Are speakers and listeners ‘only moderately Gricean’? An empirical response to Engelhardt et al. (2006)

Catherine Davies & Napoleon Katsos

This paper appeared as:

Davies, C. & Katsos, N. (2013) Are speakers and listeners ‘only moderately Gricean’? An empirical response to Engelhardt et al. *Journal of Pragmatics* 49 (2013) 78-106
http://dx.doi.org/10.1016/j.pragma.2013.01.004

Abstract

If hearers are sensitive to Gricean maxims of Quantity (Grice, 1975/1989), they should disfavour expressions which give too little or too much information for the unique identification of an intended referent. Accordingly, cooperative speakers are expected to provide all and only as much information as is necessary for their interlocutor to uniquely identify a referent. Engelhardt et al. (2006) report that speakers and hearers are sensitive to under-informativeness but not to over-informativeness. Based on this finding, the authors re-interpret the literature which claims to document pragmatic effects in language comprehension and instead attribute previous findings to structural-lexical biases. We argue that the reason why speakers and hearers seemed insensitive to over-informativeness in Engelhardt et al.’s studies was because certain aspects of their experiments favoured the use of redundant information. Our experiments 1 and 2 manipulate these factors, revealing that hearers are in fact sensitive to violations of over- as well as under-informativeness. A further production experiment shows that speakers do not under- or over-specify when the factors that favoured over-informativeness in Engelhardt et al.’s study are removed. The findings provide evidence that speakers and hearers are sensitive to both Quantity maxims, and suggest that the effects obtained in previous literature should indeed be attributed to pragmatic factors.

1. Introduction

When referring to a specific entity, speakers have a wide choice of potential forms with which to direct an interlocutor’s attention to their intended target. The choice of referring expression (RE) reflects not only properties of the target referent, but also contrasting properties of other entities in the physical or linguistic context (Clark and Bangerter, 2004; Clark et al., 1983; Pechmann, 1984, 1989). For unambiguous reference, an RE must be contrastive, containing enough information to allow a hearer uniquely to pick out the intended referent from other candidates in the context. In this way, cooperative speakers adhere to the first Gricean maxim of Quantity, which stipulates that utterances should be as informative as is required (Grice, 1975/1989), and also to the submaxim of Manner 'avoid ambiguity'. For example, a speaker would not be expected to refer to one of two mutually available apples as 'the apple'. In such a situation, a hearer would struggle to establish reference without the provision of contrastive modification.

What happens when speakers go beyond using minimally contrastive expressions and include additional modifiers in their REs? Expressions that provide more information than the communicative situation requires should still succeed in picking out a target referent. However, side-effects can result from the use of overspecifying expressions. Grice proposed a second maxim of Quantity stating that speakers should provide no more information than is needed for the purposes of the exchange. Grice conceded that this second maxim is disputable (1989:26), but added that a hearer may be misled into thinking that there is a particular point to the provision of excess information. Empirically, inferential effects have been shown to arise routinely in the processing of modified referring expressions (Sedivy, 2003; Sedivy et al., 1999, i.a.). For example, on hearing the RE 'the big apple', hearers fixate more quickly on the target in the presence of a contrast mate (i.e. a smaller apple) as they infer a rational motivation for the speaker's inclusion of the modification.

The question of whether speakers and hearers are indeed sensitive to under-informativeness (which derives from Grice's maxim of Quantity-1, hereafter Q-1) and over-informativeness (from Quantity-2, hereafter Q-2) is still under debate. This issue not only has implications for the psychological validity of pragmatic maxims, but also for the wider field of sentence processing. In particular, it is relevant to how non-syntactic constraints influence the interpretation of syntactically ambiguous constructions: potentially over-informative utterances such as 'put the apple on the towel in the box' have been widely used in investigations into the effect of referential context on incremental processing (Tanenhaus et al., 1995, i.a.). Should sensitivity to over-informativeness be found, what is its influence on sentence processing relative to other pragmatic constraints such as unique referent identification (governed by Q-1)? For example, if an RE uniquely identifies a referent, does its over-informativeness concede a processing penalty? Do nonlinguistic features such as attribute salience modulate the effect of over-informativeness? For instance, would REs such as 'the large apple' in contexts containing a single but saliently large apple cause the same inferential effects as when the apple was not saliently large? In other words, how strongly do Gricean expectations about optimal informativeness constrain comprehension and production, relative to other linguistic and nonlinguistic features?

A promising approach to this question is to measure off-line sensitivity to informativeness, and explore how this is affected by the presence or absence of relevant contextual features. In order to monitor the impact of nonlinguistic features such as salience or referential complexity, sensitivity to informativeness should initially be measured in baseline contexts without these marked features. Sensitivity should then be re-tested with contextual manipulations which add these potentially competitive constraints. The studies in this paper examine informativeness sensitivity in two distinct visual worlds (one complex, one simple) in order to explore how non-pragmatic constraints affect Gricean sensitivities. The results have implications for future research into the interplay of constraints on the form of referring expressions produced, on the associated judgement of pragmatic felicity, and on referential comprehension more widely.

The primary research question driving this paper investigates whether adults are sensitive to Q-1 and Q-2. First we present a brief review of the overarching theoretical framework and the links between the Gricean view of conversation and constraints on referential processing. We then discuss the existing literature on the comprehension and production of under- and over-informativeness. We report three new experiments: two which measure hearer sensitivity to the Gricean maxims of Quantity and one which measures speaker sensitivity. In the general discussion, we explore the implications of these results with reference to the existing literature, and identify three possible accounts of the role of Gricean considerations in referential processing. Although the current experiments cannot conclusively adjudicate between these possibilities, they provide input to future research which could address the positions more directly.

1.1. Constraint-based views of language processing

Constructions such as ‘put the apple on the towel in the box’ are incrementally ambiguous: the first prepositional phrase [PP] can momentarily function either as a destination or as a modifier of the head noun. Correct on-line parsing of such expressions has been argued to rely on expectations of context-dependent informativeness. That is, constraints from the referential context (e.g. whether there are multiple apples in the array) interact with lexical and syntactic biases to determine the most likely interpretation on-line (see Altmann, 1998 for a review). This mechanism is at the heart of the longstanding psycholinguistic debate over whether sentence processing uses structural (mainly syntactic) constraints prior to other types of linguistic and contextual information, or whether multiple sources of information influence the parser from the earliest stages of processing (Boland et al., 1990; Ferreira and Clifton, 1986; Frazier and Rayner, 1982; Frazier, 1987; MacDonald, 1994; Tanenhaus et al., 1989; Taraban and McClelland, 1988; Trueswell et al., 1993, i.a.).

For the interpretation of ambiguous sentences, the modular account asserts that initial interpretations are solely based on analyses of underlying syntactic representations. The syntactic module is ‘informationally encapsulated’ from other modules (such as referential or statistical sources of information) and no other knowledge is available to the parser until later in processing (Frazier, 1987). Conversely, on the interactive view, multiple sources of information interact and constrain interpretation from the very first stages of processing. This includes information from the extra-sentential context (e.g. the preceding discourse context or the visual referential world), the frequency of each interpretation in the language at large, semantic knowledge, thematic biases for each of the critical elements in the string, and frequency of argument structure, amongst other factors. This approach posits that multiple interpretations are activated in parallel, with activation of the various candidates dynamically changing as the sentence unfolds, shaping the eventual interpretation.

We consider that Gricean considerations should be added to the list of potential constraints shaping utterance interpretation. Provided that interlocutors are sensitive to informativeness, then if one interpretation of an ambiguous utterance is under- or over-informative, this is less likely to be the intended interpretation. For example, if one hears 'put the apple on the towel. . .' in a context containing one apple, the PP-as-modifier interpretation is over-informative, and hence the PP-as-destination interpretation is more likely to be preferred. Within the constraint-based approach, our account predicts that Gricean maxims of Quantity influence hearers' parsing decisions.

Gricean sensitivity to informativeness might in principle influence sentence processing to a greater or less extent: we might broadly characterise it as all-powerful, having no influence, or being one constraint among many. These positions are evaluated further in section 6.

1.2. Previous research on sensitivity to under- and over-informativeness

As the two submaxims of Quantity interact asymmetrically with reference resolution constraints, research on sensitivity to these submaxims must consider them separately. In the referential communication paradigm, under-informative REs such as 'the apple' in the presence of two apples not only violate Q-1, but in doing so also fail to establish unique reference. By contrast, over-informative REs such as 'the big apple' in contexts containing a single apple only fail on one of these counts: they violate Q-2, but do allow the hearer to uniquely identify the intended target. The original study which experiment 1 aims to replicate (Engelhardt et al., 2006; expt. 2) found that hearers were not sensitive to Q-2 (but were to Q-1) and thus argued that interlocutors are 'only moderately Gricean' (Engelhardt et al., 2006:572). Should sensitivity to both maxims be found, that would strengthen the claim that Gricean maxims have a powerful influence on speaker and hearer behaviour. Of course, arguing for full Gricean sensitivity in more general contexts would require a more comprehensive review of the experimental literature in such areas as scalar implicature (as well as further research into the impact of the remaining maxims). Several studies in this domain report that under-informative descriptions (e.g. saying that 'some of the horses jumped over the fence' when all of the horses did) are predominantly rejected by adults at rates that range from 60% (Noveck, 2001) to over 85% (Papafragou and Musolino, 2003; Katsos and Bishop, 2011) depending on the task. In the referential communication paradigm, Engelhardt et al. (2006) report that under-informative descriptions are consistently penalised compared to optimally informative expressions. Contrastingly, the single study performed on over-informative utterances (Engelhardt et al., 2006, experiments 2a and 2b) found that hearers do not rate them as any worse than optimal expressions, even though participants in the same task rated under-informative utterances as significantly worse than their optimal counterparts.

In Engelhardt et al.'s ratings paradigm, participants were shown photographs depicting scenes before and after a change had taken place, and judged spoken utterances that described the change. In a typical item for the one-referent condition, the first picture showed an apple on a towel, a puppet, a towel with nothing on it, and an empty box. In the second picture, the apple had been moved either to the matching location (from towel to towel), or to a different location (from towel to box). In the corresponding item for the two-referent condition, the first picture showed two apples (one on a towel, and one by itself), a towel with nothing on it, and an empty box. Again, in the second picture, the apple on the towel had been moved either onto the other towel or into the empty box. The participants' task was to judge whether the sentence uttered was an appropriate one to bring about the change between the two pictures. The sentence either contained a modifying PP after the target noun ('put the apple on the towel on the other towel'), or just the location with no modification ('put the apple on the other towel'). The resulting 2 2 design (sentence condition) thus contains under-informative, optimal, and over-informative instructions. Their results (expt. 2b) showed that hearers judged under-informative instructions to be significantly worse than their optimal counterparts, but they did not judge over-informative utterances to be any worse than the corresponding optimal expressions.

The authors conclude that hearers are 'only moderately Gricean in their adherence to the Maxim of Quantity' (p. 572) in that they are sensitive to under- but not to over-informativeness, at least in this off-line rating task. This conclusion is hard to reconcile with the Gricean account which predicts some leniency but not full tolerance or insensitivity towards over-informativeness. It is also surprising given the findings of their third experiment, which found costs in the on-line processing of over-informative expressions. In this experiment, participants' eye movements were recorded as they moved objects in response to felicitous and infelicitous instructions. The same instructions were used as in the preceding experiment, but only 1-referent contexts were tested as the authors were not investigating under-informativeness any further. Based on the previous results, the authors predicted no difference in fixation patterns between the optimal ('put the apple in the box') and over-informative ('put the apple on the towel in the box') utterances. However, differences were found between the two conditions regarding looks to the target, to the correct destination, and especially to the incorrect destination, where hearers looked to the empty towel on hearing 'on the towel'. This suggests that, in the absence of a contrast set (i.e. another apple, not on a towel), participants interpret the PP 'on the towel' as a destination rather than as a modifier of the target, since Gricean expectations lead them to reject the latter interpretation in which the utterance is over-informative. However, in the absence of sensitivity to over-informativeness in their off-line ratings study, the authors reject the pragmatically oriented explanation of the on-line experiment, and attribute these results to structural constraints alone[1].

[1] There are concerns about the comparability of these off-line and on-line comprehension experiments. One reason for the different findings may lie in the form of the required response, where the former involves metalinguistic reflective judgment on given, static stimuli and the latter necessitates a direct, action-based response. Drawing from the research tradition proposing that language processing tightly interacts with events unfolding in the real world (see e.g. Tanenhaus et al., 1995 and references therein), the metalinguistic task may not have tapped into participants' linguistic competence in the same way that the action-based one did, casting doubt on Engelhardt et al's unification of results under one theoretical explanation (see Altmann and Kamide, 1999, for differences in implicit performance based on whether participants were concurrently engaging in metalinguistic judgements or not).

