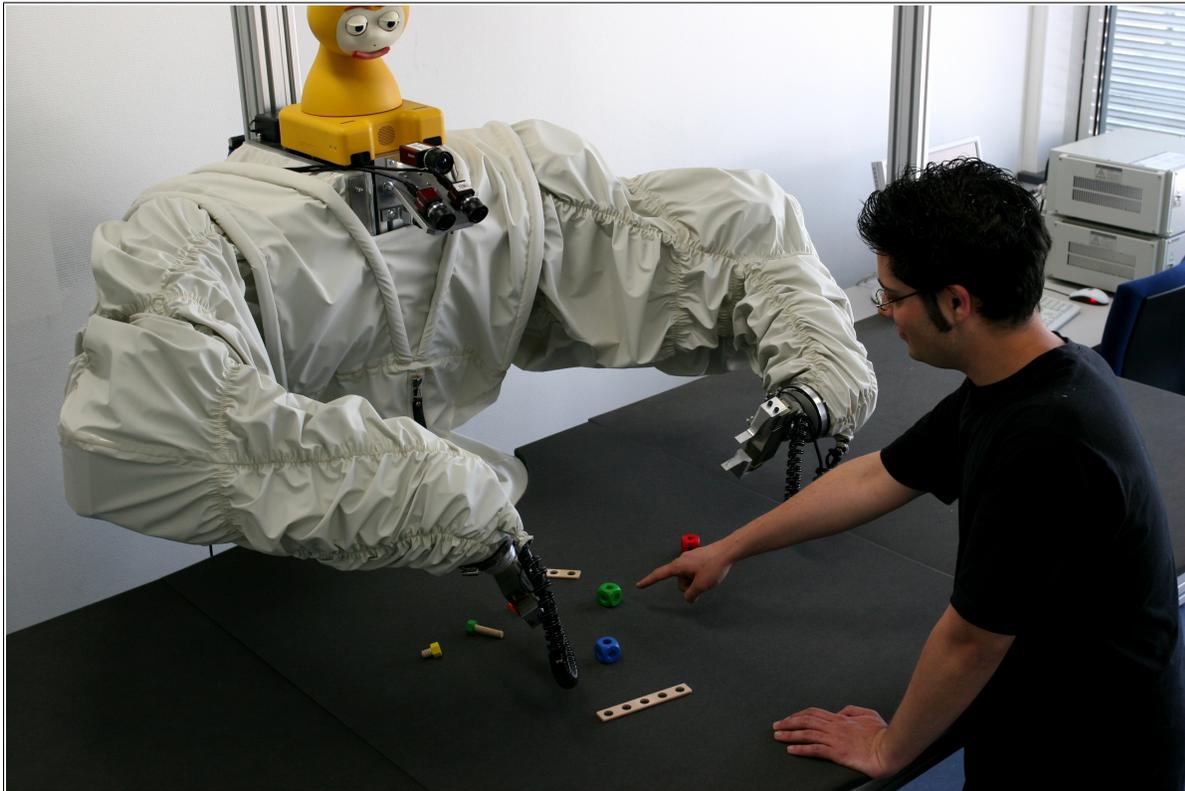


Figure 1. The JAST dialogue robot

-
1. **System** First we will build a windmill. Okay?
 2. **User** Okay.
 3. **System** To make a windmill, we must make a tower.
 4. **System** [*picking up and holding out red cube*] To make a tower, insert the green bolt through the end of this red cube and screw it into the blue cube.
 5. **User** [*takes cube, performs action*] Okay.
-

Figure 2. Scenario 1: robot instructing user (context-sensitive reference highlighted)

-
1. **System** First we will build a windmill.
 2. **User** Okay.
 3. **User** *{picks up a yellow cube, unnecessary piece for a windmill}*
 4. **System** You don't need a yellow cube to build a windmill.
 5. **System** To build a windmill, you first need to build a tower.
 6. **System** *[picking up and holding out red cube]* To build the tower, insert the green bolt through the end of this red cube and screw it into the blue cube.
 7. **User** *[takes cube, performs action]* Okay.
-

Figure 3. Scenario 2: robot and user jointly executing plan, with error detection (context-sensitive references highlighted)

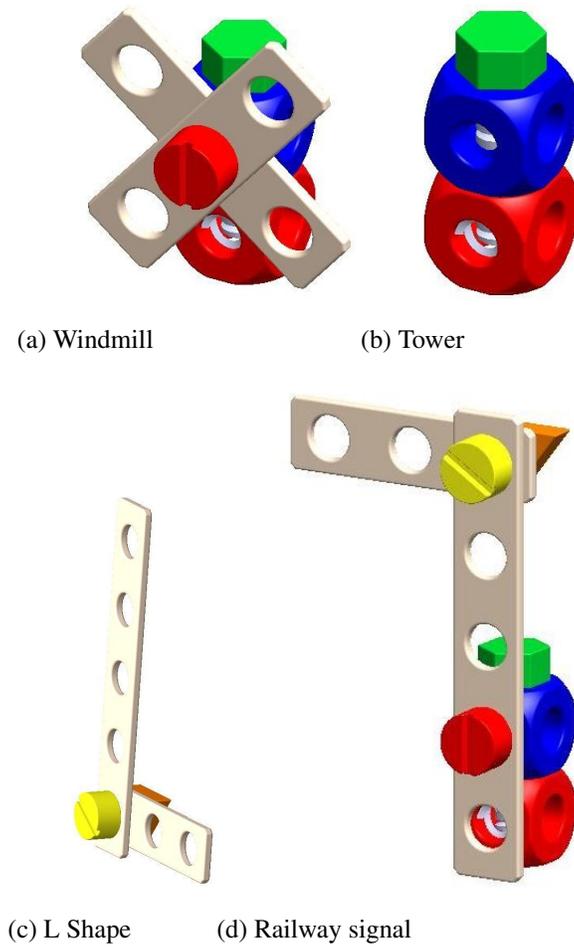


Figure 4. The target objects used in the experiments