Natural Language Engineering **17** (3): 311–329. © Cambridge University Press 2010
doi:10.1017/S1351324910000227

Natural discourse reference generation reduces cognitive load in spoken systems

E. CAMPANA[1], M. K. TANENHAUS[2], J. F. ALLEN[3] and R. REMINGTON[4]

[1]*Psychology Department, 699 S. Mill Avenue, Arizona State University, Tempe, AZ 85281, USA*
email: ellen.campana@asu.edu

[2]*Brain and Cognitive Sciences, RC Box 270268, University of Rochester, Rochester, NY 14627, USA*
email: mtan@bcs.rochester.edu

[3]*Computer Science, RC Box 270226, University of Rochester, Rochester, NY 14627, USA*
email: james@cs.rochester.edu

[4]*School of Psychology, McElwain Building, The University of Queensland, St. Lucia QLD 4072, Australia*
email: r.remington@psy.uq.edu.au

(*Received 1 October 2009; revised 7 June 2010; accepted 14 July 2010; first published online 10 September 2010*)

Abstract

The generation of referring expressions is a central topic in computational linguistics. Natural referring expressions – both definite references like 'the baseball cap' and pronouns like 'it' – are dependent on discourse context. We examine the practical implications of context-dependent referring expression generation for the design of *spoken systems*. Currently, not all spoken systems have the goal of generating natural referring expressions. Many researchers believe that the context-dependency of natural referring expressions actually makes systems *less* usable. Using the dual-task paradigm, we demonstrate that generating natural referring expressions that are dependent on discourse context reduces cognitive load. Somewhat surprisingly, we also demonstrate that practice does not improve cognitive load in systems that generate consistent (context-independent) referring expressions. We discuss practical implications for spoken systems as well as other areas of referring expression generation.

1 Introduction

Jerry Coleman, baseball radio announcer for the San Diego Padres, once gave the following commentary: 'Winfield goes back to the wall. He hits his head on the wall, and it rolls off! It's rolling all the way back to second base! This is a terrible thing for the Padres!' (Dickson 2008). Coleman's commentary is remembered long after the original broadcast because of the humorous use of the referring expression 'it'. As a radio commentator, his goal was to describe where the baseball was. Of course, he meant to convey the message that the baseball, not Winfield's head, rolled all the way to second base. However, that is not the most natural interpretation.

Such a miscommunication is possible because in natural language the specific meanings of referring expressions like 'it' and 'the baseball cap' (i.e. which entities they refer to) depend on context. To avoid miscommunication, a speaker must

https://doi.org/10.1017/S1351324910000227 Published online by Cambridge University Press

describe an object so as to distinguish it from the other objects the listener may be considering as potential referents (Olson 1970; Grice 1975; Clark and Wilkes-Gibbs 1986; Clark 1992; Paraboni, van Deemter and Masthoff 2007). For instance, 'the hat' distinguishes a baseball cap from a pair of socks, but 'the baseball cap' is necessary to distinguish that same object from other hats (e.g. a cowboy hat, a winter hat).

There is a vast literature on the computational generation of human-like referring expressions (Appelt 1985; Dale 1988; Dale and Reiter 1995; Reiter and Dale 2000; Krahmer and Thune 2002; Croitoru and van Deemter 2007; Gatt and van Deemter 2007; Paraboni, van Deemter and Masthoff 2007; van der Sluis and Krahmer 2007; Gatt and van Deemter 2009). The basic task of a referring expression generation (REG) algorithm is to determine the content (e.g. physical features, color, size) and the linguistic form (e.g. noun, pronoun, complex noun phrase) that uniquely describe the intended entity, given the context. The content and the linguistic form together determine the surface form of the referring expression (e.g. 'it', 'the hat', 'a red baseball cap').

In many algorithms, both content determination and the selection of the linguistic form vary with the evolving discourse context (Grosz and Sidner 1986; DeVault, Rich and Sidner 2004; Belz and Varges 2007; Guhe and Bard 2008), which is consistent with the literature on human language (Gundel, Hedberg and Zacharski 1993; Hajičová 1993; Grosz, Joshi and Weinstein 1995; Walker, Joshi and Prince 1998; Hajičová, Sgall and Havelka 2003). There is an important distinction between the first mention of an object and subsequent rementions of the same object. In the first mention, the referring expression includes enough information to distinguish that object from all other objects in the domain. In subsequent rementions, the referring expression can be much shorter, distinguishing the object from the smaller set of other objects that have been previously described. When an object is the most salient object in the domain, for instance because it is the topic of the previous sentence, it can be referred to using a pronoun like 'it'.

Consider a context in which there are three different objects, in this case screws. The three screws differ in material (metal and plastic) and in length (three different lengths, with the plastic one being longest of the three). Example (1) contains felicitous language that a human might produce in this context.

(1) a. The long metal screw goes perpendicular to the plastic screw.
b. Make sure the metal screw is tightened down.
c. It should go all the way through the board.

In (1a), the first time any object is mentioned, the speaker refers to that object as 'the long metal screw'. This description distinguishes it from the other two screws in the domain (the shorter metal screw and the longer plastic screw). In (1b), the speaker then refers to the same object as 'the metal screw'. This description distinguishes it from the other object that has been mentioned (the longer plastic screw), but not the object that has not been mentioned (the shorter metal screw). This is sufficient because, due to the evolving discourse context, the shorter metal

https://doi.org/10.1017/S1351324910000227 Published online by Cambridge University Press

screw is unlikely to be considered by the listener as a potential referent. Although it would certainly be possible for the speaker to include the additional modifier 'long' in the second mention, it would be redundant in this context. Speakers do not typically include redundant information (Grice 1975; Pechman 1989).

In (1c) the speaker refers to the long metal screw with the pronoun 'it'. This description unambiguously distinguishes the target object from all other objects because at this point in the discourse the long metal screw is by far the most salient object; It is the most recently mentioned object and it was the topic of the previous sentence. Barring intervening pragmatics (e.g. the listener screwing in the plastic screw only part way), 'it' cannot refer to either the plastic screw or the shorter metal screw here. Moreover, because 'it' unambiguously refers to the long metal screw, use of a definite referring expression to refer to it would likely introduce unintended implicatures (Grice 1975; Levinson 2000). These implicatures might cause the listener to wonder, for instance, if there was an unseen, unmentioned screw that was very important (e.g. blocking the goal).

