Over overinformativeness: rationally redundant referring expressions

Judith Degen, Caroline Graf, Robert X.D. Hawkins, Noah D. Goodman
{jdegen,ngoodman}@stanford.edu
Department of Psychology, 450 Serra Mall
Stanford, CA 94305 USA

September 25, 2016

Abstract

Referring is one of the most basic and prevalent uses of language. How do speakers choose from the wealth of referring expressions at their disposal? Rational theories of language use have come under attack for decades for not being able to account for the seemingly irrational overinformativeness ubiquitous in referring expressions. Here we present a novel production model of referring expressions within the Rational Speech Act framework that treats speakers as agents that rationally trade off cost and informativeness of utterances. Crucially, the assumption of deterministic meanings is relaxed. This allows us to capture a large number of seemingly disparate phenomena within one unified framework: the basic asymmetry in speakers' propensity to overmodify with color rather than size; the increase in overmodification in complex scenes; the increase in overmodification with atypical features; and the preference for basic level reference in nominal reference. The findings cast a new light on the production of referring expressions: rather than being wastefully overinformative, reference is rationally redundant. This implicates a production system geared towards communicative efficiency.

Keywords: reference; referring expressions; informativeness; probabilistic pragmatics; experimental pragmatics

Contents

1 Introduction

Reference to objects is one of the most basic and prevalent uses of language. But how do speakers choose amongst the wealth of referring expressions they have at their disposal? How does a speaker choose whether to refer to an object as *the animal*, *the dog*, *the dalmatian*, or *the big mostly white dalmatian*? The context within which the object occurs (other non-dogs, other dogs, other dalmatians) plays a large part in determining which features the speaker chooses to include in their utterance – speakers aim to be sufficiently informative to uniquely establish reference to the intended object. However, speakers' utterances often exhibit what has been claimed to be *overinformativeness*: referring expressions are often more specific than necessary for establishing unique reference, and they are so in systematic ways. However, providing a unified theory for speakers' systematic patterns of overinformativeness has so far proved elusive.

This paper is concerned with modeling precisely this choice of referring expression (RE). We restrict ourselves to definite descriptions of the form *the (ADJ?)+ NOUN*, that is, noun phrases that minimally contain the definite determiner *the* followed by a head noun. In addition, any number of adjectives may occur between the determiner and the noun.[1] A model of these REs will allow us to unify two domains in language production that have been typically treated as separate, and that have typically been treated as interesting for different reasons: the production of so-called overmodified referring expressions on the one hand, which a lot of literature in language production has been devoted to (Herrmann & Deutsch, 1976; Pechmann, 1989; Nadig & Sedivy, 2002; Maes, Arts, & Noordman, 2004; Engelhardt, Bailey, & Ferreira, 2006a; Arts, Maes, Noordman, & Jansen, 2011; Koolen, Gatt, Goudbeek, & Krahmer, 2011; Rubio-Fernandez, 2016); and the production of simple nominal expressions, which has so far mostly received attention in the concepts and categorization literature (Rosch, 1973; Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). In the following, we review some of the key phenomena and puzzles in each of these literatures which have for the most part been treated as unrelated. We then present a model of RE production within the Rational Speech Act (M. C. Frank & Goodman, 2012) framework, which treats speakers as boundedly rational agents who optimize the tradeoff between utterance cost and informativeness. Our key innovation is to relax the assumption that semantic truth functions are deterministic. [jd: one sentence here that inspires intuition, or a paragraph foreshadowing, making it seem like the obvious solution?] It is this crucial innovation that allows us to provide a unified explanation for a great number of seemingly disparate phenomena from the modified and nominal RE literature.

[1] In contrast, we will *not* provide a treatment of pronominal referring expressions, indefinite descriptions, names, or definite descriptions with post-nominal modification, though we offer some speculative remarks on how the approach outlined here can be applied to these cases.

1.1 Production of referring expressions: a case against rational language use?

How should a cooperative speaker produce referring expressions? Grice, in his seminal work, provided some guidance by formulating his famous conversational maxims, intended as a guide to listeners' expectations about good speaker behavior (Grice, 1975). His maxim of Quantity, consisting of two parts, requires of speakers to:

1. *Quantity-1:* Make your contribution as informative as s required (for the purposes of the exchange).

2. *Quantity-2:* Do not make your contribution more informative than is required.