Specifically, they invoke Minimal Attachment to explain the building of the simpler syntactic structure compatible with the destination interpretation (as assumed in the Garden Path model of parsing; Frazier, 1978). This account removes the need to consult any referential or pragmatic sources of information and so accommodates the apparent lack of maxim sensitivity in experiments 1 and 2. Engelhardt et al. reinforce this structural explanation by positing powerful influences from the verb argument structure of 'put', which requires a location-prepositional phrase. They argue that the parser saturates its argument structures as quickly as possible, and hence, upon hearing 'put', hearers rapidly interpret the following PP as a destination, which is consistent with the eye movement data. Crucially, neither of these explanations involves pragmatic reasoning about the amount of information to be conveyed. Engelhardt et al. conclude that their study raises doubts about whether well-supported effects in sentence processing should indeed be attributed to pragmatic considerations.

Despite Engelhardt et al.'s claims, on-line effects of visual/referential context are well documented, and are compatible with Gricean accounts of processing. Two types of studies are particularly relevant. The first concerns the study of temporarily ambiguous sentences, and in particular sentences where a PP is ambiguous between a modifier or a destination interpretation (as Engelhardt et al. used). A robust finding is that sentences such as 'put the apple on the towel in the box' cause garden-pathing, as hearers interpret the first PP 'on the towel' as the destination, and have to re-analyse this interpretation upon parsing the second PP. However, this garden-path is robust only in cases involving one referent. In the presence of a contrast set, the garden-path disappears (see e.g. Tanenhaus et al., 1995; Trueswell et al., 1999; Spivey et al., 2002, i.a.). This effect is explained by Referential Theory (Altmann and Steedman, 1988; Crain and Steedman, 1985) which postulates that the preferred parse of a sentence is the one which minimises the presuppositions (in this case, the number of entities that have to be additionally postulated beyond the ones that are explicitly encoded). Referential Theory is also compatible with Grice's Quantity maxims: if the relevant visual world contains only one apple, a modifier is unnecessary for the identification of the referent. To avoid a dispreferred, over-informative interpretation (which would be felicitous only if an additional apple were available) the PP is correctly analysed as a destination. In contrast, if the visual world contains two entities of the same type, then modification is necessary for resolving reference. Hence, the preferred interpretation of the PP in 2-referent contexts is as a modifier. Experimental evidence for the reality of this distinction supports the theory-critical conclusion that interlocutors use Gricean-like considerations in incremental sentence processing.

In a second strand of research documenting the effects of over-informativeness, Sedivy et al. (1999) and Sedivy (2003) consider visual displays including both a contrast set and a competitor, e.g. a big and a small glass, and a big jug. They show that, upon hearing 'big', hearers fixate on the big glass before they hear the disambiguating noun ('glass' or 'jug'). The proposed explanation for this effect is that hearers are engaging in Gricean inferencing, reasoning that if the speaker meant to refer to the big jug, they would have referred to it without a modifying adjective in order to avoid over-informativeness. The use of a modifying adjective thus signals that the intended referent is the one for which there is a contrast-mate, i.e. the big glass. Omission of the modification in that case would have led to under-informativeness. These results have been replicated by Hanna et al. (2003), Heller et al. (2008) and Grodner and Sedivy (2011).

Taken together, the findings from post-modifying PPs and pre-modifying adjectives demonstrate effects from pragmatic constraints in incremental on-line processing regardless of whether or not sentences are syntactically ambiguous. The findings from Engelhardt et al.'s on-line experiment 3 accord straightforwardly with these studies: if hearers are able to reason about referential intentions based on Quantity considerations, then they should also experience a penalty (in this case, looking at the incorrect destination) when processing explicitly over-informative descriptions.

However, Engelhardt and colleagues argue that their on-line findings cannot be attributed to sensitivity to pragmatic constraints, since hearers did not penalise over-informative descriptions in their off-line ratings experiment (2b). They also observe production tendencies to acknowledge the presence of two location objects with an expression such as 'other' (influencing PPs without 'other' to be interpreted as modifiers). On these grounds, Engelhardt et al. appeal to Minimal Attachment and verb argument structure to explain PP interpretation in their on-line data, as discussed above.

This conclusion has implications for the wider debate about the nature of the linguistic parser, concerning whether processing is initially structure-based or whether it is interactive from an early stage, as discussed in section 1.1. While the interpretation of the on-line findings given by Tanenhaus, Trueswell, Spivey, Sedivy and colleagues falls squarely within the latter kind of account, Engelhardt et al.'s analysis is in line with the former approach, where information based on pragmatic considerations is not available at the earliest stage of language processing and instead lexical-structural information takes priority.

Leaving aside for a moment the evidence to the contrary from Tanenhaus and colleagues, Engelhardt et al.'s analysis leaves us with the possibility that Gricean considerations are philosophical abstractions rather than psychologically real constraints on sentence processing. In their view (Engelhardt et al., 2006:569), speakers and hearers avoid under-descriptions that might lead to confusion about referent identity. This emphasis on unique identification appears to motivate sensitivity to under-informativeness based on mechanistic reference establishment rather than on wider Gricean pragmatics.

There are two broad reasons why this conclusion should be contested. First, research on Quantity implicature shows that, given a rich story context and ample time to respond, participants are very good at detecting under-informativeness even when this does not lead to reference assignment failure (e.g. Guasti et al., 2005, experiment 4; Papafragou and Musolino, 2003; Katsos and Bishop, 2011). Secondly, the complexity of the visual array used by Engelhardt et al. may have masked or blocked the application of Gricean constraints. In this scenario, visual processing has to deal with numerous items in a single array, potentially leaving fewer resources available for Gricean reasoning. This is a plausible factor given that sensitivity to under-informativeness decreases with available processing resources in the case of

Quantity implicatures (see Bott and Noveck, 2004; De Neys and Schaeken, 2007). Moreover, the visual complexity may itself have motivated over-informing. Several researchers have noted that increased array complexity biases speakers towards giving more information (Koolen et al., 2009; Paraboni et al., 2007; Pechmann, 1989). This arises for both hearer-directed reasons and speaker-directed reasons. According to the principle of distant responsibility (Clark and Wilkes-Gibbs, 1986), there is a pragmatic motivation for over-informing if referential clarity is at risk. This principle states that the amount of information provided by speakers depends on their estimation of the hearers' potential for misunderstanding. Over-informing might also be motivated by the speaker's ease of computation: the Minimal Redundancy hypothesis (Freedle, 1972) states that as the number of dimensions used to construct referents increases, and the number of nonreferents increases, speakers are more likely to give redundant messages. This is linked to a principle of economy, in that speakers expend less effort by encoding a redundant attribute than they would by evaluating its distinctiveness and then suppressing it should it turn out to be noncontrastive (Pechmann, 1989; Whitehurst, 1976). Producing minimally contrastive expressions is efficient in terms of linguistic material yet costly in terms of cognitive effort. In Engelhardt et al.'s task, array complexity could feasibly have led participants to build redundancy into their message (in expt. 1) and to expect a certain degree of redundancy (in expts. 2 and 3).

As well as array complexity, Engelhardt et al.'s design contains other factors that may have promoted over-informativeness. Their experimental contexts featured more than one identical object (such as the source and destination containers, e.g. the towels, in their experiment). Meanwhile, the target object was the only compositional object in the array (i.e. consisting of an object placed on another object). Consequently, hearers may have expected an overspecified RE by dint of the increased salience of both the multiply-occurring container and the compositional container-object (an instance of a speaker-oriented effect). Such an effect was documented by Carbary and Tanenhaus (2007), where the presence of a particular feature elsewhere in the array increased mentions of the same redundant feature in references to the target. In fact, over-informing as a result of visual salience can have a beneficial communicative effect. For example, when a partially discriminative dimension is more easily perceived than a fully discriminating one, overspecified REs lead to faster identification times (Arts, 2004; Arts et al., 2010; Mangold and Pobel, 1988; Paraboni et al., 2007). In Engelhardt et al.'s design, the compositional object (e.g. the apple on the towel) may have been the most salient object in the 'before' photograph, leading interlocutors to produce or expect an exhaustive RE such as 'the-apple-on-the-towel'. These considerations may go some way towards explaining the 'surprisingly common' rate (30%) of over-descriptions observed in Engelhardt et al.'s production experiment (expt. 1). Similarly, in expt. 2, the hearers may not have penalised the over-informative instructions either because they were aware of the pressures on the speaker or because they found the 'redundant' information helpful.

Finally, methodological factors could also be at play in Engelhardt et al.'s study. In their ratings experiment, participants were asked to view scenes before and after a change had taken place and rate spoken utterances on a scale of 1--5. A rating of 1 on this scale was an 'incorrect' instruction (for example, a wrong category label, as featured in half of the control items, or an under-informative referring expression), 3 indicated an 'adequate' instruction and 5 an instruction that 'could not be better'. Such instructions focus strongly on the semantic truth-values of the utterances and hence may have masked pragmatic nuances in the stimuli. Furthermore, such a scale may have encouraged participants to rate any utterance which was more than minimally contrastive as 'more than adequate', disposing them to give over-informative utterances a score of 4 or 5.

Overall, the literature suggests a strong likelihood that hearers give favourable ratings to speakers who over-inform when (i) the discourse context is complex, (ii) the relevant property is salient, or (iii) the instructions given bias them to do so. The constraints favouring over-informativeness vary in origin and orientation: the principle of distant responsibility and the maxim of Manner are pragmatic norms of conversation and hearer-oriented in nature, whereas the salience constraint is perceptual and most likely speaker-oriented. Critically, it is not clear whether the lack of a penalty for the over-informative utterances in Engelhardt et al.'s off-line study reflects a lack of sensitivity to over-informativeness (as the authors claim) or an actual preference/tolerance for over-informativeness when these factors are at play. The former explanation denies the effect of Gricean maxims in comprehension, whereas the latter argues that these aspects of Gricean reasoning are present but overruled by other constraints, either Gricean in nature (minimise ambiguity), non-Gricean and perceptual (salience) or purely circumstantial (bias in instructions).

Experiments 1--3 in this paper were designed to address these issues. The first two experiments follow a ratings methodology and probe participants' comprehension by manipulating the factors which may have led participants to prefer over-informative utterances in Engelhardt et al.'s study. Experiment 3 measures rates of informativeness in the production of REs, using similar stimuli to those used in experiment 2.

The experiments also inform the overarching question about the role of Gricean pragmatics in sentence processing. The major positions are discussed in section 6 on future research directions, but can be summarised as: (i) pragmatics is the main constraint in sentence processing, (ii) pragmatics is a theoretical abstraction and plays no part in sentence processing, and (iii) pragmatics is one constraint amongst others on processing. Whereas Engelhardt et al. argue for (ii) with respect to Q-2, a review of the sentence processing literature on Q-1 leads us to support (iii) at this point. If the experiments provide further support in this direction, then research can continue towards the goal of establishing a ranked list of constraints involved in sentence processing, including Gricean-pragmatic and nonlinguistic factors.

2. Experiment 1: sensitivity to informativeness violations: replication of Engelhardt et al. (2006)

Experiment 1 aims to replicate the findings of Engelhardt at al.'s experiment 2b in order to provide a comparable baseline measure of informativeness sensitivity in a complex referential context, for comparison with the simpler contexts which feature in experiment 2. As in Engelhardt et al.'s study, subjects used a 5-point rating scale to judge spoken utterances alongside a visual context depicting a static scene before and after a change had taken place. Reaction time data (which were not collected in Engelhardt et al.'s study) were also gathered for the ratings. Assuming that processing infelicitous utterances is more costly than processing optimal ones, longer latencies are predicted for the under- and over-informative utterances than for the pragmatically optimal items. The only significant modification of the original methodology is in the wording of the instructions and ratings criteria. In the original experiment, participants were asked to rate the utterances on a scale of correctness. For the reasons discussed in section 1.2, the new instructions ask participants to rate each utterance according to 'how naturally' it described the change from the first to the second scene. Ratings according to naturalness should encompass both semantic and pragmatic violations (cf. the solely semantic judgement induced by 'correct' in Engelhardt et al.'s experiment). This criterion also does not induce the expectation that more information warrants higher ratings (as 'could not be better' is likely to have done), and is appropriate for both our experimental and the syntactically erroneous filler items.

2.1. Method

2.1.1. Participants

24 university students participated in the experiment (mean age 20 years; 5 males and 19 females). All were native speakers of English and did not participate in experiments 2 and 3.

2.1.2. Design

Following Engelhardt et al. (2006), the experiment comprised a 2 (number of referents) 2 (modification) 2 (destination) design, giving rise to eight conditions (see Table 1). Visual displays contained either one or two possible target referents, the instruction type either contained a postnominal prepositional phrase modifier or did not, and the destination type either matched or did not match the location-object on which the target originally rested. All variables were manipulated within subjects, but number of referents was a between-items variable and the other two variables were within-item. There was no theoretical motivation for including destination type as a factor in the current design, as it was not significant in Engelhardt et al.'s study; it was included merely to make the design of the replication as similar as possible to the original experiment.