As Example (1) demonstrates, identifying an effective description depends on predicting the perceived set of alternatives in the listener's mind. In human–human communication, prediction is supported by cognitive and perceptual mechanisms that are common to the speaker and the listener. Many computational natural language generation (NLG) systems explicitly model the relevant cognitive and perceptual processes to improve REG (Grosz and Sidner 1986; Greenbacker and McCoy 2009; Guhe 2009).

There are many open questions regarding how human perceptual and cognitive processes affect reference generation, and thus how systems might make use of the same information. While reference generation algorithms can produce many of the referring expressions humans produce, they do not cover the full range. Veithen and Dale (2006) argue that choice of referential strategy (e.g. using a landmark or a grid location in a grid-based domain) is an important but underexplored factor. In the cognitive science literature it is clear that pragmatics (Campana, Brown-Schmidt and Tanenhaus 2002; Brown-Schmidt, Campana and Tanenhaus 2005; Guhe and Bard 2008), attentional limits (Brown-Schmidt and Tanenhaus 2008), and the details of visual perception (Krauss and Weinheimer 1964; Krauss and Weinheimer 1966; Krauss and Weinheimer 1967) affect the form and content of referring expressions. Understanding the effects of cognitive and perceptual factors, and how they should influence REG, is an increasingly active area of research.

We focus on the practical implications of context-dependent referring expressions specifically for the design of *spoken systems*. After nearly 40 years of research and development, spoken systems are now in widespread use for tasks ranging from checking bank balances to interacting with devices like phones, music players, and global positioning systems. It is becoming clear to even those outside the computational linguistics community that there are demonstrable benefits to spoken systems. It is a time of great promise for research in NLG via spoken systems, with more complex contexts to investigate and a wider community of users. However, not all researchers in spoken systems agree that information about the referring expressions humans naturally produce is relevant to the design of usable spoken

https://doi.org/10.1017/S1351324910000227 Published online by Cambridge University Press

systems. Some have even argued that spoken systems should deliberately AVOID human-like referring expressions in order to improve usability (Schneiderman 1980).

2 Approaches to spoken system design

We now provide an overview of the two basic approaches to the design of spoken systems, the NATURAL approach and the STANDARDIZED approach. Both have solid theoretical bases and active research and development communities built around them. When designing spoken systems, developers must choose an approach based on intuition, beliefs, and available tools. Our goal is to support decision-making with empirical data. We also expect the present research to be useful for guiding future research efforts in spoken systems and REG more generally.

2.1 The NATURAL approach

The NATURAL approach to spoken system design emphasizes powerful human abilities to produce and understand unrestricted natural human–human language. Natural language occurs in all cultures, and has co-evolved with human cognitive capacities. It is likely that natural language is optimized for human communication, and that deviations from natural language in spoken systems will bear costs for users. Moreover, people have a lifetime of practice talking to one another to draw on. It is possible that the complexity of human language is necessary for communicating what needs to be communicated for any even slightly complex domain. People may also simply prefer systems that are natural.

One strength of NATURAL spoken systems is their flexibility, and therefore their expressive power. A complete NATURAL spoken system would be able to handle anything in any domain, just as people do. It would not require users to learn, and it would be as easy to use as talking to another person, something we do effortlessly every day. There are also drawbacks to the NATURAL approach. One is the large gap between theory and practice. The goals may be enticing, but in the current state-of-the-art there are many errors and many open questions that need to be addressed. Addressing them requires behavioral research, coupled with system innovation. Even for well-understood domains, development cost is high; systems are generally developed by hand for each domain by experts in language and computation. Runtime cost also tends to be high, due to the complexity of natural language. Proponents of the NATURAL approach are aware of these issues, and are making progress on them (e.g. emphasizing domain-general architectures and code re-use). While they view these as important challenges, they also feel that they are acceptable costs, given the potential payoffs.

The NATURAL approach has implications for all aspects of spoken system design, including reference generation. Here we focus specifically on systems that are natural in the sense that they generate referring expressions that are dependent on discourse context. Even this limited case can introduce considerable complexity

https://doi.org/10.1017/S1351324910000227 Published online by Cambridge University Press

and cost into the design of spoken systems, which at a minimum must retain a history of the conversation, including all referring expressions and which objects they referred to. Given the costs, some people prefer to seek alternative methods that allow for context-independent referring expressions.

2.2 The STANDARDIZED approach

The STANDARDIZED approach emphasizes powerful human capacities for learning and adaptation, rather than the optimality of human language *per se*. Humans can adapt to a wide range of graphical user interfaces, provided the inputs and outputs are consistent. Even within human–human language, concise subset languages have developed for domains in which accurate communication is critical like air traffic control (Kittredge and Lehrberger 1982; Churcher, Atwell and Souter 1996). Subset languages for computational systems are attractive because they reduce the set of alternatives that must be considered while preserving some analogue with human–human language (Churcher, Atwell and Souter 1996; Sidner and Forlines 2002). Some have argued that even this link with human–human language is unnecessary. Instead, they predict, people will prefer systems that explicitly avoid sounding like humans, so they will be reminded that the system is a system and therefore it will have certain limitations (Schneiderman 1980).

One strength of STANDARDIZED systems is low development cost; systems can be quickly extended to new domains without requiring human–human data collection and/or experts in human language. For example, developers with no formal linguistic training were able to extend the Universal Speech Interface to a new domain in fifteen minutes (Toth *et al.* 2002). STANDARDIZED systems are also associated with lower run-time processor costs and lower memory costs. There are some drawbacks to the STANDARDIZED approach, too. There is a learning curve for users who do not have experience with the system. Even after training and experience, users continue to have problems with compliance especially in complex domains, as they do with human–human subset languages (Rantanen and Kokayeff 2002). Proponents of the STANDARDIZED approach are aware of these issues. They are working to reduce the learning curve, and to select standardized languages that are habitable, or easy for users to comply with. However, it is thought that once the optimal set of design principles has been identified, and interactions have been standardized across domains, these issues will shrink in importance.