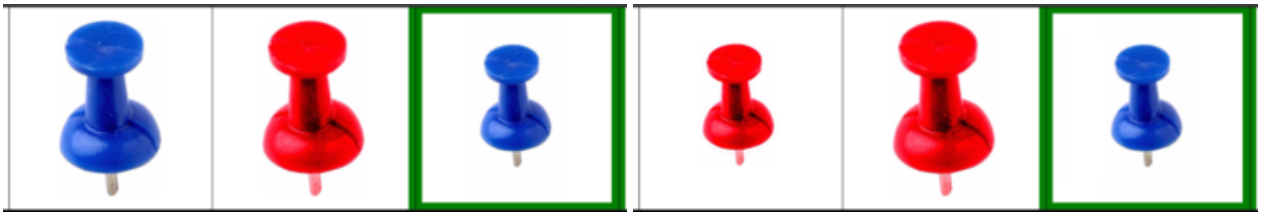
That is, speakers should aim to produce neither under- nor overinformative utterances. While much support has been found for the former (?, ?), speakers seem remarkably happy to systematically violate Quantity-2. In modified referring expressions, they routinely produce modifiers that do not uniquely establish reference (e.g., *the small blue thumbtack* instead of *the small thumbtack* in contexts like Figure 1a (?, ?)). In simple nominal expressions, speakers routinely choose to refer to an object with a basic level term even when a superordinate level term would have been sufficient for establishing reference (e.g., *the dog* instead of *the animal* in contexts like Figure 2d (?, ?)).

These observations have posed a challenge for theories of language production, especially those positing rational language use (including the Gricean one): why this extra expenditure of useless effort? Why this seeming blindness to the level of informativeness requirement? Many have argued from these observations that speakers are in fact not economical (?, ?). Some have derived a built-in preference for referring at the basic level from considerations of [jd: bla] and [jd: bla] (Rosch et al., 1976). Others have argued for salience-driven effects on willingness to overmodify (?, ?). In all cases, it is argued that informativeness cannot be the key factor in determining the content of speakers' referring expressions.

Here we revisit this claim and show that systematically relaxing the requirement of a deterministic semantics for referring expressions also systematically changes the informativeness of utterances. This results in a reconceptualization of what have been termed *overinformative referring expressions* as *rationally redundant referring expressions*. We begin by reviewing the phenomena of interest that a revised theory of definite referring expressions should be able to account for.

1.2 Modified referring expressions

Most of the literature on overinformative referring expressions has been devoted to the use of overinformative modifiers in modified referring expressions. The prevalent observation is that speakers frequently do not include only the minimal modifiers required for establishing unique reference, but often also include redundant modifiers (Pechmann, 1989; Nadig & Sedivy, 2002; Maes et al., 2004; Engelhardt et al., 2006a; Arts et al., 2011; Koolen et al., 2011). However, not all modifiers are created equal: there are systematic differences in the overmodification patterns observed for size adjectives (e.g., *big, small*), color adjectives (e.g., *blue, red*), material adjectives (e.g., *plastic, wooden*), and many others. Here we review some of the intriguing patterns of overmodification that have plagued that literature, focusing for the most part on size and color.



(a) Size sufficient. (b) Color sufficient.

Figure 1: Example contexts where size vs. color is sufficient for unique reference. A green border marks the intended referent.