Two lists were created. Items were rotated by instruction type (modified/unmodified), i.e. each visual display appeared in both a modified and an unmodified version, paired as felicitous-infelicitous conditions. Lists were used between subject groups, randomly allocated (see Appendix 1 for full lists).

2.1.3. Materials and procedure

The visual stimuli consisted of photographs of scenes before and after a change had taken place, presented side by side on a computer screen (see Fig. 1a and b for example arrays). The photographs were taken specially for the experiment. A full list of items can be found in Appendix 2.

The left panel shows the scene before and the right panel shows the scene after the change has taken place. Associated utterances were 'Put the peg on the other glove' (optimal-1 condition) and 'Put the peg on the glove on the other glove' (over-informative condition).

Again, the left panel shows the scene before and the right panel shows the scene after the change has taken place. Associated utterances were 'Put the button on the sock' (under-informative condition) and 'Put the button on the sponge on the sock' (optimal-2 condition).

Novel audio stimuli were created using the method described by Engelhardt et al. The sound files were recorded by the native-speaker experimenter. The unmodified instructions were created by digitally removing the modifier from modified instructions using Praat speech synthesis software (Boersma and Weenink, 2010). For example, a sound file was created containing the instruction 'put the button on the sponge on the sock', forming a modified utterance which was optimal for a 2-referent display (as in Fig. 1b) and over-informative for a 1-referent display. The same file then had the PP 'on the sponge' removed, forming the unmodified counterpart for the same visual stimuli: under-informative for a 2-referent display and optimal for a 1-referent display. The creation of these utterances ensured that unmodified items for both matching and different destinations had similar prosody, thus avoiding prosodic cues which could affect ratings of the sound stimuli.

The experiment was programmed using SuperLab 4.5 experimental software. Auditory stimuli were played through headphones and responses were made using a USB response pad with five active buttons. Participants completed the experiment on a laptop in a purpose-designed testing suite with the experimenter in an adjacent room.

Before the experiment began, participants sequentially saw isolated images of all of the objects that would be used in the experiment, and simultaneously heard the noun phrase (NP) that would be used for each, e.g. 'button'. This phase familiarised the participants with the specific label to be used for each item. This manipulation of the original design was intended to ensure that neither the choice of lexical item nor any difficulty in recognising the objects would influence ratings and RTs. Participants then read the instructions for the experiment, including details of the ratings scale. They were instructed to rate the spoken command for its naturalness within the visual context on a scale of 1--5, where 1 represented a very unnatural command for the visual context, 3 was neither natural nor unnatural, and 5 was a very natural command for the context.

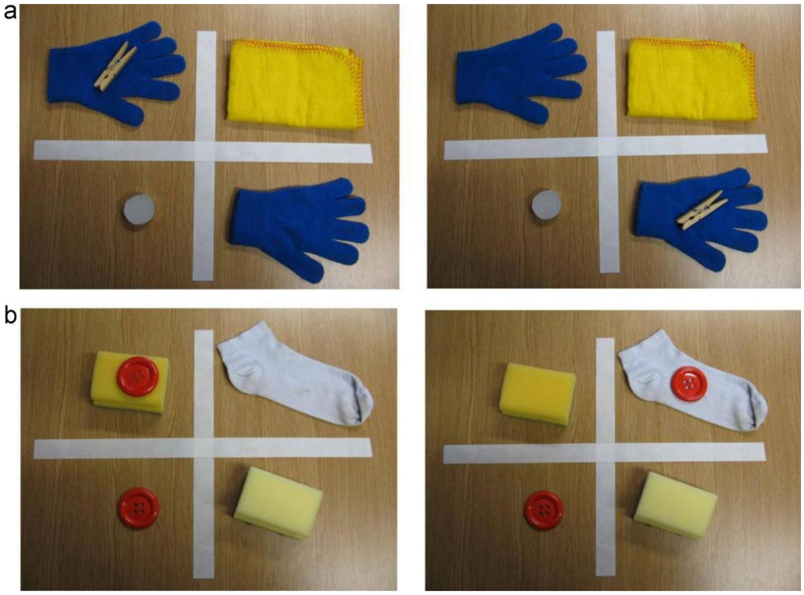


Fig. 1. (a) Experiment 1: example visual stimulus (1-referent conditions: item C05). (b) Experiment 1: example visual stimulus (2-referent conditions: item C22).

Each trial began with a fixation cross displayed for 1000 ms, followed by the visual display which was shown for 2000 ms before the utterance began. The display remained on the screen throughout the audio stimulus and until the participant responded. RTs were recorded, monitoring latencies between the offset of the spoken utterance and the onset of the participants' response. The order of trials was randomly assigned for each participant and participants were tested individually. The experiment took around 15 min to complete.

Each participant completed ten randomised practice trials exemplifying each of the control and critical conditions which followed in the experiment. There were 72 trials in the session, consisting of 24 critical trials with 3 in each condition (see Table 1 for the factors within each condition) and 48 control trials (which also functioned as fillers to disguise the focus on informativeness). In line with the original design, half of all trials were 'good' and half were 'bad'. The bad trials each involved a clear violation of expectations, where the description was semantically false (mentioning an erroneous object), or pragmatically under-informative. Full stimulus lists can be found in Appendix 1.

2.2. Results

2.2.1. Ratings

Results of experiment 1 (ratings) are shown in Table 1. Although the discontinuous nature of the 1--5 ratings scale typically requires nonparametric statistical analyses, parametric statistics are used in line with Engelhardt et al's analysis.

Table 1

Experiment 1: conditions and mean ratings (standard error of the mean).

	1-Referent		2-Referent	
	Matching destination	Different destination	Matching destination	Different destination
	Over-informative	**Over-informative**	**Optimal-2**	**Optimal-2**
Modified	4.23 (.19)	3.98 (.16)	4.37 (.17)	4.14 (.20)
	Optimal-1	**Optimal-1**	**Under-informative**	**Under-informative**
Unmodified	4.72 (.10)	4.70 (.10)	3.52 (.21)	3.04 (.15)

The result of a 2 x 2 x 2 repeated measures ANOVA revealed a main effect of number of referents, with higher ratings in favour of 1-referent contexts, *F1*(1,23) = 56.82, *p* < .001[2], and a main effect of destination, with higher ratings in favour of matching contexts, *F1*(1,23) = 11.18, *p* < .005. Crucially, there was an interaction of number of referents and modification, *F1*(1,23) = 46.48, *p* < .001. No other interactions were found to be significant.

Since there is no interaction between destination and either of the other factors, the matching and different conditions are collapsed in the subsequent analysis resulting in a 2 x 2 analysis crossing number of referents by modification. The resulting four conditions are: over-informative (1-referent, modified), under-informative (2-referents, unmodified), and two optimal conditions: optimal-1 (1-referent, unmodified) and optimal-2 (2-referents, modified). Pairwise analyses now allow us directly to compare the penalties for straightforwardly under-informative and over-informative utterances. Ratings for the four experimental conditions plus the two control conditions are shown in Fig. 2.

Further pairwise planned comparisons by means of t-tests were conducted with a Bonferroni correction applied, These comparisons reveal that over-informative utterances (M = 4.10, SE = .16) were rated significantly lower than their optimal-1 counterparts (M = 4.71, SE = .09), *t*1(23) = 4.19, *p* < .001. Under-informative utterances (M = 3.28, SE = .15) were rated significantly lower than their optimal-2 counterparts (M = 4.26, SE = .17), *t*1(23) = 4.59, *p* < .001. Comparing the two modified instruction conditions (over-informative vs. optimal-2) revealed no significant differences. However, comparing the two unmodified conditions revealed that under-informative utterances were rated lower than optimal-1 utterances, *t*1(23) = 8.43, *p* < .001. When comparing by infelicity type, results show that under-informative utterances were rated significantly lower than over-informative utterances, *t*1(23) = 4.32, *p* < .001.

Post hoc analyses between the controls and the infelicitous conditions reveal that the semantically true controls (M = 4.71, SE = .07) were rated significantly higher than both the under-informative, *t*1(23) = 9.10, *p* < .001, and the over-informative utterances, *t*1(23) = 4.34, *p* < .001. The semantically false controls (M = 1.53, SE = .08) were rated significantly lower than both the under-informative utterances,

[2] By-items statistics are not reported since item variability was controlled in each experiment by counterbalancing lists across groups of participants (Raaijmakers et al., 1999).

*t*1(23) = 9.79, *p* < .001, and the over-informative utterances, *t*1(23) = 13.85, *p* < .001.

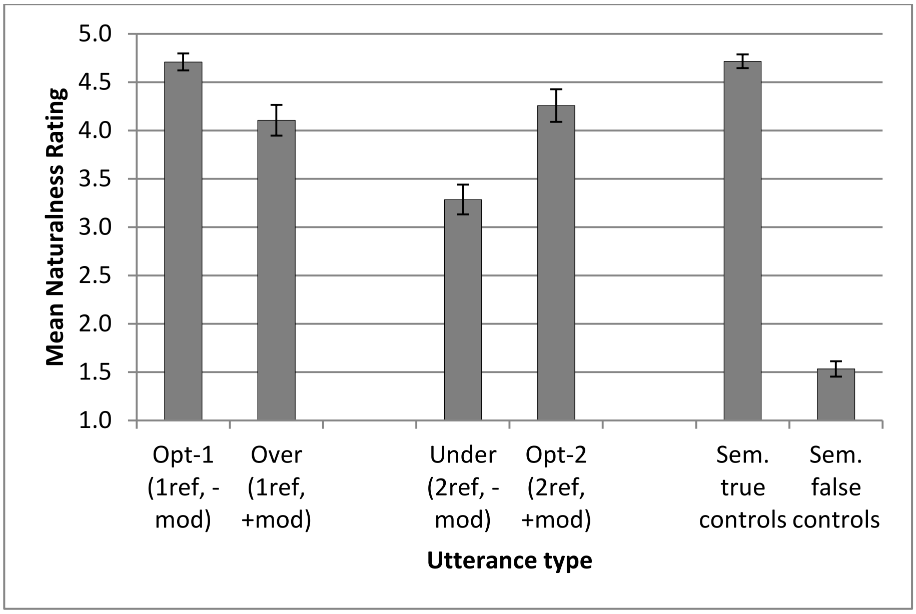


Fig. 2. Experiment 1: mean ratings (matching/different conditions collapsed). Error bars show standard error of the mean.

2.2.2. Reaction times

Mean RTs by condition are presented in Fig. 3. Latencies were measured from the offset of the verbal instruction to the participant's response. Any RTs longer than two standard deviations above participant means were discarded. There were no RTs lower than 2SDs below participant means for any participant, but those lower than 100 ms were also discarded. Altogether, 99 of the 1728 RT data points were discarded (6%). By condition, these were 25/432 for semantically true control items, 42/720 for semantically false control items, 2/144 for optimal-1 items, 11/144 for optimal-2 items, 4/144 for over-informative items, and 15/144 for under-informative items. Mean latencies for each condition were then calculated, and are shown in Fig. 3. As in the analysis of ratings, the matching and different conditions are collapsed, resulting in a 2 2 design.

A 2 x 2 repeated measures ANOVA for the four critical conditions reveals a main effect of number of referents, *F*1 (1,23) = 19.31, *p* < .001, a main effect of modification *F*1(1, 23) = 7.50, *p* < .05, and a significant interaction between number of referents and modification *F*1(1, 23) = 9.24, *p* < .01.

Further pairwise planned comparisons by means of t-tests were conducted with a Bonferroni correction applied. There were no significant differences in RTs between over-informative utterances (M = 1890, SE = 216) and their optimal-1 counterparts (M = 1907, SE = 145). Under-informative utterances (M = 2829, SE = 268) elicited significantly longer RTs than their optimal-2 counterparts (M = 2033, SE = 182), *t*1(23) = 4.07, *p* < .001. Comparing the two modified instruction conditions (over-informative vs. optimal-2) revealed no significant differences, but comparing the two unmodified conditions revealed that under-informative utterances yielded longer reaction times than optimal-1 utterances, *t*1(23) = 4.95, *p* < .001. When comparing by infelicity type, under-informative utterances elicited significantly longer RTs than over-informative utterances, *t*1(23) = 4.42, *p* < .001.