The STANDARDIZED approach has implications for all aspects of system design, including reference generation. Here we focus specifically on systems that are standardized in the sense that they generate referring expressions that are consistent, regardless of discourse context. Because the references are used in all discourse contexts, and must always uniquely specify a single entity, they are generally overinformative by Gricean standards (e.g. the red enemy helicopter will always be called 'the red enemy helicopter' even when it is the only helicopter or the only enemy helicopter). In domains where errors are costly this overinformativity is often seen as an advantage because it is thought to reduce the likelihood of misunderstandings.

https://doi.org/10.1017/S1351324910000227 Published online by Cambridge University Press

3 Evaluating and comparing spoken systems

Spoken systems are sometimes compared in their entirety, as robust end-to-end implemented systems. More commonly, however, subcomponents are evaluated individually. Spoken systems are generally comprised of components for at least the following capabilities: speech recognition, parsing, dialogue management, generation, and speech synthesis. Three of these are involved in spoken output. The *dialogue manager* decides which message to convey and when, the *generation component* selects the specific wording for the message, and the *synthesizer* gives those words a voice. There are good reasons to evaluate these components individually. End-to-end systems often have many differences across modules, making it difficult to tell what aspects of the system contribute to evaluation outcomes (e.g. different dialogue management algorithms send different input to reference generation). Component-level evaluation allows much cleaner comparisons. Moreover, the individual components tend to operate somewhat independently within the end-to-end systems. Component-level evaluation allows developers to select and combine the best of each. For the generation of referring expressions specifically, component-level evaluation has been facilitated by organized shared-task challenges which allow groups of researchers to apply their algorithms to the same problem and compare results directly (Belz, Gatt, Reiter and Viethen 2007).

There are four categories of measures that are used evaluate generation components: task-based measures, subjective user ratings, statistical measures, and combined measures (for comprehensive reviews see Paraboni, van Deemter and Mastoff 2009; Reiter and Belz 2009). For *task-based measures*, users are asked to use the output of the generator to do some task (e.g. booking a flight from Phoenix to Rochester). The speed and accuracy of the task is used as a measure of system quality. In some domains, task-based performance can also be measured in terms of outcomes (e.g. learning in tutoring systems, health improvement in medical domains). Task-based measures are collected in the background, as users interact with the system normally. This contrasts with *subjective user ratings*, in which users are asked to respond to sentences like 'The system is easy to use' with a numeric rating (Hone and Graham 2001). Subjective ratings are costly to collect, and there is growing interest in *statistical measures*, which are less costly. Statistical measures compare generated output to corpora of human-generated language, similarly to the BLEU method in machine translation (Papineni *et al.* 2002). Finally, *combined measures* use statistical methods to combine results from the other methods. For instance, the PARADISE measure uses task completion rate and speed to infer usability ratings (Walker *et al.* 1998).

The present work compares usability across two fundamentally different approaches to system design, focusing specifically on different ways that referring expressions can be related to discourse context in spoken language. We expect that the usability of the systems may change over time, and that different contexts may be associated with different usability costs across systems. None of the measures we have discussed so far are well suited to this comparison. Task speed measures are biased in favor of the NATURAL approach, which supports the use of pronouns and therefore

https://doi.org/10.1017/S1351324910000227 Published online by Cambridge University Press

systematically generates shorter utterances.[1] Task-based accuracy is uninformative because we are comparing systems of equal quality. Therefore, accuracy should be uniformly high across systems. Outcome measures are not appropriate because a domain-general method is needed. Subjective user ratings at the level of the whole interaction are not fine-grained enough to allow us to investigate change over time, or the usability of individual utterances. It is possible to collect user ratings more frequently; however, in this case subjective user ratings at the level of individual utterances would be biased in favor of the STANDARDIZED approach because they would interrupt the discourse context with new verbal information. Corpus-based statistical measures are by definition biased in favor of the NATURAL approach because they are based on similarity to language produced by humans. Combined measures would be difficult to interpret because the measurements they combine are inappropriate. Because all of the existing measures are in some way inadequate for our purposes, the present comparison required us to move beyond these measures to find a direct, fine-grained measure of usability that was not biased toward either of the two approaches.

3.1 Dual-task comparison

To address the need for an unbiased, direct, fine-grained measure of usability we adopted the dual-task methodology. The dual-task methodology is based on many findings from psychology demonstrating that human cognitive resources – memory, attention, vision, central executive functioning, etc. – have limited capacity (Miller 1956; Cowan 2005). People can only track a certain number of moving objects at the same time, they can only pick out so many individual instruments in a piece of music, and they can only keep a certain number of grocery list items in mind. The capacity for each cognitive resource includes all tasks that a person is doing. If one task requires a lot of resources, people have little left over for other tasks. This explains why people crash more often when talking on mobile phones, even hands-free phones (Hanowski *et al.* 2006). The task of conversing uses cognitive resources which might otherwise be used for driving.

Limited-capacity cognitive resources are closely related to usability. A system that requires the user to focus all attention on simply understanding what is said would not be usable. It would require a context of use with no distractions, and it would not leave resources for the user to think about other aspects of the task. Even with a simple task like looking up movie times, users need to think about things other than simply understanding speech that is generated by the system (e.g. considering whether they can make it to the theater in time). Systems will be most usable if understanding the speech that they generate consumes the fewest cognitive resources for the user, leaving more for other things.

[1] There are ways to measure timing that are not as heavily biased toward the natural case (e.g. Reiter *et al.* 2005).

https://doi.org/10.1017/S1351324910000227 Published online by Cambridge University Press

The dual-task methodology is a general method for investigating how much of a cognitive resource is consumed by a given task, based on the limited-capacity observation (Norman and Bobrow 1975). It involves two tasks, a *primary task* and a *secondary task*, which are done at the same time. The primary task is the one of interest and the secondary task is a simple one that is well understood and might plausibly be combined with system use in the real world. Performance on the secondary task is used to measure the resource consumption of the primary task in different conditions. For evaluation of spoken systems, the primary task is interacting with the system (in NATURAL and STANDARDIZED conditions), and a good secondary task is something like watching to see if the gas light comes on in your car. When the primary task condition is a highly usable system it will require less attention, leaving more to do well on the secondary task.

Dual-task comparisons are common in cognitive psychology, applied psychology, and evaluation of multimodal interfaces. There have been some recent studies that apply it to spoken interfaces as well. However, it has not yet been applied specifically to evaluating *generation components* of spoken interfaces. The dual-task methodology is an ideal method for the current research because it has a direct link to a specific theoretical definition of usability that does not rely on subjective reporting. It is also sensitive to subtle differences between systems due to its fine temporal grain; secondary task events can occur often, and at theoretically relevant times, which allows for investigation of overall performance as well as information about how resource consumption changes over time. The method does all of this while avoiding disruption of the discourse context with other spoken or written information. It is also not biased toward either approach.