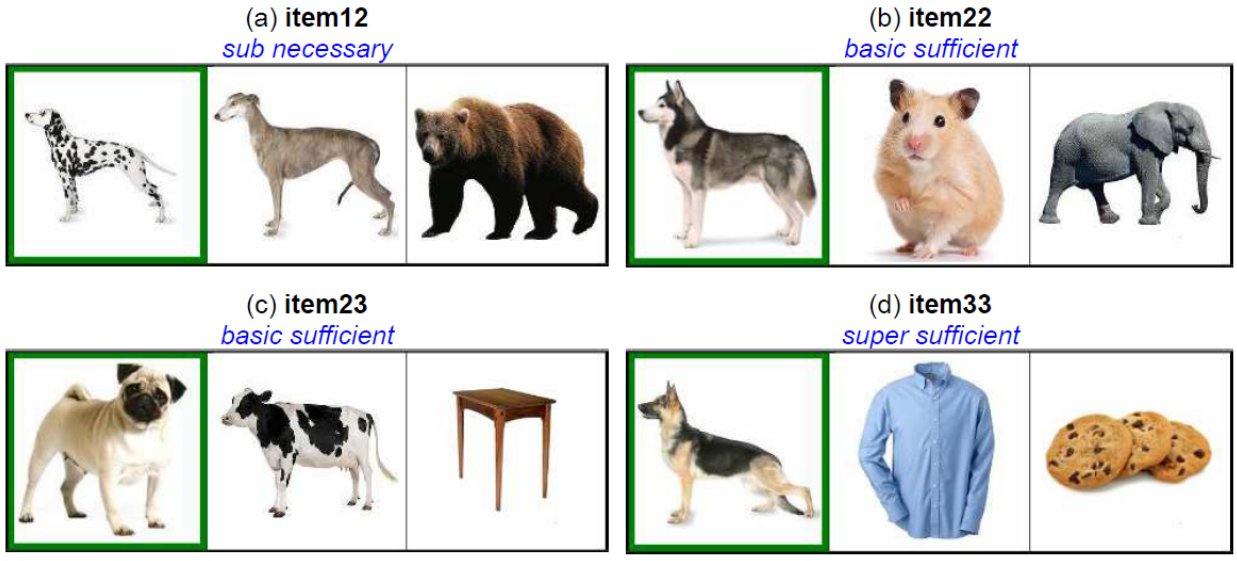


Figure 2: Example contexts in which different levels of reference are necessary for establishing unique reference to the target marked with a green border: sub (*dalmatian*, a), basic (*dog*, b, c), or super (*animal*, d).

Table 1: Proportions of minimally informative *color* or *size* and overinformative *color_size* mentions in color-sufficient vs. size-sufficient conditions across experiments.[jd: keep filling in.]

		Color sufficient			Size sufficient		
Study	Language	*color*	*size*	*color_size*	*color*	*size*	*color_size*
Pechmann (1989)	Dutch	99	0	1	9	36	55
Gatt et al. (2011)	English	92	0	8	3	17	80
Gatt et al. (2011)	Dutch	90	0	10	0	21	79
Our baseline study	English	94	0	6	2	52	46

1.2.1 Asymmetry in redundant use of color and size adjectives

In Figure 1a, singling out the object with the green border requires only mentioning its size, as in *the small thumbtack*. But it is now well-documented that speakers routinely include redundant color adjectives as in *the small blue thumbtack*, which do not uniquely single out the intended referent in these kinds of contexts (Pechmann, 1989; Belke & Meyer, 2002; Gatt, van Gompel, Krahmer, & van Deemter, 2011). However, the same is not true for size: in contexts like Figure 1b, where color is sufficient for unique reference (*the blue thumbtack*), speakers overmodify much more rarely with size. Table 1 shows proportions of color, size, and (overinformative) color-and-size mentions in conditions like those depicted in Figure 1 across different experiments. In all cases there is a preference for overmodifying with color but not with size.[2]

Explanations for this asymmetry have varied. Pechmann (1989) was the first to take the asymmetry as evidence for speakers following an incremental strategy of object naming: speakers initially start to articulate an adjective denoting a feature that listeners can quickly and easily recognize (i.e., color) before they have fully inspected the display and extracted the sufficient dimension. However, this would predict that speakers routinely should produce expressions like *the blue small thumbtack*, which violate the preference for size adjectives to occur before color adjectives in English (?, ?). While Pechmann did observe such violations in his dataset, most cases of overmodification did not constitute such violations, and he himself concludes that incrementality cannot (on its own) account for the asymmetry in speakers' propensity for overmodifying with color vs. size.

Another explanation for the asymmetry is that speakers try to produce modifiers that denote features that are reasonably easy for the listener to perceive, so that, even when a feature is not fully distinguishing in context, it at least serves to restrict the number of objects that could plausibly be considered the target. Indeed, there has been some support for the idea that overmodification can be beneficial to listeners by facilitating target identification (Arts et al., 2011; Rubio-Fernandez, 2016; Paraboni, van Deemter, & Masthoff, 2007).

[jd: try to find a quote from someone who says it's all just a matter of cost?]