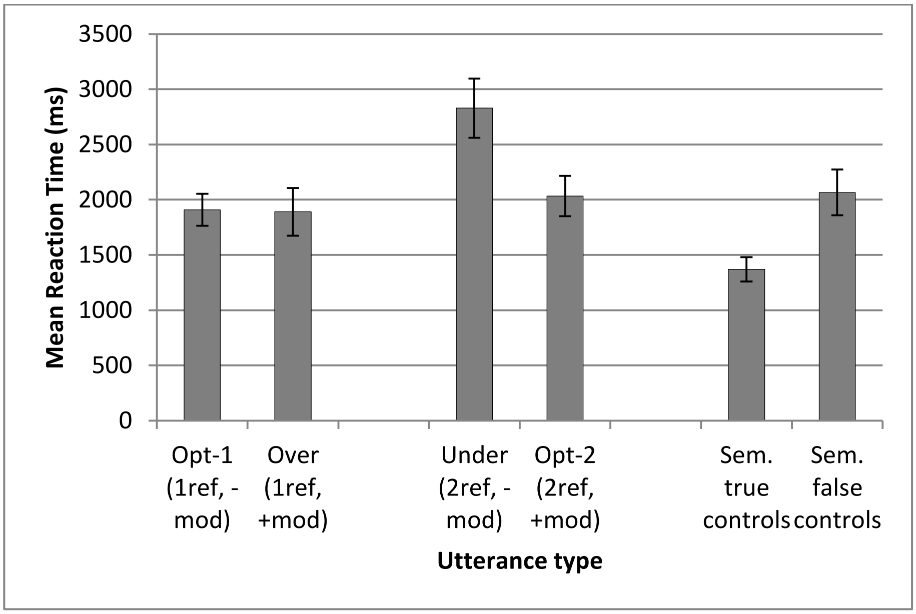


Fig. 3. Experiment 1: mean reaction times (matching/different conditions collapsed). Error bars show standard error of the mean.

Post hoc analyses reveal that the true controls (M = 1370, SE = 109) elicited significantly shorter RTs than the under-informative items, *t*1(23) = 7.04, *p* < .001, and the over-informative items, *t*1(23) = 2.64, *p* < .05. The false controls (M = 2065, SE = 208) elicited significantly shorter RTs than the under-informative items, *t*1(23) = 4.41, *p* < .001. There were no significant differences between the false controls and the over-informative items.

Overall there is a symmetry between the ratings and the reaction time measures: those conditions which received lower ratings also yielded longer reaction times. The exception is the false control items, which were rated lower than both types of pragmatically infelicitous items, but were responded to faster than the under-informative items and comparably to the over-informative items. This may be due to the clearly dispreferred nature of the semantically false controls, relative to the more subtle pragmatic violations.

2.3. Discussion

While our rating results accord with Engelhardt et al.'s results for sensitivity to under-informativeness, experiment 1 yields the novel finding that over-informative utterances are rated significantly lower in naturalness than their optimal-1 counterparts. In line with Engelhardt et al.'s (2006) experiment 2, experiment 1 obtained significantly lower ratings for under-informative utterances compared to pragmatically optimal utterances. RTs for under-informative items were also significantly slower than for the optimal items. This is good evidence that hearers are indeed sensitive to under-informativeness. As well as contravening expectations of full informativeness, the latency for the under-informative item may be partly due to a process of elimination. Recall that the display shows, e.g. a button on a sponge and a button not on a sponge, and the utterance instructs to 'put the button on the sponge'. In this case, the hearer would need to rule out the button which already fulfils the instruction. We argue that this process, if it occurs, is still Gricean, albeit related to the maxim of Relation rather than Quantity. That said, we maintain that the elimination process is not the whole story: some consideration of what would be felicitous (i.e. describing a novel and relevant situation) precedes it.

Also in line with Engelhardt et al., numerically lower ratings were obtained for over-informative utterances as compared to their optimal counterparts, one-referent unmodified utterances (4.3 vs. 4.6 in Engelhardt et al.'s study; 4.1 vs. 4.7 in the current study). However in the present experiment, unlike theirs, this difference reached significance, suggesting sensitivity to Q-2.

Nevertheless, comparing the two modified instruction conditions (over-informative vs. optimal-2) revealed no significant differences. This suggests that the significant interaction may be driven by the short instructions yielding some leniency towards over-informativeness in this design. Furthermore, there were no significant differences in reaction times between over-informative utterances and either of their optimal counterparts. Therefore the evidence for a penalty towards over-informativeness is not yet robust -- certainly not as robust as the evidence that was obtained for under-informative utterances, which were both penalised and slower relative to the optimal utterances, as well as to the over-informative ones. The difference between the two types of informativeness violation was expected, since they give rise to different kinds of communicative failure in the referential communication paradigm.

Although we do not have unequivocal evidence in favour of sensitivity to Q-2, the new ratings scale provides a greater indication that participants are sensitive to over-informativeness. As argued in section 1.2, the lack of a robust penalty for over-informative utterances using these materials may not reflect a lack of sensitivity, but rather may indicate a preference for additional information in conditions where it is motivated. To disentangle preference from lack of sensitivity, truly gratuitously over-informative utterances are required. These can be created by simplifying the visual display and removing any complexity-related preference for over-informativeness. If ratings and reaction times for over-informative utterances and optimal utterances do not show unequivocal differences in such a design, then there would indeed be evidence for a lack of sensitivity to informativeness. Another related possibility is that pragmatic considerations can only exert an effect when they are allocated sufficient processing resources. These two considerations motivate experiment 2 which uses a simple referential world to remove the contextual motivation for over-informativeness and to free up resources for potential Gricean reasoning.

3. Experiment 2: sensitivity to informativeness violations with prenominal modification

Experiment 2 tests sensitivity to informativeness using simplified stimuli. This attempts to reduce the complexity and visual salience issues which may have influenced the ratings and RTs for over-informative utterances in experiment 1 and in Engelhardt et al.'s 2006 study. Task demands are reduced, since participants only have to process a single display rather than considering the difference between two, and the displays show a somewhat simpler visual world. Should sensitivity to Q-2 be demonstrated, through lower ratings and longer RTs for over-informative than for optimal utterances, this would suggest (together with the established Q-1 penalties) that interlocutors are sensitive to both Gricean maxims of Quantity, and that these operate alongside referential constraints.

The auditory stimuli in experiment 2 contain NPs modified with prenominal adjectives rather than with postnominal PPs as in experiment 1. While the latter are ambiguous (modifier of the noun or destination of the movement), the syntactic role of prenominal adjectives is unambiguous. The visual stimuli are simpler than those used in experiment 1, comprising four single objects, rather than compositional objects-on-objects. The movement-to-destination component was also removed. The fundamental 2 2 design of experiment 1 was maintained, and participants were again tested for their sensitivity to violations of informativeness.

3.1. Pretesting for default descriptions

A pre-test was performed on potential stimuli, in which items showing attributes from several dimensions, including scalar and absolute contrasts (e.g. size, material), were presented in isolation on-screen. English-speaking adult participants (n = 31) described each item in response to the written prompt 'What's this?' Items which were frequently referred to using an unmodified NP in the pre-test were added to the stimulus lists for subsequent use, and those attracting modification in more than 60% of elicited REs were discarded. This was done to rule out penalisation on the grounds that the particular stimulus possessed such a salient attribute that it would seem marked to omit it, even when appearing in isolation (cf. Karmiloff-Smith's descriptor/determinor functions; 1979:45). For example, one of the discarded images depicted a broken cup for which 88% of the elicited REs included the modifier 'broken'. The pre-test helps ensure that the stimuli in experiments 2 and 3 are likely to have an unmodified default description, which means that hearers should not have strong expectations about the use of an adjective.

3.2. Method

3.2.1. Participants

24 university students participated in the experiment (mean age 25 years; 6 males and 18 females). All were native speakers of English and did not participate in experiments 1 and 3.

3.2.1.1. Design.

Experiment 2 had a 2 (number of referents) 2 (presence of adjectival modification) within-subjects design, creating four conditions: over-informative (1-referent, modified), under-informative (2-referent, unmodified) and two optimal conditions; optimal-1 (1-referent, unmodified) and optimal-2 (2-referent, modified). These are illustrated in Fig. 4a-d.

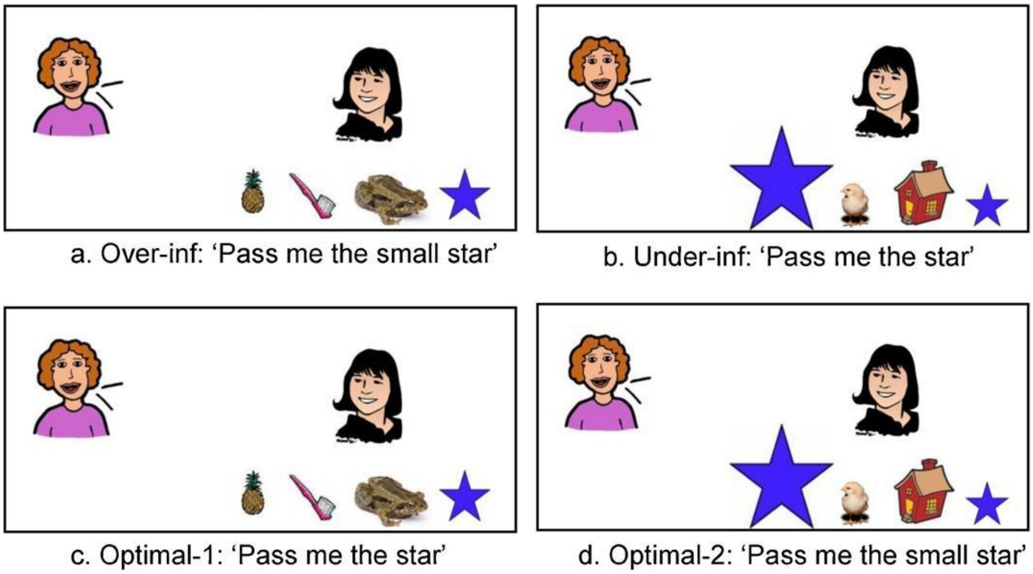

a. Over-inf: 'Pass me the small star'
b. Under-inf: 'Pass me the star'
c. Optimal-1: 'Pass me the star'
d. Optimal-2: 'Pass me the small star'

Fig. 4. (a--d) Experiment 2: example stimuli.

3.2.2. Materials and procedure

The experiment was programmed using SuperLab 4.5 experimental software. A static display on a computer screen showed two characters, one of whom had four items in her vicinity. In these simple arrays where there is at most one contrasting item of the same category as the target, participants must expend little effort deciding whether an attribute is minimally contrastive or not, and so should have sufficient resources to make this calculation and then make their rating. Participants heard the speaker-character asking the hearer-character to 'pass me the [referring expression]' on pre-recorded sound files, and then were asked to rate how natural the instruction was in the visual context using a 5-point scale. Responses were made using a USB response pad with five active buttons, with 1 signifying a 'very unnatural' utterance and 5 signifying 'very natural'. Participants completed the experiment in a purpose-designed testing suite with the experimenter in an adjacent room. The experiment took around 15 min to complete, and formed part of an hour-long testing session containing unrelated experiments.

There were 40 critical items, 10 in each condition, plus 20 syntactically infelicitous items, e.g. 'pass me the cup plastic'. These control items were included to ensure that participants were able to penalise straightforwardly inappropriate utterances (see Appendix 2 for a sample list of items). Items were randomly presented and the position of the target referent and the contrasting referent was rotated between items in all conditions. A Latin square design was used to counterbalance item effects: every target item appeared in only one of the four conditions for a given participant. The same syntactically infelicitous items were used across versions.

3.3. Results

3.3.1. Ratings

The mean ratings (n = 24) for the four experimental conditions together with those for the syntactically infelicitous controls are presented in Fig. 5. As no comparison is made with the rating study of Engelhardt et al. (2006), non-parametric statistics appropriate to the discrete Likert scales are now used.

Friedman's ANOVAs for nonparametric data revealed a significant difference between critical conditions, $x^2_1(3) = 41.41$, $p < .001$. Further pairwise planned comparisons were conducted using Wilcoxon signed-rank Tests with a Bonferroni correction applied. Comparing the two 1-referent conditions revealed that the over-informative utterances (M = 3.95, SE = .15) were rated lower than optimal-1 utterances (M = 4.87, SE = .04), *Z*1 = 4.20, $p < .001$. Likewise, comparing the two 2-referent conditions revealed that the under-informative utterances (M = 3.51, SE = .19) were rated lower than optimal-2 utterances (M = 4.70, SE = .07), *Z*1 = 3.82, $p < .001$. Comparing the two modified instruction conditions also yielded significant differences: over-informative utterances were rated lower than optimal-2 utterances, *Z*1 = 3.73, $p < .001$, and comparing the two unmodified conditions revealed that under-informative utterances were rated lower than optimal-1 utterances, *Z*1 = 4.11, $p < .001$. Comparing by infelicity type, under-informative utterances were rated lower than over-informative utterances, *Z*1 = 2.71, $p < .01$. There were no significant differences between the two optimal conditions.