4 Our experiment

We used the dual-task paradigm to compare NATURAL and STANDARDIZED approaches, specifically as they relate to discourse context. The system generated simple directives (e.g. 'Put the big blue bowl to the left of the small yellow cup'). The users' primary task was to follow these instructions as quickly as possible, moving pictures to different locations on a 5×5 grid with a mouse. While doing this task, users also did the secondary task of monitoring a row of 'lights' in the graphical user interface (GUI). Whenever one of the nine lights flickered, users pressed a key on the keyboard as quickly as possible. This simple flicker-detection task is well-understood in psychology and it is a reasonable analogue to something one might do while using a spoken system in the real world (e.g. paying attention to traffic signals while following spoken driving directions from a GPS).

4.1 Method

We collected data from twenty adult participants from the University of Rochester community. All were native speakers of English who were not colorblind, deaf, or hard-of-hearing. At the start of the experiment participants read a set of instructions about the primary and secondary tasks. The instructions emphasized that both were equally important. Participants practiced each task separately and

https://doi.org/10.1017/S1351324910000227 Published online by Cambridge University Press

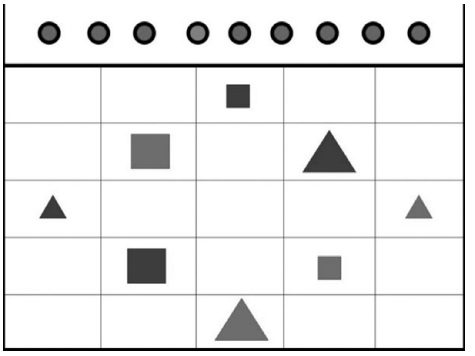


Fig. 1. A screen shot from the start of a trial. Each trial had a unique set of objects, and included three system-generated instructions.

then combined. Throughout the experiment accuracy and reaction times for both tasks were recorded.

Each participant followed a total of 408 directives generated by the system. These were divided up into 136 three-directive trials. Each trial was associated with a unique set of eight objects in the GUI (Figure 1). The objects were colorized (red, blue and green) and scaled (large and small) versions of a set of images that are widely used in psychology research (Snodgrass and Vanderwart 1980; Rossion and Pourtois 2001). Every set of eight objects contained two types of objects (e.g. bowl, cup). The size and color of these objects varied such that, without additional context, a unique reference to any object would need to include size, color, and object type. Participants could move objects and click on them at any point during the trial.[2] In order to minimize variance, both the tasks that participants were instructed to do and the directives generated by the system were predetermined and consistent for all participants. They were prerecorded by a female native speaker of American English to ensure fluency and natural prosody. Prosody can dramatically affect the interpretation of referring expressions during human–human interaction (Pierrehumbert and Hirschberg 1990). We used a human speaker rather than a speech synthesizer to ensure the experiment addressed the ideal form of referring expressions, rather than the current state-of-the-art in speech synthesis.[3]

Half of the participants interacted with the NATURAL version of the system and half of the participants interacted with the STANDARDIZED version. The NATURAL version generated human-like referring expressions that were dependent on discourse context. In contrast, participants in the STANDARDIZED condition heard instructions containing referring expressions that were consistent across all

[2] The timing of the instructions was not dependent on the timing of participant actions. This minimized the extent to which differences in the speed of the primary task would affect performance in the secondary task.

[3] The NATURAL instructions contained natural variations in stress, while the STANDARDIZED instructions were neutral and context-independent. This was accomplished by recording the NATURAL instructions in order within trials, and the STANDARDIZED instructions in random order.

https://doi.org/10.1017/S1351324910000227 Published online by Cambridge University Press

discourse contexts. The visual contexts and the set of actions the participants were asked to perform were identical across the two systems.

Over the course of the experiment participants were asked to move objects on the screen and to click on them. The actions and orders varied, to reduce task predictability. During half of the directives, one of the 'lights' at the top of the screen would change to a lighter color for 100 ms. Participants watched for these brief flickers and indicated when they had seen them by pressing a key. The presence or absence of these flickers could not be predicted from the actions, the visual context, or the discourse context. Embedded within the experiment were thirty-six Discourse Context (DC) trials that allowed us to examine three specific discourse contexts that are frequent and well documented in the literature. All of the DC trails were accompanied by a flicker in the secondary task. To prevent participants from predicting these flickers, we included an identical set of thirty-six filler trials that were not accompanied by flickers during the second sentence.[4]

The DC trials included three specific types of entities: discourse-new (NEW) entities, discourse-given but nonfocused (GF−) entities, and discourse-given and focused (GF+) entities. NEW entities have not been mentioned before. In human–human language they are referred to using a definite noun phrase that includes enough information to uniquely identify the referent given the full set of potential referents in the domain, but usually not more than that. GF+ have been mentioned before, and are currently in focus. In our experiment, GF+ references were always the direct object of the main verb 'put' in the previous instruction and in human–human language they would be referred to using the pronoun 'it'. GF− entities are in between the other two. They have been mentioned before, but they are not the most salient entity of the set of entities that have been mentioned before. In human–human communication, discourse-given but nonfocused entities are referred to using a definite referring expression that includes enough information to uniquely identify the referent given *the set of entities that have been previously mentioned*, but usually not more than that.