There have been various attempts to capture the color-size asymmetry in computational natural language generation models. The earliest contenders for models of definite referring expressions like the Full Brevity algorithm (Dale, 1989) or the Greedy algorithm (Dale, 1989) focused only on discriminatory value – that is, an utterance's informativeness – in generating referring expressions, which resulted in an inability to capture the color-size asymmetry: the models only produced

[2]There is quite a bit of variation in the actual numbers. We will discuss this variation in the Discussion of Section 3.

the minimally specified expressions. Subsequently, the Incremental algorithm (Dale & Reiter, 1995) incorporated a preference order on features, with color ranked higher than size. The order is traversed and each encountered feature included in the expression if it serves to exclude at least one further distractor. This results in the production of overinformative color but not size adjectives. However, the resulting asymmetry is much greater than that evident in human speakers, and is deterministic rather than exhibiting the probabilistic production patterns that human speakers exhibit. More recently, the PRO model (Gatt, van Gompel, van Deemter, & Krahmer, 2013) has sought to integrate the observation that speakers seem to have a preference for including color terms with the observation that a preference does not imply the deterministic inclusion of said color term. The model is specifically designed to capture the color-size asymmetry: in a first step, the uniquely distinguishing property (if there is one) is first selected deterministically. In a second step, an additional property is added probabilistically, depending on both a salience parameter associated with the additional property and a parameter capturing speakers' eagerness to overmodify. If both properties are uniquely distinguishing, a property is selected probabilistically depending on its associated salience parameter. The second step proceeds as before.

However, while the PRO model – the most state-of-the-art computational model of human production of modified referring expressions – can capture the color-size asymmetry in and of itself, it is neither flexible enough to be extended straightforwardly to other modifiers beyond color and size, nor can it straightforwardly be extended to capture the more subtle systematicity with which the preference to overmodify with color changes based on various features of context. We delve into these more subtle patterns in the next two sections before presenting our alternative model within the Rational Speech Act framework.

1.2.2 Scene variation

Speakers' propensity to overmodify with color is highly dependent on features of the distractor objects in the context. In particular, as the variation present in the scene increases, so does the probability of overmodifying with color (Davies & Katsos, 2013; Koolen, Goudbeek, & Krahmer, 2013). How exactly scene variation is quantified differs between papers. One very clear demonstration of the scene variation effect was given by Koolen et al. (2013), who quantified scene variation as the number of feature dimensions along which objects in a scene vary. Over the course of three experiments, they compared a low-variation condition in which objects never differed in color with a high-variation condition in which objects differed in type, color, orientation, and size. They consistently found higher rates of overmodification with color in the high-variation (28-27%) than in the low-variation (4-10%) conditions.

The effect of scene variation on propensity to overmodify has typically been explained as the result of the demands imposed on visual search: in low-variation scenes, it is easier to discern the discriminating dimensions than in high-variation scenes, where it may be easier to simply start naming features of the target that are salient (Koolen et al., 2013).

The PRO model does not have a straightforward way of capturing the effect of scene variation on probability of overmodification. One way of doing so is to make the salience and overmodification parameters directly dependent on the amount of variation in the scene. However, this requires additional free parameters and makes the model prone to overfitting. [jd: elaborate? throw out?]

1.2.3 Feature typicality

Overmodification with color has been shown to be systematically related to the typicality of the color for the object. Building on work by ? (?), Westerbeek, Koolen, and Maes (2015) (and more recently, Rubio-Fernandez (2016)) have shown that the more typical a color is for an object, the less likely it is to be mentioned when not necessary for unique reference. For example, speakers never refer to a yellow banana as *the yellow banana*, but they sometimes refer to a brown banana as *the brown banana*, and they almost always refer to a blue banana as *the blue banana*. In fact, color typicality and probability exhibit a linear negative correlation (Westerbeek et al., 2015). Similar typicality effects have been shown for other (non-color) properties. For example, Mitchell (2013) showed that speakers are more likely to include an atypical than a typical property (either shape or material) when referring to everyday objects like boxes when mentioning at least one property was necessary for unique reference.

Whether speakers are more likely to mention atypical properties over typical properties because they are more salient to *them* or because they are trying to make reference resolution easier for the listener, for whom presumably these properties are also salient, is an open question (Westerbeek et al., 2015). Some support for the audience design account comes from a study by Huettig and Altmann (2011), who found that listeners, after hearing a noun with a diagnostic color (e.g., *frog*), are more likely to fixate objects of that diagnostic color (green), indicating that typical object features are rapidly activated and aid visual search. Nevertheless, the benefit for listeners and the salience for speakers might simply be a happy coincidence and speakers might not, in fact, be designing their utterances for their addressees. We will remain agnostic about the underlying reason for typicality effects [jd: will we, though? the model assumes that typicality affects the literal listener, who speakers reason about, so in a sense we're making a strong audience design claim.]