Post hoc analyses reveal that ratings for the syntactically violated controls (M = 1.69, SE = .10) were significantly lower than those for both the under-informative items, *Z*1 = 4.06, $p < .001$, and the over-informative items, *Z*1 = 4.23, $p < .001$.

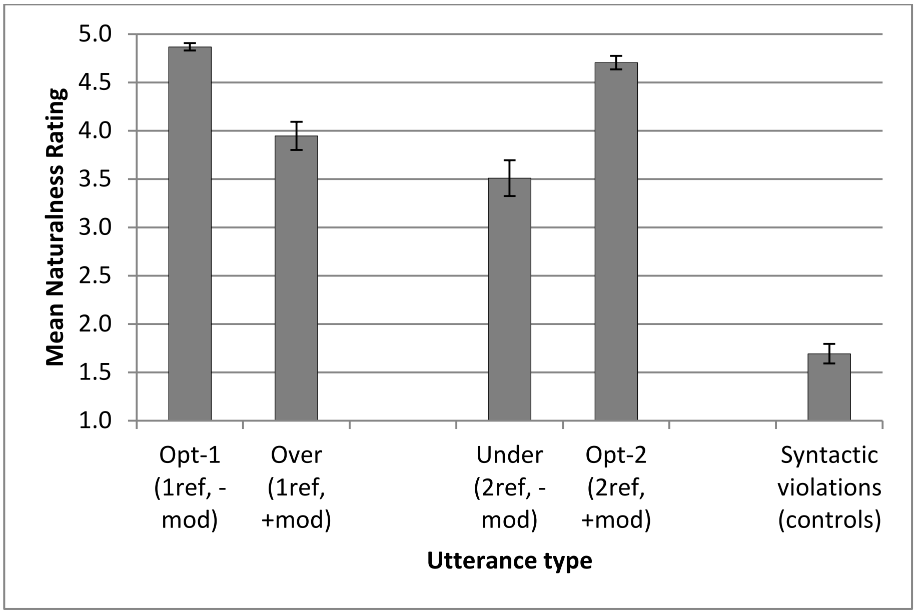


Fig. 5. Experiment 2: mean ratings. Error bars show standard error of the mean.

3.3.2. Reaction times

Mean RTs by condition are presented in Fig. 6. Latencies were measured from the offset of the verbal instruction to the participants' responses. Any RTs longer than 2 SDs above participant means were discarded. There were no RTs lower than 2SDs below means for any participant, but those lower than 100 ms were also discarded. In total, 147 of the 1440 RT data points were discarded (10%): 43/480 responses to the control items, 13/240 in the optimal-1 condition, 37/240 in the optimal-2 condition and 28/240 in the over-informative condition. Mean latencies for each condition were then computed.

A 2 x 2 repeated measures ANOVA for the four critical conditions reveals no main effect of number of referents or of modification. There was however a highly significant interaction between number of referents and modification, *F*1(1, 23) = 21.85, $p < .001$.

Further pairwise planned comparisons were conducted by means of t-tests with a Bonferroni correction applied. RTs were longer for the over-informative utterances (M = 1421, SE = 128) than for optimal-1 utterances (M = 929, SE = 88), *t*1(23) = 5.18, $p < .001$. RTs were longer for the under-informative utterances (M = 1320, SE = 117) than for optimal-2 utterances (M = 968, SE = 121), marginally significant at *t*1(23) = 2.45, $p = .02$. A similar pattern holds for the by-modification comparisons: RTs were longer for over-informative than for optimal-2 utterances, *t*1(23) = 4.89, $p < .001$, and longer for under-informative than for the optimal-1 utterances, *t*1(23) = 3.28, $p < .005$. RTs were not significantly different for under-informative and over-informative utterances, nor for optimal-1 and optimal-2 utterances.

Post hoc analyses revealed that the syntactically infelicitous controls (M = 1194, SE = 86) elicited significantly shorter RTs than the over-informative items only, t1(23) = 3.04, p < .01: the more salient syntactic violation attracts a quicker response, whereas the Q-2 violation is arguably more subtle and yields greater response latency.

Note also that the mean RTs for critical conditions in experiment 2 (1162 ms) are numerically shorter than those in experiment 1 (2165 ms). We take this to indicate the lower task demands required for processing the simpler visual array, the lack of syntactic ambiguity, and the simpler modificational frame in experiment 2.

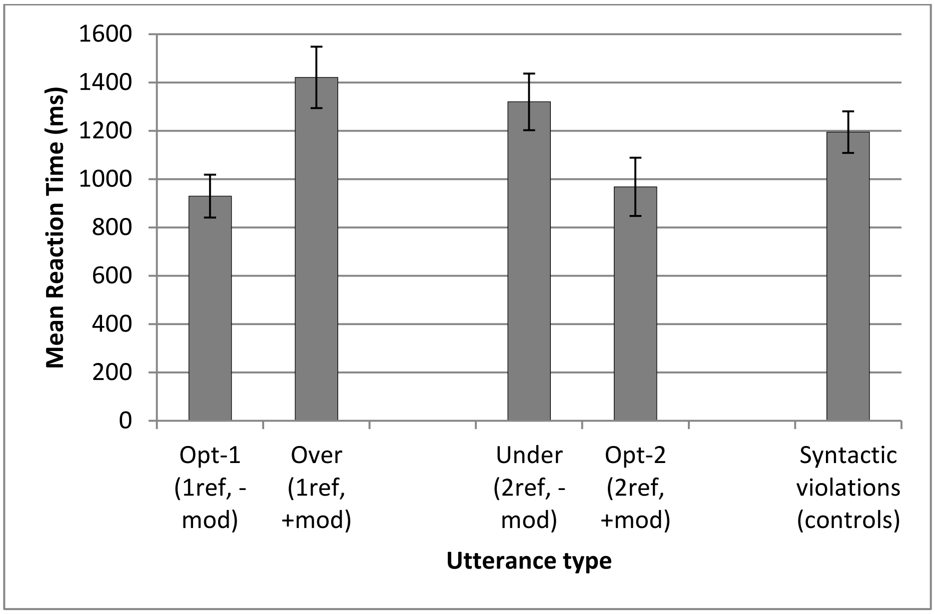


Fig. 6. Experiment 2: mean reaction times. Error bars show standard errors.

3.4. Discussion

Experiment 2 revealed sensitivities to violations of both Quantity maxims: under- and over-informative utterances were penalised relative to optimally-informative utterances, and also elicited longer RTs. As in experiment 1, under-informativeness was rated significantly lower than over-informativeness, in line with Gricean predictions of relative lenience for violations of the second Quantity maxim, and with reference resolution constraints. Our results therefore suggest that in simple visual worlds, where there is little motivation for over-informing, hearers are sensitive to violations of Q-2. As there was minimal visual complexity in experiment 2, and no other entity shared the adjectival property, modification in the over-informative condition was straightforwardly redundant and served no clarifying function, unlike in Engelhardt et al's study. While over-informative descriptions may facilitate processing in complex arrays (Arts, 2004; Arts et al., 2010; Mangold and Pobel, 1988), the longer RTs found in this experiment indicate that this does not hold in simple arrays. On the basis of these data, we tentatively hypothesise that the time may be spent looking for a function for the adjective. Accordingly, over-informative descriptions were penalised relative to both optimal conditions in this experiment (contra expt. 1).

If interlocutors adhere to Gricean maxims in the hearer role, we should expect them also to do so in the speaker role (as was found in Engelhardt et al.'s expt. 1 and 2). Experiment 3 tests this prediction using the same stimuli as used in experiment 2.

4. Experiment 3: production of referring expressions

Experiment 2 suggests that hearers are Gricean in their expectations of informativeness in referring expressions. Experiment 3 seeks to document rates of under- and over-informing in production contexts similar to those used in experiment 2.

Previous studies using the referential communication paradigm document reasonably high (though variable) rates of over-informing in production. Pechmann (1989) found that 21% of adults' expressions in a referential task contained redundant descriptions, i.e. attribute(s) of the intended referent that were not necessary for unique identification. Engelhardt et al. (2006) found that 30% of the REs in their production study were over-informative, and Viethen and Dale (2006) recorded 25% of their participants' REs as redundant. In all three studies, the referential arrays were more complex than those used in the current experiment, containing more objects in a single array, and more dimensions along which the target and competitor objects varied. Complexity stemming from (i) numbers of objects, (ii) their compositional nature, and (iii) the presence of identical destinations may have motivated over-informing in Engelhardt et al.'s work (2006), as discussed in section 1.2.

These representative studies suggest that around a quarter of REs are over-informative, and that speakers commonly give more information than is minimally required. It is much less common for speakers to under-inform, i.e. to give less information than is required to identify an entity (Engelhardt et al., 2006; expt. 1). However, the factors in this experiment that favoured over-informativeness (see experiments 1 and 2) may bias speakers as well as hearers. Experiment 3 examines whether speakers and hearers are matched in rates of informativeness, using the simple referential array from experiment 2. As in the comprehension experiment, it is anticipated that the simplicity of the arrays will lead to lower rates of over-informing, as participants can easily complete a full scan of the array before articulating their RE.

4.1. Method

4.1.1. Participants

24 university students participated in the experiment (mean age 26 years; 13 males and 11 females). All were native speakers of English and did not participate in experiments 1 and 2.

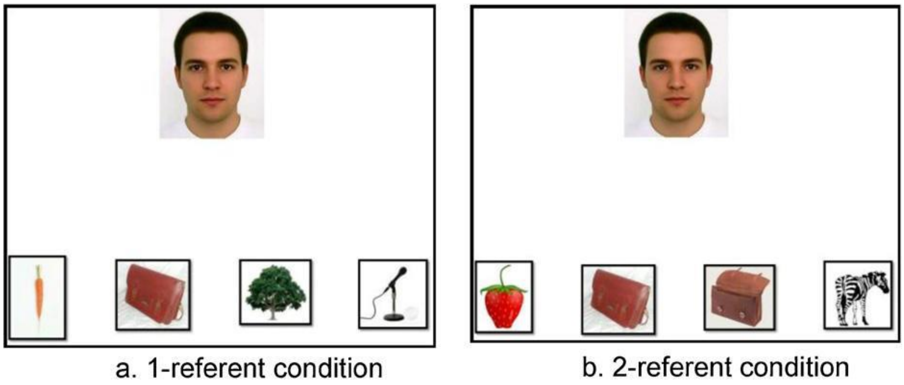
4.1.2. Design

Experiment 3 used a repeated measures design, where presence/absence of a contrast set was manipulated. Speakers saw visual stimuli containing either one or two referents of the same type, accompanied by either two (2-referent condition) or three (1-referent condition) non-referents. See Fig. 7a and b for sample arrays.

4.1.3. Materials and procedure

The visual stimuli were similar to those used in experiment 2. Forty arrays were created, each in two versions; a 1-referent display (no contrast set), and a 2-referent display (contrast set). Half of the participants saw the first group of items in the 1-referent display and the second group of items in the 2-referent display, while this was reversed for the other participants. See Appendix 3 for a sample list of items. The order of items within each group was randomised and the relative position of the target referent and the contrasting referent was rotated between items. There were two practice items at the beginning of the experiment, exemplifying each of the two conditions.

Participants were presented with the stimuli via a static laptop display using Microsoft PowerPoint software and were asked to instruct an on-screen hearer (depicted by a photograph) to 'pass one of these objects in a way that he would easily understand'. They were told that the on-screen items were visible both to themselves and the on-screen hearer. The target items were cued in a separate physical booklet visible only to the participant, which showed the same displays as those on-screen but with the target item (one per array) highlighted by an arrow. The participants were instructed to proceed by advancing the visual display to the next slide using the laptop keyboard, inspecting the complete visual array, and then turning the page in their booklet to reveal the identity of the target and requesting that object. The task was administered in a purpose-designed testing suite by a single experimenter who stayed in an adjacent room after the instructions and practice items were successfully completed. The participants were tested individually. Their responses were voice-recorded and later transcribed and coded. The production study took around 5 min to administer, and formed part of a longer experimental session involving additional experiments for unrelated research projects.



a. 1-referent condition

b. 2-referent condition

Fig. 7. (a and b) Experiment 3: sample on-screen visual stimuli. The target item was cued in a separate booklet given to the participants (2nd item from the left in this case in both illustrated items).

4.1.4. Coding the responses

All responses were recorded, transcribed and classified as under-informative, optimally informative or over-informative depending on the visual display. For example, in the array depicted in Fig. 7a, an optimally informative expression would be ‘pass me the bag’, and an over-informative expression would be ‘pass me the closed bag’ or ‘pass me the leather bag’. ‘Pass me the bag’ would be under-informative in Fig. 7b, with e.g. ‘pass me the closed bag’ coded as optimal and ‘pass me the closed leather bag’ as over-informative. The over-informative expressions were coded for the type of attribute used, and the modified expressions were also coded for syntactic form, e.g. comparative or postnominal modification.