All of the DC trials had a similar structure. The first instruction directed the participant to place one object to the left of, to the right of, above, or below another object on the screen. The second instruction directed the participant to move an object to one of the corners or to the center. The third instruction directed the participant to click on an object not mentioned in previous instructions. The different types of DC trials differed only in the second instruction, which referred to a NEW, GF+, or GF− entity. The NATURAL and STANDARDIZED versions of the system also differed only in the reference that was generated to refer to this entity (see Table 1). All other sentences were identical. For NEW trials, there were no differences between the two systems. Secondary task flickers were timed to occur at the offset of the referring expression for all DC trials, to avoid bias and allow for a clearer interpretation of the data. For example, for a GF+ trial participants in the

[4] Flickers accompanied the first and third instructions on half of these trials as well. Flickers during the first and third instructions were equally distributed across the DC trials and the matched filler trials.

https://doi.org/10.1017/S1351324910000227 Published online by Cambridge University Press

Table 1. *Trial types for experiment*

Type	#*	NATURAL†	STANDARDIZED‡
GF+	1	Put the big red triangle to the right of the small blue square.	Put the big red triangle to the right of the small blue square.
	2	Now put **it** in the center.	Now put **the big red triangle** in the center.
	3	Now click on the small red square.	Now click on the small red square.
GF−	1	Put the big red triangle to the right of the small blue square.	Put the big red triangle to the right of the small blue square.
	2	Now put **the square** in the center.	Now put **the small blue square** in the center.
	3	Now click on the small red square.	Now click on the small red square.
NEW	1	Put the big red triangle to the right of the small blue square.	Put the big red triangle to the right of the small blue square.
	2	Now put **the small red triangle** in the center.	Now put **the small red triangle** in the center.
	3	Now click on the small red square.	Now click on the small red square.

* Three spoken directives appeared in order for each trial.
† The NATURAL system generated references that were dependent on discourse context, as in human–human communication.
‡ The STANDARDIZED system generated consistent references that were not dependent on discourse context.

NATURAL condition might hear the instruction 'Now put it in the center' while participants in the STANDARDIZED condition would hear 'Now put the big red triangle in the center'. For both groups the flicker was timed to coincide with the start of the word 'in'.

4.2 Results

Throughout the experiment, we recorded accuracy and reaction times for both the primary task (following directives generated by the system) and secondary task (detecting light flickers). As expected, primary task accuracy was high and did not differ significantly between NATURAL and STANDARDIZED versions of the system (see Table 2). Both design approaches are expected to result in systems that are accurate. Our research question relates to the costs of each in terms of cognitive load, assuming the ideal can be realized. Thus, equally good primary task performance simply demonstrates that there was no bias in the quality of the two systems. Primary task reaction times were marginally faster for the NATURAL system (see Table 2). This is not surprising because the NATURAL system generated shorter instructions. For instance, the NATURAL system generated the pronoun 'it' at times when the STANDARDIZED system generated a full definite noun phrase (e.g. 'the big red square'). These differences could explain faster reaction times in the primary task for the NATURAL case, yet it may not reflect a difference in usability. For this reason performance on the secondary task is more informative.

https://doi.org/10.1017/S1351324910000227 Published online by Cambridge University Press

Table 2. *Overall performance and ANOVA results*

		N $\mu(\sigma)$*	S $\mu(\sigma)$†	F(1,18)	P
Primary ALL	RT‡	10.33 (0.26)	11.00 (0.26)	3.3754	<0.10
	Accuracy§	88.0% (2.3%)	83.7% (3.1%)	n.s.	n.s.
Secondary ALL	RT‖	612 (20)	740 (30)	6.7953	<0.05
	Recall¶	0.982 (0.005)	0.946 (0.009)	11.3205	<0.01
	Precision#	0.912 (0.017)	0.845 (0.025)	4.9799	<0.05
Secondary DC**	RT	590 (20)	770 (50)	8.6401	<0.05
	Recall	0.98 (0.01)	0.92 (0.02)	n.s.	n.s.
	Precision	0.92 (0.01)	0.85 (0.02)	7.0682	<0.05

* Mean and standard error for NATURAL system.
† Mean and standard error for STANDARDIZED system.
‡ Reaction Time (RT) for the primary task includes all three instructions and is measured in seconds from the start of the first instruction. Lower numbers are better. Incorrect trials are excluded.
§ Accuracy in the primary task is the per cent of trials in which there were no errors in following any of the three instructions. Higher numbers are better.
‖ RT for the secondary task is measured in milliseconds from the onset of the flicker.
¶ Recall is the number of correct responses to flickers divided by the total number of responses (correct + false alarm). Higher numbers are better (fewer false alarms).
# Precision is the number of correct responses to flickers divided by the total number of flickers that occurred (correct + miss).
** DC includes only individual NEW, GF+, and GF− instructions, a more systematic measure of reference-related cognitive load.

As described in Section 3.1, secondary task performance provides information about the usability of different systems in the primary task. This is because attention is a limited-capacity resource. When the primary task is easy there is plenty of attention left over to perform well on the secondary task. When the primary task is difficult it consumes all of the user's attention, leaving little for the secondary task. Thus, performance on the secondary task suffers when the primary task is difficult. This is related to usability; systems that require more attention simply to understand what is being said are less usable; at the extremes, they require a context of use with no distractions, and they limit the users' ability even to do other aspects of the task at hand (e.g. considering options, planning, reasoning, speaking). In our analysis, we compared both accuracy and reaction time in the secondary task for NATURAL and STANDARDIZED versions of the system. By definition, higher accuracy and lower reaction times in the secondary task reflect lower cognitive load, and thus better usability, of the system in the primary task.

By this measure, our experiment provided clear evidence that the NATURAL system was easier to use than the STANDARD system. The NATURAL system was associated with higher accuracy and lower reaction times in the secondary task (see Table 2). This difference can be attributed directly to the use of human-like discourse references, because that was the only difference between the two systems.

https://doi.org/10.1017/S1351324910000227 Published online by Cambridge University Press

Table 3. *Practice and primary task accuracy*

	1st $\mu(\sigma)^*$	2nd $\mu(\sigma)^\dagger$	F(1,18)‡	P§
NATURAL	88% (4%)	88% (2%)	6.8728	<0.05
STANDARDIZED	87% (3%)	80% (4%)		

* Mean and standard error for the first half.
† Mean and standard error for the second half.
‡ F-values given correspond to the interaction between half and system. There were no main effects.
§ P-values given correspond to the interaction between half and system. There were no main effects.

Stronger evidence for that claim comes from the DC trials, which included just the discourse contexts that were of interest. We compared performance on the secondary task specifically for flickers that occurred in these discourse contexts, revealing an even larger advantage for the NATURAL system over the STANDARDIZED system. This supports our claim that the difference in speed and accuracy on the secondary task can be attributed to the relationship between discourse context and the form and content of referring expressions.