Irrespective of the source of typicality effects, it is unclear how the PRO model could accommodate them. As for the scene variation effects, it is possible to make the salience and overmodification eagerness parameters directly dependent on the typicality of the feature value for the object the speaker wants to refer to. However, as mentioned above, in the absence of a principled motivation for the way in which these parameters interact, this is simply an exercise in model-fitting without adding explanatory value. In addition, one is left with the task of explaining how scene variation and typicality should interact.

1.3 Nominal referring expressions

A problem related to the issue of how many additional features to include in a modified referring expression, but which has received much less attention in the language production literature, is that of deciding at which taxonomic level to refer to an object to in a simple nominal expression. That is, even in the absence of adjectives, a referring expression can be more or less informative: *the dalmatian* communicates more information about the object in question than *the dog*, which in turn is globally more informative than *the animal*. Thus, this choice can be considered analogous to the choice of adding more modifiers – in both cases, the speaker has a choice of being more or less specific about the intended referent. However, the choice of reference level in simple nominal referring expressions is also interestingly different from that of adding modifiers in that there is no additional word-level cost associated with being more specific – the choice is between different one-word utterances, not between utterances of different lengths (in words).

Table 2: List of effects a theory of referring expression production should account for.

Effect	Description
Color/size asymmetry	More redundant use of color adjectives than size adjectives
Scene variation	More redundant use of color adjectives with increasing scene variation
Color typicality	More redundant use of color adjectives with decreasing color typicality
Basic level preference	Preference for basic level term when superordinate level term sufficient
Subordinate level mention	Unnecessary use of sub level term when basic or super level sufficient

Nevertheless, cost affects the choice of reference level: in particular, speakers prefer more frequent nouns over less frequent ones (?, ?), and they prefer shorter ones over longer ones (?, ?). This may go part of the way towards explaining the well-documented effect from the concepts and categorization literature that speakers prefer to refer at the *basic level* (Rosch et al., 1976; Tanaka & Taylor, 1991). That is, in the absence of other constraints, even when a superordinate level term would be sufficient for establishing reference (as in Figure 2d), speakers prefer to say *the dog* rather than *the animal*.

However, there are nevertheless cases of contexts where either the superordinate (Figure 2d) or the basic level (Figure 2b and Figure 2c) term would be sufficient for unique reference, where speakers prefer to use the subordinate level term *the dalmatian*. This is the case when the object is a particularly good instance of the subordinate level term or a particularly bad instance of the basic level term. For example, penguins, which are rated as particularly atypical birds, are often referred to at the subordinate level *penguin* rather than at the basic level *bird*, despite the general preference for the basic level (Jolicoeur, Gluck, & Kosslyn, 1984).

1.4 Summary

In sum, the production of modified and simple nominal referring expressions is governed by a rich interplay of many factors, including an utterance's informativeness, its cost relative to alternative utterances, and the typicality of an object or its features. We are here especially interested in cases where speakers appear to be overinformative – either by adding more modifiers or by referring at a more specific level than necessary for establishing unique reference. A summary of the effects we will focus on in the remainder of the paper is provided in Table 2.

To date, there is no theory to account for all of these different phenomena; and no model has attempted to unify overinformativeness in the domain of modified and nominal referring expressions. We touched on some of the explanations that have been proposed for these phenomena. We also highlighted where computational models have been proposed for individual phenomena, and how they fall short. In the next section, we present the Rational Speech Act modeling framework, within which we will provide precisely the kind of theory that can account for at least all of the phenomena listed here and holds great promise for scaling up to many other overinformativeness phenomena.