4.2. Results

4.2.1. Quantitative analysis

The proportions of under-informative, optimal, and over-informative referring expressions for 1-referent and 2-referent arrays are presented in Fig. 8.

With regard to the 1-referent displays, where only two types of RE were documented (optimal and over-informative, M = 18.08 (max frequency 20), SE = .42 and M = 1.92, SE = .42 respectively), a Wilcoxon signed-rank test for non-parametric data revealed a significant difference between optimal and over-informative descriptions, $Z1 = 4.31$, $p < 0.001$. With regard to the 2-referent displays, where under-, optimal-, and over-informative expressions were elicited; M = .63, SE = .20; M = 17.92, SE = .36; M = 1.46, SE = .34, respectively, a Friedman’s ANOVA for non-parametric data revealed a significant difference between conditions, Z1 ($x^2(2) = 39.06$, $p < .001$). Still with the 2-referent displays, further pairwise comparisons by means of Wilcoxon signed-rank tests revealed significant differences between under- and optimally-informative utterances, $Z1 = 4.30$, p < .001, and optimal and over-informative utterances, $Z1 = 4.30$. p < .001. Rates of under- and over-informative utterances in the 2-referent displays were not significantly different.

Overall, in displays both with and without a contrast set, speakers used a minimally contrastive referring strategy in around 90% of their REs, i.e. they were highly optimal. In the 2-referent condition, where both under- and over-informative expressions were documented, the rates of occurrence of these expressions did not significantly differ. Rates of optimal-and over-informativeness were similar across both types of displays, despite over-informative utterances requiring at least one adjective in the 1-referent condition and at least two in the 2-referent. Unsurprisingly, under-informativeness was attested only in the 2-referent condition, since under-informativeness in the 1-referent condition would involve omitting the noun. Crucially, rates of over-informativeness (which averaged out to 8.4% across both conditions) were lower than the rates of 30% reported by Engelhardt et al. (2006, expt. 1).

Results show that in a simple visual world, speakers do not often produce over-informative expressions, which parallels the results from experiment2 showing that hearers are sensitive to this type of speaker behaviour. In both the speaker and the hearer role, interlocutors appear to be sensitive to violations of both Quantity maxims. This is evidence that interlocutors are aware of expectations of informativeness other than those minimally required to achieve reference resolution?

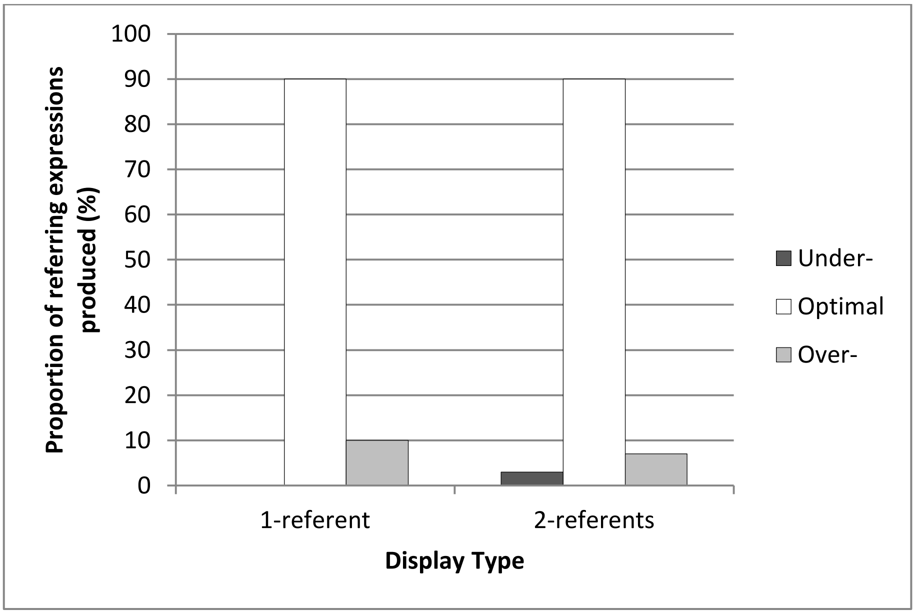


Fig. 8. Experiment 3 results: production of under-, optimal, and over-informative referring expressions.

4.3. Qualitative analysis of the over-informative expressions

The 81 tokens of overspecified reference that were elicited from 960 REs were analysed for type. Of these overspecified tokens, the attribute most frequently provided redundantly was colour (54%), followed by size (20%), material (8%), and on-screen location (5%). The remaining 13% of overspecified tokens were coded as 'other' and included salient attributes such as 'desktop' in relation to a singleton computer, 'cordless' for a singleton phone and 'sleeping' for a singleton baby. The results regarding the colour over-modification accord with previous work which suggests that colour attributes are commonly used redundantly as well as contrastively (Eikmeyer and Ahlsen, 1998; Mangold and Pobel, 1988; Pechmann, 1989; Rubio-Fernández and Glucksberg, 2010; Schriefers and Pechmann, 1988; Weiss and Mangold, 1997). In particular, Sedivy (2002) found that colour modifiers are frequently encoded in default descriptions, probably due to their visual salience and absolute (rather than scalar) nature[3]. This suggests that at least some of the attested over-informativeness could be attributed to a nonlinguistic constraint, the salience of colour, overriding pragmatic considerations, rather than to a general lack of sensitivity to pragmatic maxims.

The form of optimal REs was also coded for the 2-referent condition (in which modification was required). The most frequent syntactic pattern (83%) was prenominal modification (all of the items allowed attributive use, e.g. 'the big star', 'the unsliced bread'). However, the data revealed three further trends in the form of modification. 8% of optimal utterances were comparative, e.g. 'the larger of the two vases', 'the more modern phone'. 4% of optimal utterances were appositively modified as in 'the glass, the full one' and 'the phone, the new phone'. A further 4% were modified using relative clauses such as 'the briefcase that's shut' and 'the rug that's got tassels'. The postnominal character of the latter two forms suggests that REs were initiated before the visual scan of the array was complete and so utterance planning and visual scanning took place simultaneously (Pechmann, 1989). The use of comparative expressions reveals that the participants were actively contrasting the target items with the contrast-mate.

4.4. Discussion

Simplifying the visual array removes motivation for over-informing, as documented by the high rates of optimality and relatively low rates of over-informing in the referring expressions elicited in experiment 3. As discussed above, previous experiments using more complex displays have documented higher rates of over-informing, allowing us to reconcile the current results from the simplified stimuli with previous patterns in the literature.

In addition to differences in referent numbers/dimensions between the current study and its forerunners, differences in procedure are likely to have impacted upon the amount of detail given in REs. In work comparing referring expressions and visual scanning processes (Eikmeyer and Ahlsen, 1998; Pechmann, 1989), overspecifications are cast as by-products of incremental on-line processes wherein REs are encoded at the same time as scanning the visual array. Appositively modified items in

[3] Relatedly, in comprehension, colour modification does not trigger contrastive inference, at least for objects with predictable colour modification (Sedivy, 2003).

our data lend support for the occurrence of such processing. The current procedure differed from previous studies in that the participants completed a full scan of the array before the target was revealed to them. This means that if participants anticipated the target to be one member of the contrast set (which could have been learned over the course of the experiment), they had time to identify the distinguishing feature of the contrast set in the 2-referent displays and to plan their expression accordingly. This forced separation of difference-detection and articulation could account for the relatively low rates of over-informing in the data, relative to studies that have elicited incremental, naturalistic REs. However, a comparison with a similar production experiment which did not separate difference-detection and articulation (Davies and Katsos, 2010, expt. 1) reveals relatively stable rates of overinforming: 6% in Davies and Katsos (2010) versus 8% in the new experiment (1- and 2-referent conditions combined). Moreover, the new data contained some postnominal modifications for which articulation may have started before the speakers identified the distinguishing feature: these constituted 8% of the optimal expressions in the 2-referent condition.

Production results from experiment 3 suggest that, when arrays are simple and speakers are given time to plan their REs, they produce highly optimal utterances. This load/efficiency trade-off reveals a substantial cognitive component in object reference. However, pragmatic processes are also at work. The hearer-character in the current experiment was deemed to be unaware of the identity of the target referent until the speaker-participant specified it. Consequently, the participant was responsible for providing a cooperative utterance, i.e. one which conformed to Gricean expectations of minimal informativeness. Furthermore, the hearer in experiment 3 was clearly not a live interlocutor: there remains the possibility that using a live confederate might increase the incidence of redundancy in referring expressions. The different rates of optimal productions in the current study compared to previous work do not rule out Gricean pragmatic requirements as a constraint on language production and comprehension. It is possible that when situations are complex, or when aspects of the referent are salient, speakers build in extra redundancy (as some hearer-oriented realisation of the maxim of Manner for the former, or due to a speaker-oriented effect for the latter). When arrays are simple and speakers have time to plan efficient utterances, they adhere more strictly to Quantity-2.

5. General discussion

The experiments reported provide support for the reality of interlocutors' sensitivity to both Gricean maxims of Quantity and thus for Gricean expectations to be considered as a constraint on sentence processing.

In summary, experiment 1 provides tentative evidence for hearers' sensitivity to Q-2 and replicates the robust sensitivity to Q-1 that had been found previously (Engelhardt et al., 2006). Experiment 2 unequivocally documents sensitivity to Q-2, as reflected both by ratings and reaction time data, in simplified worlds devoid of extraneous motivation for over-informing. This strongly suggests the operation of both Q-1 and Q-2 constraints in sentence comprehension. It is important to point out that 'pragmatics' should not be seen as a unitary concept: Gricean maxims have variable influence, as shown by stronger sensitivity to Q-1 than Q-2 throughout the experiments, which relates to the need to establish reference. Experiment 3 documents that speakers show the same pattern of sensitivity as found in experiment 2, i.e. they tend not to under- or over-inform in their production of REs. This adds weight to the argument for the psychological reality of Quantity-based pragmatic constraints in sentence processing across speaker and hearer roles.

The current work challenges the conclusion drawn by Engelhardt et al. (2006:572) that speakers and hearers are only 'moderately Gricean'. They argued this point on the basis that speakers produced many over-informative descriptions (their expt. 1) which hearers did not penalise (their expt. 2), whereas both speakers and hearers were much more sensitive in avoiding and penalising under-informativeness. Engelhardt et al.'s finding that over-informative descriptions nevertheless trigger a penalty in on-line visual world sentence processing (their expt. 3) led them to re-analyse this penalty as lexical-structural rather than pragmatic in motivation. However, we argue that there are pragmatic as well as non-pragmatic factors at play in their experiments which may render over-informativeness acceptable. Engelhardt and colleagues had to reconcile discrepant data from their production and ratings studies with their eyetracking results. The former two studies documented low rates of and penalties for over-informing, while the latter elicited clear on-line penalties for the same type of infelicity (although recall the concern about the unification of off-line and on-line results, footnote 1).

However, supposing that Engelhardt et al.'s off-line and on-line studies did tap into the same kind of competence, the discrepancy in the findings can still be explained. The authors were reluctant to interpret their eyetracking data as reflecting sensitivity to pragmatic constraints, as participants in their second experiment did not show sensitivity to over-informativeness in the off-line ratings, and participants in experiment 1 produced many over-informative descriptions. However, given the findings from our experiments 1--3, we suggest that it is not the results of Engelhardt et al.'s experiment 3 that need re-conceptualising as the outcome of non-pragmatic factors. Rather, their experiments 1 and 2 can be re-analysed as involving situations where the pragmatic maxim of Q-2 is overridden by the pragmatic maxim of Manner ('avoid ambiguity/be clear') as well as other non-pragmatic constraints such as visual salience.

Overall, contrary to the claim that people are only reliably sensitive to the first maxim of Quantity and 'only moderately Gricean', our experiments suggest that speakers and hearers are Gricean with regard to both under- and over-informativeness in comprehension and production. Our results suggest that people can be more Gricean when task demands are sufficiently low, and that Gricean processes may conflict with other constraints. Several clarifications are in order.

First, experiments 2 and 3 show merely that interlocutors do not expect, and avoid producing, under- and over-informative REs. They do not address whether the actual psycholinguistic competence demonstrated is of the kind Grice envisaged, namely that interlocutors are sensitive to considerations of their others' intentions and communicative relevance. It remains possible on these data that interlocutors establish what would be optimally informative by using some local metric of distinctiveness between referents, and that Gricean considerations are either theoretical rationalisations or, if psycholinguistically real, late-arriving metalinguistic constraints. That said, evidence is emerging that sensitivity to under-informativeness does indeed take into account speakers' intentions (in terms of their conversational goal; see Breheny et al., 2006). Moreover, Q-2 sensitivity is affected by considerations of interlocutor reliability and cooperativeness (see Grodner and Sedivy, 2011, which demonstrates that the contrastive inference from the use of a modifying adjective disappears when the speaker is explicitly unreliable in the other instructions that they give). Clearly this is a matter for further investigation.