One question that immediately arises concerning these findings is the role of practice. People have a lot of practice with natural referring expressions from everyday human–human communication. They have less practice with context-independent referring expressions. With enough practice, would the cognitive load differences we observed disappear, or even reverse? Our results suggest it would not. With practice with the STANDARDIZED system, performance actually degraded. Table 3 shows the performance during the first and second halves of the experiment. There were no changes in secondary task performance over time for either system; however, there were differences in primary task accuracy. Participants in the NATURAL condition maintained consistently high accuracy, while participants in the STANDARDIZED condition got less accurate over time. We are certain that the two halves were identical because when designing the experiment we began by constructing two sets of 68 trials, each containing twelve NEW, twelve GF+, and twelve GF− trials, which differed only in that they used different item types (e.g. bowls, cups). During the experiment we presented these two sets of trials in different orders to different participants. Half of the participants in the NATURAL condition got Set A followed by Set B, and the other half of the participants got Set B followed by Set A. Likewise, for participants in the STANDARDIZED condition, Table 3 contains data from all of these participants.

Another question of interest is whether the effects we observed were due to just one discourse context. For instance, perhaps the advantage of the NATURAL system arose from the pronoun 'it'. We analyzed the results for NEW, GF+ and GF− discourse contexts separately; the NATURAL system had an advantage over the STANDARDIZED system for each context (see Table 4). Somewhat counterintuitively, in the NEW context the NATURAL and STANDARDIZED systems generated *exactly the same* set of instructions (e.g. 'the big red triangle'),

https://doi.org/10.1017/S1351324910000227 Published online by Cambridge University Press

Table 4. *Specific discourse context and secondary task performance*

		N	S	F(1,18)	P
GF+*	Recall	0.97 (0.01)	0.95 (0.02)	n.s.	n.s.
	Precision	0.95 (0.02)	0.84 (0.03)	6.5436	<0.05
	RT	570 (20)	740 (60)	5.4912	<0.05
GF–†	Recall	0.97 (0.02)	0.97 (0.01)	n.s.	n.s.
	Precision	0.94 (0.02)	0.86 (0.02)	5.5173	<0.05
	RT	610 (30)	790 (70)	3.3873	<0.05
NEW‡	Recall	1.00 (0.01)	0.98 (0.01)	n.s.	n.s.
	Precision	0.89 (0.02)	0.85 (0.03)	n.s.	n.s.
	RT	600 (30)	780 (50)	8.1318	<0.05

* Reference to the discourse-given entity that is currently in focus. NATURAL instructions use 'it' and STANDARDIZED instructions use a definite noun phrase.
† Reference to a discourse-given entity that is not in focus. STANDARDIZED instructions include more adjectives than NATURAL instructions.
‡ Reference to a discourse-new entity, one that has not been mentioned before. NATURAL and STANDARDIZED instructions are exactly the same.

yet participants in the NATURAL condition performed better on the secondary task. One explanation relates to Gricean implicature (Grice 1975). In natural language people tend to obey certain conversational maxims, including the maxim of brevity (i.e. be concise). Because of the systematicity of these maxims in natural communication, the meaning of referring expressions can go beyond literal content, implying a particular referential domain (i.e. the set of potential referents).

For instance, if people hear 'the big red. . . ', the implicature causes them to consider all big red items *except* those that can be described uniquely without mention of size or color (e.g. via 'the bowl' or 'it'), given the context. Implicatures are integral to meaning and are thought to be automatic (Levinson 2000). Overriding them may consume cognitive resources. This would explain our findings: As people gain experience with the STANDARDIZED system they learn that they must override these implicatures to complete the task. In the NEW context, upon hearing 'the small red . . . ,' participants in the STANDARDIZED condition must consciously override the implicature that the item is previously unmentioned, even though in this case that implicature is correct. Moreover, they must consider all eight objects as potential referents, while participants in the NATURAL condition can exclude some objects (those that have been previously mentioned).

5 Discussion and future work

Our findings suggest that in the long term it is worthwhile to invest research, development, and runtime resources in accurately generating human-like referring expressions in spoken systems, especially for domains that involve evolving discourse context. We demonstrated (1) that systems that generate referring expressions that are dependent on discourse context impose less cognitive load than systems that generate consistent referring expressions regardless of discourse context and (2)

https://doi.org/10.1017/S1351324910000227 Published online by Cambridge University Press

that this difference is not due to practice. This effect occurred across all three of the discourse contexts we examined, including the NEW context, which involved exactly the same speech for both context-dependent (NATURAL) and context-independent (STANDARDIZED) systems.

The processing difficulties we observed in the STANDARDIZED case were likely due in part to the introduction of temporary ambiguity, which was later resolved. However, it is important to remember that all of the referring expressions we examined were globally unambiguous. Practically speaking, this suggests that when it is possible to accurately generate such references, developers should choose a NATURAL approach. However, generating context-independent referring expressions may still be appropriate in situations where human speech is globally ambiguous (Reiter *et al.* 2005). It might also be appropriate when it is impossible to make accurate predictions about the set of objects listeners might be considering, given the state-of-the-art, yet a solution is needed in the near term. Developers should bear in mind that this choice may result in additional load for listeners that will not be reduced with practice. Future research (as described below) may improve the usability of such systems.

Although we investigated spoken language, not text, it is likely that our findings apply to text-based GRE as well. Local ambiguities introduced by conversational implicature have the effect of slowing down reading speeds and some readers can have difficulty overcoming them on a first read. Thus the NATURAL approach to GRE is likely to result in more usable text-based systems for applications in which reading slow-downs are problematic. However, for some applications additional processing load for the reader is not the only cost to consider. For instance, many controlled languages disallow the use of pronouns altogether (Reuther 2003; O'Brien 2003). This does not necessarily conflict with our findings. Many of these controlled language systems are designed to support machine translation (e.g. technical documents written in Chinese being automatically translated to English) and readability by non-native speakers (Reuther 2003). Cross-language variation in pronouns causes problems especially when translating from a language with an impoverished pronoun system to one with a rich pronoun system; additional properties of the referent need to be inferred in order to accurately select a pronoun for the target language. For these sorts of controlled languages some processing difficulty for the reader may justified, given the considerable benefits of more accurate machine translation. When readability is the primary concern, naturalness may have more of a role, in controlled languages as in other domains of NLG (Clark *et al.* 2009).