2 Modeling speakers' choice of referring expression

Here we propose an extension to the production component of the Rational Speech Act (RSA M. C. Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013) modeling framework. This extension provides a principled explanation for the phenomena reviewed in the previous section and holds promise for being generalizable to many further production phenomena related to overinformativeness, which we discuss in the General Discussion. We proceed by first presenting the general framework in Section 2.1, and show why the most basic model cannot account for any of the phenomena outlined above, due to its strong focus on maximizing the informativeness of one-word expressions under a deterministic semantics. In Section 2.2 we introduce the crucial innovation: relaxing the assumption of a deterministic semantics. We show that the model can qualitatively account both for speakers' asymmetric propensity to overmodify with color rather than with size and (in Section 2.3) for speakers' propensity to overmodify more with increasing scene variation. In Section 3 we report an interactive reference game experiment which functions as a quantitative test of the model. In Section 4 we explore how the model captures feature typicality effects. In Section 5 we apply the model to the choice of simple nominal referring expressions and show that the qualitative preference for referring at the basic level (and the exceptions from that rule) emerges from the interaction of informativeness, utterance cost, and typicality. We test the model on a second interactive reference game experiment that provides data for a quantitative test of the model. For all cases we report – modified and nominal referring expressions – we find that introducing non-determinism into the semantic truth functions results in excellent quantitative fits to the data.

2.1 Basic RSA

As has been pointed out by Gatt et al. (2013), the basic Rational Speech Act model as formulated by M. C. Frank and Goodman (2012) cannot generate overinformative referring expressions for two reasons: first, it trivially cannot do so because it is limited to one-word utterances (see also Baumann, Clark, & Kaufmann, 2014). But even when allowing two-word (or n-word) utterances, the speaker's utility will never allow for producing more redundant than minimal referring expressions as long as words contribute non-negative costs to the overall utterance cost. To see this, and as a basis for the innovations introduced in Section 2.2 and Section 5.2 it is useful to reiterate the basic form of the model.

Intuitively, the production component of RSA aims to soft-maximize the utility of utterances, where utility is defined in terms of the contextual informativeness of an utterance, given each utterance's literal semantics. Formally, this is treated as a pragmatic speaker S_1 reasoning about a literal listener L_0, who can be described by the following formula:

$$P_{L_0}(o|u) \propto \ [\![u]\!](o). \tag{1}$$

The literal listener L_0 hears an utterance u from the set of available one-word utterances U in the context of a set of objects O and forms a distribution over the referenced object, $o \in O$. Here, $[\![u]\!](o)$ is the deterministic lexical meaning of the utterance u when applied to object o. That is, $P_{L_0}(o|u)$ returns a uniform distribution over all o denoted by u. For example, in the context shown in Figure 1a, $U = \{\textit{big}, \textit{small}, \textit{blue}, \textit{red}\}$ and $O = \{o_{\text{big\_blue}}, o_{\text{big\_red}}, o_{\text{small\_blue}}\}$. The values of $P_{L_0}(o|u)$ for each u are shown on the left in Table 3.

Table 3: Literal listener distributions $P_{L_0}(o|u)$ for each utterance u in the context depicted in Figure 1a, allowing only one-word utterances (left) or one- and two-word utterances (right).

	$o_{\text{big\_blue}}$	$o_{\text{big\_red}}$	$o_{\text{small\_blue}}$
big	.5	.5	0
small	0	0	1
blue	.5	0	.5
red	0	1	0

	$o_{\text{big\_blue}}$	$o_{\text{big\_red}}$	$o_{\text{small\_blue}}$
big	.5	.5	0
small	0	0	1
blue	.5	0	.5
red	0	1	0
big blue	1	0	0
big red	0	1	0
small blue	0	0	1

The pragmatic speaker in turn produces an utterance proportional to the utility of that utterance, where utility is a function of both the utterance's *informativeness* with respect to the literal listener $\ln P_{L_0}(o|u)$ and the utterance's *cost* $c(u)$:

$$P_{S_1}(u|o) \propto e^{\lambda \ln P_{L_0}(o|u) - \beta_c c(u)} \tag{2}$$

Both the informativeness and the cost term receive a weight.[3] Informativeness is weighted by λ. To understand the effect of λ, assume that costs are equal and the cost function can thus be disregarded. As λ approaches infinity, the speaker increasingly only chooses utterances that maximize informativeness; if λ is 0, informativeness is disregarded and the speaker chooses randomly from the set of all available utterances; if λ is 1, the speaker probability-matches. For our example in Table 3, if the speaker wants to refer to $o_{\text{small\_blue}}$ she has two semantically possible utterances, *small* and *blue*, where *small* is twice as informative as *blue*. She will produce *small* with the following probabilities as λ varies: $P_{S_1}(\textit{small}|o_{\text{small\_blue}}; \lambda = \infty) = 1$, $P_{S_1}(\textit{small}|o_{\text{small\_blue}}; \lambda = 1) = \frac{2}{3}$, $P_{S_1}(\textit{small}|o_{\text{small\_blue}}; \lambda = 0) = \frac{1}{4}$. Similarly, if we ignore informativeness and focus only on costs, any asymmetry in costs will be exaggerated with increasing β_c, such that the speaker will choose the least costly utterance with higher and higher probability as β_c increases.