Second, how far does the general claim hold that participants are sensitive to over-informativeness? Recall that sensitivity to over-informativeness was only found in experiments 2 and 3, which used pre-modifying syntactically unambiguous adjectives (as in the work by Sedivy et al., 1999 and Sedivy, 2003) and not in experiment 1, which used temporarily ambiguous PPs (cf. Spivey et al., 2002; Tanenhaus et al., 1995; Trueswell et al., 1999). The former result is sufficient evidence against Engelhardt et al.'s claim that participants only moderately adhere to Quantity maxims and are 'not troubled' when they encounter over-descriptions (2006:572). However, could it be that a more modest claim holds, namely that interlocutors are not sensitive to over-informativeness in cases of syntactic ambiguity? This is not implausible, although the burden of proof lies with the proponents of such a proposal. Given the results of experiment 2, we consider that there are strong reasons to believe that the lack of penalisation in experiment 1 is not due to a lack of sensitivity, but rather due to a preference for over-informativeness in certain contexts, as discussed earlier (e.g. section 1.2, paragraph 10).

Third, the RT data raise questions about the nature of the delay involved in making judgements on infelicitous utterances. Data from the current study do not tell us whether participants were simply noticing the over-informativeness as distinct from concise and optimal expressions, or whether they were actively inferring that the speaker using the RE 'the big star' also has another star in mind. That is, the current data cannot distinguish whether the delay is due to sensitivity to a departure from the norm, or to the generation of a contrastive inference. This is a matter which has arisen in the under-informativeness literature as well, especially sentence-verification studies such as Noveck and Posada (2003) and Bott and Noveck (2004). In these studies it has been found that participants are slower at rejecting pragmatically under-informative utterances such as 'some elephants have trunks'. However, it is not possible to know from the sentence-verification paradigm whether the delay is due to the participants noticing that the utterance is infelicitous (the speaker said 'some' whereas she should have said 'all') or whether participants generated the inference that some but not all elephants have trunks. Investigations using the visual-world paradigm may be fruitful in further investigating this matter.

Fourth, we cannot say with certainty whether the relative leniency with which hearers treat violations of Q-2 to Q-1 is based on purely theoretical considerations (as Grice proposed), or due to a stronger intolerance of those utterances which fail to establish reference to a unique target. Disentangling these issues requires a study of how sensitive participants are in cases where neither violation of Q-1 nor of Q-2 leads to reference failure. This can be done using modifications to the paradigm used in the Quantity implicature literature. For example, participants could be presented with a situation in which a protagonist performs an action, and then hear a sentence that attempts to describe it. Occasionally, the sentence is under-informative (e.g. they hear that 'the client bought the computer' in a situation where the client bought both the computer and the modem) or over-informative ('the client bought the fast computer' in a situation where there was only one computer in the display). In these cases we are dealing with a violation of Q-1 and Q-2 respectively, which, unlike previous work, do not lead to reference failure.

Fifth, no firm conclusions can be made at this point as to the causal relation between the lack of motivation for over-informing and the elicited sensitivity to Q-2. To draw such conclusions, we would require further studies that systematically manipulate the posited motivation for over-informing. The pilot experiment summarised in section 6 initiates this research strand with respect to one suggested reason for over-informing; future work should continue to load these factors one by one while testing sensitivity after each manipulation. Relevant factors might include the complexity of the array and the presence of ambiguity. If such a manipulation led to increased production of over-informativeness and decreased sensitivity to over-informing, this would constitute evidence both for the masking of Gricean sensitivity in Engelhardt et al.'s study and for the relative weakness of Gricean constraints compared to perceptual constraints in sentence processing.

6. Future directions

A remaining concern over Engelhardt et al.'s (2006) methodology was that certain attributes in their visual display (including attributes of the target) were deemed more salient that others. This could have led speakers to encode the attribute redundantly or hearers to expect such encoding and thus not penalise over-informative items. If one of the item-types in an array appears more frequently than other item-types, it is reasonable to assume that its salience is increased. Similarly, if one of the items is clearly different from its array-mates, for example because it appears in a container, its salience may also increase. In Engelhardt et al.'s (2006) study, target items were privileged in both these respects. Such configurations may play a major causal role in the production of over-informative descriptions, as salience would lead to increased perceptual activation of the attribute regardless of its contrastiveness. It is not unreasonable to further expect that in these conditions, hearers will also expect superfluous detail in referring expressions.

Salience effects in the production of referring expressions have been demonstrated by Carbary and Tanenhaus (2007), who found that in arrays in which one of the non-referents shared an attribute of the target (e.g. a striped cat non-referent and a striped shirt target referent) 25% of REs to the target were overspecified, compared with an 11% baseline rate when non-referents shared no attributes with the target. The authors conclude that the salience of a particular attribute is increased by the presence of that attribute elsewhere in the array, increasing the likelihood of it being encoded in an over-informative RE. We conducted a pilot production study alongside experiment 3, which used similar instructions but increased the salience of one of the attributes of the target referent by placing it next to a distractor which shared that attribute, e.g. an open bag next to an open box. The resulting REs exhibited a numerical difference towards increased over-informativeness, climbing from 14% in the baseline condition to 20% in the shared-attribute condition. This trend, coupled with Carbary and Tanenhaus's (2007) findings, tentatively suggests that visual salience of an attribute may encourage speakers to mention a non-discriminating adjective.

This finding may help account for the relatively high incidence of over-informativeness found in Engelhardt et al.'s (2006) production expt. 1. Specific containers could be argued to possess increased salience within this paradigm, due to multiple instantiations of them in the arrays, as well as the compositional nature of the target + container items. Furthermore, we note that nonlinguistic features added to the visual display numerically affect the degree of informativeness produced. Such factors may be at play in the expectation of increased levels of informativeness in Engelhardt et al.'s expt. 2 and in our replication reported as experiment 1. The numerical trend in our pilot study suggests that factors identified as favouring over-informativeness (visual salience, array complexity etc.) have modest effects in isolation, but taken together could considerably influence levels of informativeness in referring expressions. Future work from both the speaker and the comprehender's perspective should continue in the same spirit, testing the effect of individual factors on over-informativeness and thus contributing to a model of naturalistic conversation which takes these into account.

In some ways, the current study raises more questions than it answers. Having established that speakers and hearers are sensitive to under- and over-informativeness, experiments 2 and 3 rescue Gricean reasoning as a candidate for the list of constraints influencing sentence processing. Where Gricean considerations lie relative to other constraints is still very much up for debate, but three basic positions can be proposed:

(i) Gricean considerations are the main factor constraining the interpretation of ambiguous utterances. This prioritisation of Grice is projected from theoretical predictions relating to what the idealised interlocutor is expected to do.

(ii) Gricean considerations are a philosophical abstraction. Under-informativeness only appears to constrain referential interpretations because it is actually a reference resolution constraint. Over-informativeness is seldom relevant.

(iii) Gricean considerations form one constraint amongst many, with their relative weight not yet tested.

> One possibility is that Q-1 may have more weight than Q-2 due to their differential interaction with reference resolution constraints. In addition, if Gricean considerations have relatively little impact, this may be because they are only active when sufficient processing resources are available.

Position (i) can be eliminated due to the wealth of research documenting manifold constraints on sentence processing, e.g. referential world, syntactic frequency, minimal attachment, thematic roles etc. Experiments 2 and 3 above provide evidence against (ii) due to their documentation of sensitivity to Q-2, which does not coincide with reference resolution processes. Thus position (iii) is favoured at this point. However, extensive work is required along the lines discussed above in order to shed light on the interaction of Gricean considerations with other constraints. Such research will naturally feed into a mechanistic account of how people use referring expressions under real conversational conditions, potentially involving a ranked list of constraints including Gricean and nonlinguistic factors.

Acknowledgements

We are grateful to Paul Engelhardt and colleagues for sharing their stimuli and for valuable comments on this work, and also thank Chris Cummins, Henriëtte Hendriks, Ira Noveck, John Williams, and an anonymous reviewer for comments and suggestions on an earlier draft. The research reported in this article was submitted as part of the first author's Ph.D. thesis, and was funded by a Domestic Research Scholarship from the University of Cambridge.

Appendix 1. Item list for experiment 1.

Please see published paper or contact c.n.davies@leeds.ac.uk for appendix 1.

Appendix 2. Sample stimuli for experiment 2 (list 1 from 4)

Item	Target	Distractor / Competitor	Distractor	Distractor	Utterance: 'Pass me *X*'
Under-informative					
under-01	long skirt	short skirt	football	toilet	the skirt
under-02	closed bag	open bag	strawberry	zebra	the bag
under-03	old newspaper	recent newspaper	cherry	dog	the newspaper
under-04	short sock	long sock	lobster	spoon	the sock
under-05	stripy cup	spotty cup	tv	cucumber	the cup
under-06	tall vase	short vase	key	bus	the vase
under-07	new phone	old phone	hat	flower	the phone
under-08	small star	big star	house	chick	the star
under-09	sleeping baby	feeding baby	butterfly	tap	the baby
under-10	full glass	empty glass	iron	broom	the glass
Over-informative					

over-01	thin nail	fork	apple	cat	the thin nail
over-02	tall boot	carrot	tree	microphone	the tall boot
over-03	new phone	bike	compass	feather	the modern phone
over-04	old computer	watch	guitar	mirror	the old computer
over-05	intact egg	onion	camera	hairdryer	the unbroken egg
over-06	closed umbrella	onion	ice-cream	stool	the closed umbrella
over-07	open book	hamburger	fish	teddy	the open book
over-08	unsliced bread	toothbrush	toad	pineapple	the unsliced bread
over-09	unlit cigarette	tiger	plane	soap	the unlit cigarette
over-10	modern rug	banana	crab	toaster	the modern rug
Optimal + 1-referent					
opt1-01	banana	cube	fork	watch	the banana
opt1-02	comb	onion	hamburger	toothbrush	the comb
opt1-03	hammer	apple	tiger	tree	the hammer
opt1-04	pear	camera	ice-cream	guitar	the pear
opt1-05	sausage	toad	plane	crab	the sausage
opt1-06	boat	cat	microphone	feather	the boat
opt1-07	drum	stool	hairdryer	teddy	the drum
opt1-08	corkscrew	toaster	duck	soap	the corkscrew
opt1-09	flower	razor	anchor	drum	the flower
opt1-10	onion	cup	fridge	vase	the onion
Optimal + 2-referents					
opt2-01	tall jug	short jug	kettle	panda	the tall jug
opt2-02	glass mug	china mug	duck	corkscrew	the glass mug
opt2-03	adult penguin	baby penguin	razor	lamp	the adult penguin
opt2-04	square pan	round pan	pear	fridge	the square pan
opt2-05	fresh apple	rotten apple	comb	sausage	the fresh apple
opt2-06	big cookie	small cookie	anchor	hammer	the big cookie
opt2-07	big cube	small cube	drum	vase	the big cube
opt2-08	big hat	small hat	shoe	cube	the small hat
opt2-09	small anchor	big anchor	drum	boat	the small anchor
opt2-10	dry stone	wet stone	cup	door	the dry stone
Fillers					
clefts					
f1	big circle	small circle	cube	hammer	the big circle, pass me
f2	open door	closed door	carrot	cat	the open door, pass me
f3	cookie	tree	feather	compass	the cookie, pass me
f4	hairdryer	bike	microphone	mirror	the hairdryer, pass me
f5	kettle	watch	guitar	onion	the kettle, pass me
zero article					
f6	short sock	long sock	camera	elephant	pass me short sock
f7	big bottle	small bottle	fish	hamburger	pass me big bottle
f8	feather	ice-cream	pineapple	toothbrush	pass me feather

f9	pineapple	banana	toilet	toad	pass me pineapple
f10	onion	tiger	plane	soap	pass me onion
adjective-noun reversal					
f11	plastic cup	paper cup	stool	teddy	pass me the cup plastic
f12	tall jug	small jug	toaster	strawberry	pass me the jug tall
f13	broken chair	intact chair	cherry	crab	pass me the chair broken
f14	closed eye	open eye	giraffe	spoon	pass me the eye closed
f15	wooden table	glass table	pencil	candle	pass me the table wooden
scrambled word order					
f16	small shoe	big shoe	chick	bus	me shoe pass small the
f17	big car	small car	house	key	me car pass big the
f18	teddy	cucumber	hotdog	iron	me teddy pass the
f19	soap	hat	tap	flower	me soap pass the
f20	watch	cake	football	dog	me watch pass the