In terms of future work, the cognitive load measure we have introduced is a valuable addition to the toolbox of evaluation metrics. It could be used to investigate understanding, as well as generation, of referring expressions, and the match between generation and understanding. Understanding typically lags behind generation in spoken systems (e.g. a system might be able to generate 'it' in a context where it would not be able to identify the referent of 'it'). Is it better to generate the most natural referring expressions, even if the system cannot understand them? Or is a match between generation and understanding critical for usability?

https://doi.org/10.1017/S1351324910000227 Published online by Cambridge University Press

We would also like to investigate the usability implications of generating referring expressions that are based on other types of context, beyond discourse context. For instance, it has long been known that visual similarity of competitor objects affects natural referring expressions (Krauss and Weinheimer 1964; Krauss and Weinheimer 1966; Krauss and Weinheimer 1967). Task-based pragmatic context can also have effects (Campana, Brown-Schmidt and Tanenhaus 2002; Brown-Schmidt, Campana and Tanenhaus 2005; Viethen and Dale 2006). Do violations of natural patterns in these contexts lead to the same types of processing difficulties we observed with violations of discourse context? These are interesting questions because the contexts we examined in this work are associated with very systematic variation in referring form and content. Violations of these particular natural reference patterns are associated with strong, automatic implicatures that must be overcome. Other types of context may have a more graded effect on reference generation that may be easier for listeners to adapt to, with practice (Guhe and Bard 2008; Viethen *et al.* 2010). Would we still see processing differences due to differences in information content (e.g. context-dependent systems allow for probabilistic narrowing of the referential domain while context-independent ones do not)? If so, does *any* source of systematic context-based variation lead to improvements, or is it important that this variation match human variation?

Acknowledgements

This work was supported by NASA GSRP, NIH grant HD27206, NSF grant 0748942, and ONR grant N000140510314. We are indebted to Dana Subik for managing the data collection for this experiment, and to Stephanie Packard and Tom Covey for preparing stimuli.

References

Appelt, D. E. 1985. Planning English referring expressions. *Artificial Intelligence* **26**(1): 1–33.

Belz, A., and Varges, S. 2007. Generation of repeated references to discourse entities. In *Proceedings of the 11th European Workshop on Natural Language Generation (ENLG'07)*, pp. 9–16.

Belz, A., Gatt, A., Reiter, E., and Viethen, J. 2007. Referring expression generation (Chapter 3). In M. White and R. Dale (eds.), *Report of the Workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation*, pp. 17–30.

Brown-Schmidt, S., Campana, E., and Tanenhaus, M. K. 2005. Real-time reference resolution by nave participants during a task-based unscripted conversation. In J. Trueswell and M. K. Tanenhaus (eds.), *Approaches to Studying World-Situated Language Use*: *Bridging the Language-as-Product and Language-as-Action Traditions*, pp. 1–22. Boston, MA: MIT Press.

Brown-Schmidt, B. S., and Tanenhaus, M. K. 2008. Real-time investigation of referential domains in unscripted conversation: a targeted language game approach. *Cognitive Science* **32**: 643–684.

Campana, E., Brown-Schmidt, S., and Tanenhaus, M. K. 2002. Reference resolution by human partners in a natural interactive problem-solving task. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP)*, pp. 2057–2060, Denver, CO, USA.

https://doi.org/10.1017/S1351324910000227 Published online by Cambridge University Press

Churcher, G., Atwell, E., and Souter, C. 1996. Dialogues in air traffic control. In S. Luperfoy, A. Nijholt and G. Veldhuijzen van Zanten (eds.), *Dialogue Management in Natural Language Systems, Proceedings of the 11th Twente Workshop on Language Technology*.

Clark, H. H. 1992. *Arenas of Language Use*. Chicago, IL: University of Chicago Press.

Clark, P., Harrison, P., Murray, W. R., and Thompson, J. 2009. Naturalness vs. predictability: a key debate in controlled languages. *Presented at the Workshop on Controlled Natural Language (CNL 2009)*, Marettimo Island, Italy.

Clark, H. H., and Wilkes-Gibbs, D. 1986. Referring as a collaborative process. *Cognition* **22**: 1–39.

Cowan, N. 2005. *Working Memory Capacity*. Hove, East Sussex, UK: Psychology Press.

Croitoru, M., and Kees van Deemter 2007. A conceptual graph approach to the generation of referring expressions. In *Proceedings of the 2007 International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 2456–2461, Hyderabad, India.

Dale, R. 1988. *Generating Referring Expressions in a Domain of Objects and Processes*. PhD dissertation, Edinburgh, UK: University of Edinburgh.

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

DeVault, D., Rich, C., and Sidner, C. L. 2004. Natural language generation and discourse context: computing distractor sets from the focus stack. In *Proceedings of the International Florida Artificial Intelligence Research Symposium (FLAIRS)*.

Dickson, P. 2008. *Baseball's Greatest Quotations: An Illustrated Treasury of Baseball Quotations and Historical Lore (Revised Edition)*. New York: HarperCollins.

Hanowski, R. J., Olson, R. L., Hickman, J. S., and Dingus, T. A. 2006. The 100-car naturalistic driving study: a descriptive analysis of light vehicle-heavy vehicle interactions from the light vehicle drivers perspective. Technical Report FMCSA-RRR-06-004, National Technical Information Service, US Department of Transportation.

Gatt, A., and van Deemter, K. 2007. Lexical choice and conceptual perspective in the generation of plural referring expressions. *Journal of Logic, Language and Information (JoLLI)*, **16**(4): 423–443.

Gatt, A., and van Deemter, K. 2009. Generating plural NPs in discourse: evidence from the GNOME corpus. In *Proceedings of the CogSci 2009 Workshop on the Production of Referring Expressions: Bridging the Gap between Computational and Empirical Approaches to Reference (PRE-CogSci 2009)*.

Greenbacker, C. F., and McCoy, K. F. 2009. Feature selection for reference generation as informed by psycholinguistic research. In *Proceedings of the CogSci 2009 Workshop on the Production of Referring Expressions: Bridging the Gap between Computational and Empirical Approaches to Reference (PRE-CogSci 2009)*.

Grice, H. P. 1975. Logic and conversation. In P. Cole and J. L. Morgan (eds.), *Syntax and Semantics, Volume 3*: *Speech Acts*, pp. 43–58. New York: Academic Press.

Grosz, B. J., Joshi, A. K., and Weinstein, S. 1995. Centering: a framework for modeling the local coherence of discourse. *Computational Linguistics* **21**: 203–225.

Grosz, B. J., and Sidner, C. L. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics* **12**(3): 175–204.

Guhe, M. 2009 Generating referring expressions with a cognitive model. In *Proceedings of the CogSci 2009 Workshop on the Production of Referring Expressions: Bridging the Gap between Computational and Empirical Approaches to Reference (PRE-CogSci 2009)*.

Guhe, M., and Bard, E. G. 2008. Adapting referring expressions to the task environment. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society*. pp. 2404–2409. Austin, TX.

Gundel, J. K., Hedberg, N. and Zacharski, R. 1993. Cognitive status and the form of referring expressions. *Language* **68**: 274–307.

Hajičová, E. 1993. *Issues of Sentence Structure and Discourse Patterns*. Prague, Czech Republic: Charles University Press.

https://doi.org/10.1017/S1351324910000227 Published online by Cambridge University Press

Hajičová, E., Sgall, P. and Havelka, J. 2003. Discourse semantics and the salience of referents. *Journal of Slavic Language*: 127–140.

Hone, K. S., and Graham, R. 2000. Towards a tool for the subjective assessment of speech system intefaces (SASSI). *Natural Language Engineering* **6**(3–4): 287–303.

Kittredge, R. and Lehrberger, J. 1982. *Sublanguage: Studies of Languages in Restricted Semantic Domains*. New York: Walter de Gruyter.

Krahmer, E. and Theune M. 2002. Efficient context-sensitive generation of referring expressions. In K. van Deemter and R. Kibble (eds.), *Information Sharing: Givenness and Newness in Language Processing*, pp. 223–264. Stanford, CA: CSLI Publications.

Krauss, R. M., and Weinheimer, S. 1964. Changes in reference phrases as a function of frequency of usage in social interaction: a preliminary study. *Psychonomic Science* **1**: 113–114.

Krauss, R. M., and Weinheimer, S. 1966. Concurrent feedback, confirmation and the encoding of referents in verbal communication. *Journal of Personality and Social Psychology* **4**: 343–346.

Krauss, R. M., and Weinheimer, S. 1967. Effects of referent similarity and communication mode on verbal encoding. *Journal of Verbal Learning and Verbal Behavior* **6**: 359–363.

Levinson, S. 2000. *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. Cambridge, MA: MIT Press.

Miller, G. A. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* **63**: 81–97.

Norman, D. A., and Bobrow, D. G. 1975. On data-limited and resource-limited processes. *Cognitive Psychology* **7**: 44–64.

O'Brien, S. 2003. Controlling controlled english: an analysis of several controlled language rule sets. In *Proceedings of the Joint Conference combining the 8th International Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop (EAMT-CLAW 2003)*.

Olson, D. 1970. Language and thought. Aspects of a cognitive theory of semantics. *Psychological Review* **77**: 257–273.

Papineni, K., Roukos, S., Ward, T., and Zhu, W. J. 2002. BLEU: a method for automatic evaluation of machine translation. In *40th Annual meeting of the Association for Computational Linguistics (ACL)*, pp. 311–318. Philadelphia, PA, USA.

Paraboni, I., Van Deemter, K., and Masthoff, J. 2007. Generating referring expressions: Making referents easy to identity. *Computational Linguistics* **33**(2).

Pechmann, T. 1989. Incremental speech production and referential overspeciification. *Linguistics* **27**: 89–110.

Pierrehumbert, J. and Hirschberg, J. 1990. The meaning of intonational contours in the interpretation of discourse. In P. R. Cohen, J. Morgan and M. E. Pollack (eds.), *Intentions in Communication*, pp. 271–311. Cambridge, MA: MIT Press.

Reiter, E., and Belz, A. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics* **35**(4): 529–558.

Reiter, E. and Dale, R. 2000. *Building Natural Language Generation Systems*, pp. 144–158. Cambridge, UK: Cambridge University Press.

Reiter, E., Sripada, S., Hunter, J., Yu, J., and Davy, I. 2005. Choosing words in computer-generated weather forecasts. *Artificial Intelligence* **167**: 137–169.

Reuther, U. 2003. Two in one – Can it work? Readability and translatability by means of controlled language. In *Proceedings of the Joint Conference combining the 8th International Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop (EAMT-CLAW 2003)*.

Rantanen, E. M., and Kokayeff, N. K. 2002. Pilot error in copying air traffic control clearances. In *Proceedings of the 46th Annual Meeting of the Human Factors and Ergonomics Society*.

https://doi.org/10.1017/S1351324910000227 Published online by Cambridge University Press

Rossion, B., and Pourtois, G. 2001. Revisiting Snodgrass and Vanderwart's object database: color and texture improve object recognition. *Journal of Vision* **1**(3): 413, 413a.

Schneiderman, B. 1980. Natural vs. Precise concise language for operation of computers: Research issues and experimental approaches. In *Annual Meeting of the Association of Computing Machinery (ACL)*, pp. 139–141, Philadelphia, PA, USA.

Sidner, C. L., and Forlines, C. 2002. Subset languages for conversing with collaborative interface agents. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP)*, pp. 281–284, Denver, CO, USA.

Snodgrass, J. G., and Vanderwart, M. 1980. A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory* **6**: 174–215.

Toth, A. R., Harris, T. K., Sanders, J., Shriver, S., and Rosenfeld, R. 2002. Towards every-citizen's speech interface: an application generator for speech interfaces to databases. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP)*, pp. 1497–1500, Denver, CO, USA.

van der Sluis, I., and Krahmer, E. 2007. Generating multimodal references. *Discourse Processes* **44**(3): 145–174.

Viethen, J., and Dale, R. 2006. Algorithms for generating referring expressions: do they do what people do? In *Proceedings of the 4th International Conference on Natural Language Generation*, pp. 63–70, Sydney, Australia.

Viethen, J., Zwarts, S., Dale, R., and Guhe, M. 2010. Generating subsequent referring expressions in a visual domain. In *Proceedings of the 7th Language Resources and Evaluation Conference (LREC)*. Malta.

Walker, M. A., Joshi, A. K., and Prince, E. 1998. *Centering Theory in Discourse*. Oxford, UK: Oxford University Press.

Walker, M. A., Litman, D. J., Kamm, C. A., and Abella, A. 1998 Evaluating spoken dialogue agents with PARADISE: two case studies. *Computer Speech and Language* **12–3**.

https://doi.org/10.1017/S1351324910000227 Published online by Cambridge University Press