Appendix 3. Sample stimuli for experiment 3 (list 1 from 2)

Item	Target	Distractor/competitor (2-ref condition only)	Distractor	Distractor
1-ref01	thin nail	fork	apple	cat
1-ref02	tall boot	carrot	tree	microphone
1-ref03	new phone	bike	compass	feather
1-ref04	old computer	watch	guitar	mirror
1-ref05	intact egg	onion	camera	hairdryer
1-ref06	closed umbrella	onion	ice-cream	stool
1-ref07	open book	hamburger	fish	teddy
1-ref08	unsliced bread	toothbrush	toad	pineapple
1-ref09	unlit cigarette	tiger	plane	soap
1-ref10	modern rug	banana	crab	toaster
1-ref11	banana	cube	fork	watch
1-ref12	comb	onion	hamburger	toothbrush
1-ref13	hammer	apple	tiger	tree
1-ref14	pear	camera	ice-cream	guitar
1-ref15	sausage	toad	plane	crab
1-ref16	boat	cat	microphone	feather
1-ref17	drum	stool	hairdryer	teddy
1-ref18	corkscrew	toaster	duck	soap
1-ref19	flower	razor	anchor	drum
1-ref20	onion	cup	fridge	vase
2-ref01	closed bag	open bag	strawberry	zebra
2-ref02	old newspaper	recent newspaper	cherry	dog
2-ref03	stripy cup	spotty cup	tv	cucumber
2-ref04	new phone	old phone	hat	flower
2-ref05	sleeping baby	feeding baby	butterfly	tap
2-ref06	full glass	empty glass	iron	broom
2-ref07	long skirt	short skirt	football	toilet

2-ref08	short sock	long sock	spoon	lobster
2-ref09	tall vase	short vase	key	bus
2-ref10	small star	big star	house	chick
2-ref11	glass mug	china mug	duck	corkscrew
2-ref12	adult penguin	baby penguin	razor	lamp
2-ref13	square pan	round pan	pear	fridge
2-ref14	fresh apple	rotten apple	comb	sausage
2-ref15	dry stone	wet stone	cup	door
2-ref16	tall jug	short jug	kettle	panda
2-ref17	big cookie	small cookie	anchor	hammer
2-ref18	big cube	small cube	drum	vase
2-ref19	small hat	big hat	shoe	cube
2-ref20	small anchor	big anchor	drum	boat

References

Altmann, Gerry, 1998. Ambiguity in sentence processing. Trends in Cognitive Sciences 2, 146--152.

Altmann, Gerry, Kamide, Yuki, 1999. Incremental interpretation at verbs: restricting the domain of subsequent reference. Cognition 73, 247--264. Altmann, Gerry, Steedman, Mark, 1988. Interaction with context during human sentence processing. Cognition 30, 191--238.

Arts, Anja, 2004. Overspecification in instructive texts. Ph.D. Thesis, Tilburg University, The Netherlands. Wolf Publishers, Nijmegen.

Arts, Anja, Maes, Alfons, Noordman, Leo, Jansen, Carel, 2010. Overspecification facilitates object identification. Journal of Pragmatics 43, 361-- 374.

Boersma, Paul, Weenink, David, 2010. Praat: Doing Phonetics By Computer [Computer Program]. Version 5.1.40. , Retrieved from http://www. praat.org/ (13.07.10).

Boland, Julie, Tanenhaus, Michael, Garnsey, Susan, 1990. Evidence for the immediate use of verb control information in sentence processing. Journal of Memory and Language 29, 413--432.

Bott, Lewis, Noveck, Ira, 2004. Some utterances are underinformative: the onset and time course of scalar inference. Journal of Memory and Language 51, 437--457.

Breheny, Richard, Katsos, Napoleon, Williams, John, 2006. Are generalised scalar implicatures generated by default? An on-line investigation into the role of context in generating pragmatic inferences. Cognition 100, 434--463.

Carbary, Kathleen, Tanenhaus, Michael, 2007. Syntactic priming in an unscripted dialogue task. Poster presented at the 20th CUNY Conference on Human Sentence Processing, UCSD Center for Research in Language, La Jolla, California, March 2007.

Clark, Herbert, Bangerter, Adrian, 2004. Changing conceptions of reference. In: Noveck, I., Sperber, D. (Eds.), Experimental Pragmatics. Palgrave Macmillan, Basingstoke, pp. 25--49.

Clark, Herbert, Wilkes-Gibbs, Deanna, 1986. Referring as a collaborative process. Cognition 22, 1--39.

Clark, Herbert, Schreuder, Robert, Buttrick, Samuel, 1983. Common ground and the understanding of demonstrative reference. Journal of Verbal Learning and Verbal Behavior 222, 245--258.

Crain, Stephen, Steedman, Mark, 1985. On not being led up the garden path: the use of context by the psychological syntax processor. In: Dowty, D.R., Karttunen, L., Zwicky, A.M. (Eds.), Natural Language Parsing: Psychological, Computational, and Theoretical Perspectives. Cambridge University Press, Cambridge.

Davies, Catherine, Katsos, Napoleon, 2010. Over-informative children: production/comprehension asymmetry or tolerance to pragmatic violations? Lingua 120, 1956--1972.

De Neys, Wim, Schaeken, Walter, 2007. When people are more logical under cognitive load: dual task impact on scalar implicature. Experimental Psychology 54, 128--133.

Eikmeyer, H.-J., Ahlsen, E., 1998. The cognitive process of referring to an object: a comparative study of German and Swedish. In: Haukioja, T. (Ed.), Papers from the 16th Scandinavian Conference of Linguistics, vol. 60, Turku, Finland, pp. 75--86.

Engelhardt, Paul, Bailey, Karl, Ferreira, Fernanda, 2006. Do speakers and listeners observe the Gricean Maxim of Quantity? Journal of Memory and Language 54, 554--573.

Ferreira, Fernanda, Clifton, Charles, 1986. The independence of syntactic processing. Journal of Memory and Language 25, 348--368. Frazier, Lyn, 1978. On comprehending sentences: syntactic parsing strategies. Ph.D. Thesis, University of Connecticut.

Frazier, Lyn, 1987. Theories of sentence processing. In: Garfield, J. (Ed.), Modularity in Knowledge Representation and Natural Language Processing. MIT Press/Bradford Books, Cambridge, MA, pp. 493--522.

Frazier, Lyn, Rayner, Keith, 1982. Making and correcting errors during sentence comprehension: eye movements in the analysis of structurally ambiguous sentences. Cognitive Psychology 14, 178--210.

Freedle, Roy, 1972. Language users as fallible information-processors: implications for measuring and modelling comprehension. In: Freedle, R., Carroll, J. (Eds.), Language Comprehension and the Acquisition of Knowledge. Winston, Washington, DC.

Grice, H. Paul, 1975. Logic and conversation. In: Cole, P., Morgan, J.L. (Eds.), Syntax and Semantics, vol. 3. Academic Press, New York, reprinted in Studies in the Way of Words (1989). Harvard University Press, Cambridge, MA.

Grodner, Dan, Sedivy, Julie, 2011. The effects of speaker-specific information on pragmatic inferences. In: Pearlmutter, N., Gibson, E. (Eds.), The Processing and Acquisition of Reference. MIT Press, Cambridge, MA.

Guasti, Maria Teresa, Chierchia, Gennaro, Crain, Stephen, Foppolo, Francesca, Gualmini, Andrea, Meroni, Luisa, 2005. Why children and adults sometimes (but not always) compute implicatures. Language and Cognitive Processes 20, 667--696.

Hanna, Joy, Tanenhaus, Michael, Trueswell, John, 2003. The effects of common ground and perspective on domains of referential interpretation. Journal of Memory and Language 49, 43--61.

Heller, Daphna, Grodner, Daniel, Tanenhaus, Michael, 2008. The role of perspective in identifying domains of reference. Cognition 108, 831--836. Karmiloff-Smith, Annette, 1979. A Functional Approach to Child Language. CUP, Cambridge.

Katsos, Napoleon, Bishop, Dorothy, 2011. Pragmatic tolerance: implications for the acquisition of informativeness and implicature. Cognition 20, 67--81.

Koolen, Ruud, Gatt, Albert, Goudbeek, Martijn, Krahmer, Emiel, 2009. ‘Need I say more?’ On factors causing referential overspecification. In: van Deemter, K., Gatt, A., van Gompel, R., Krahmer, E. (Eds.), Proceedings of the Workshop on the Production of Referring Expressions: Bridging the gap between computational and empirical approaches to reference (PRE-CogSci 2009), Amsterdam, July 2009.

MacDonald, Maryellen, 1994. Probabilistic constraints and syntactic ambiguity resolution. Language and Cognitive Processes 9, 157--201. Mangold, Roland, Pobel, Rupert, 1988. Informativeness and instrumentality in referential communication. Journal of Language and Social Psychology 7, 181--191.

Noveck, Ira, 2001. When children are more logical than adults. Cognition 86, 253--282.

Noveck, Ira, Posada, Andres, 2003. Characterizing the time course of an implicature: an evoked potentials study. Brain and Language 85, 203--210.

Papafragou, Anna, Musolino, Julien, 2003. Scalar implicatures: experiments at the semantics/pragmatics interface. Cognition 86, 253--282. Paraboni, Ivandré, van Deemter, Kees, Masthoff, Judith, 2007. Generating referring expressions: making referents easy to identify. Computational Linguistics 33, 229--254.

Pechmann, Thomas, 1984. Accentuation and redundancy in children and adult’s referential communication. In: Bouma, H., Bouwhuis, D.G. (Eds.), Attention and Performance X: Control of Language Processes. Erlbaum, NJ, (Chapter 25).

Pechmann, Thomas, 1989. Incremental speech production and referential overspecification. Linguistics 27, 89--110.

Raaijmakers, Jeroen G.W., Schrijnemakers, Joseph M.C., Gremmenv, Frans, 1999. How to deal with ‘‘The language-as-fixed-effect fallacy’’: common misconceptions and alternative solutions. Journal of Memory and Language 41, 416--426.

Rubio-Fernández, Paula, Glucksberg, Sam, 2010. The cognitive costs and benefits of encoding and interpreting colour adjectives. Paper presented at the Euro-XPrag Workshop, Leuven, June 2010.

Schriefers, Herbert, Pechmann, Thomas, 1988. Incremental production of noun-phrases by human speakers. In: Zock, M., Sabah, G. (Eds.), Advances in Natural Language Generation. Pinter, London, pp. 172--179.

Sedivy, Julie, 2002. Invoking discourse-based contrast sets and resolving syntactic ambiguities. Journal of Memory and Language 46, 341--370. Sedivy, Julie, 2003. Pragmatic versus form-based accounts of referential contrast: evidence for effects of informativity expectations. Journal of Psycholinguistic Research 321, 3--23.

Sedivy, Julie, Tanenhaus, Michael, Chambers, Craig, Carlson, Gregory, 1999. Achieving incremental semantic interpretation through contextual representation. Cognition 712, 109--147.

Spivey, Michael, Tanenhaus, Michael, Eberhard, Kathleen, Sedivy, Julie, 2002. Eye movements and spoken language comprehension: effects of visual context on syntactic ambiguity resolution. Cognitive Psychology 45, 447--481.

Tanenhaus, Michael, Carlson, Gregory, Trueswell, John, 1989. The role of thematic structures in interpretation and parsing. Language and Cognitive Processes 4, 211--234.

Tanenhaus, Michael, Spivey-Knowlton, Michael, Eberhard, Kathleen, Sedivy, Julie, 1995. Integration of visual and linguistic information in spoken language comprehension. Science 268, 1632--1634.

Taraban, Roman, McClelland, James, 1988. Constituent attachment and thematic role assignment in sentence processing: influences of content-based expectations. Journal of Memory & Language 27, 597--632.

Trueswell, John, Tanenhaus, Michael, Kello, Christopher, 1993. Verb-specific constraints in sentence processing: separating effects of lexical preference from garden-paths. Journal of Experimental Psychology: Learning, Memory and Cognition 19, 528--553.

Trueswell, John, Sekerina, Irina, Hill, Nicole, Logrip, Marian, 1999. The kindergarten-path effect: studying on-line sentence comprehension in young children. Cognition 73, 89--134.

Viethen, Jette, Dale, Robert, 2006. Algorithms for generating referring expressions: do they do what people do? In: Proceedings of the Fourth International Natural Language Generation Conference, Sydney, Australia, July 2006, pp. 63--70.

Weiss, Ronald, Mangold, R., 1997. Meant with color but said without: when is the color of an object not named? Sprache & Kognition 16, 31--47. Whitehurst, Grover, 1976. Development of communication -- changes with age and modeling. Child Development 47, 473--482.