As noted above, this model cannot generate redundant referring expressions for multiple reasons. One of these is trivial: U only contains one-word utterances. We can ameliorate this easily by allowing complex two-word utterances. We assume an intersective semantics for complex utterances u_{complex} consisting of two sub-utterances $u_{\text{size}} \in \{\textit{big}, \textit{small}\}$ and $u_{\text{color}} \in \{\textit{blue}, \textit{red}\}$, such that $[\![u_{\text{complex}}]\!] = [\![u_{\text{size}}]\!] \wedge [\![u_{\text{color}}]\!]$. The resulting literal listener distributions are shown on the right in Table 3.

Does this now allow for generating redundant referring expressions? To answer this, let's turn again to the case where the speaker wants to communicate the small blue object. There are now two utterances, *small* and *small blue*, which are both more informative than *blue* and equally informative to each other, for referring to the small blue object. Because they are equally informative in context, what we need is for the complex utterance to be the *cheaper* one in order to tilt the scales in its

[3] In fact, M. C. Frank and Goodman (2012) did not include a cost weight in their formulation and since they ultimately assumed equal costs for all utterances, they made no use of the cost function. Subsequent work has shown that taking into account utterance cost is necessary for modeling certain interpretation phenomena like cost-based quantity implicatures (Degen, Franke, & Jäger, 2013) and M-implicature (?, ?). The cost function will become important for our purposes in a little while.

favor. While this achieves the desired effect mathematically, the cognitive plausibility of complex utterances being cheaper than simple utterances is highly dubious. But this is the only circumstance under which overinformative referring expressions will be produced with a greater probability than minimally specified referring expressions. Thus, unless we want to introduce a highly dubious cost assumption into the model, we must look elsewhere to account for overinformativeness: to the computation of informativeness itself. This is what we turn to next.

2.2 RSA with non-deterministic semantics – emergent color-size asymmetry

Here we introduce the crucial innovation: rather than assuming a deterministic truth-conditional semantics that returns 1 (true) or 0 (false) for any combination of expression and object, we assume a non-deterministic semantics that can return intermediate values. That is, rather than assuming that an object is unambiguously big or unambiguously blue, we allow for a non-deterministic semantics, capturing that objects count as big or blue to varying degrees [jd: mention prototype theory and cite?]. In particular, consider some of the notable differences between color and size adjectives: color adjectives are considered *absolute adjectives* while size adjectives are inherently *relative* (?, ?). That is, while both size and color adjectives are vague, size adjectives are arguably context-dependent in a way that color adjectives are not – whether an object is big depends inherently on its comparison class; whether an object is red does not.[4] In addition, color as a property has been claimed to be inherently salient in a way that size is not (?, ?, ?). Finally, we have shown in recent work that color adjectives are much less subjective in their interpretation than size adjectives (?, ?). We use these observations as motivation for exploring the effects of the assumption that the semantics of size adjectives is inherently noisier than the semantics of color adjectives.

Formally, $[\![u]\!](o) = \exp(\text{fidelity}(u, o))$, where fidelity(u,o) returns a number between 0 and 1. The higher an utterance type's fidelity, the less noisy it is and the more likely it is to correctly pick out objects with the denoted property. The lower an utterance type's fidelity, the noisier the utterance is and the more likely it is to incorrectly pick out objects that don't exhibit the denoted property. We defer a discussion of the meaning of these fidelity values to the Discussion in Section 3.3. [jd: or straight to the GD?]

The effects of assuming non-deterministic truth functions in contexts like those depicted in Figure 1a and Figure 1b are visualized in Figure 3.[5] To orient the reader to the graph: the standard truth-functional semantics of the utterances are approximated where both fidelities are close to 1 (.999, right-most edge of each graph). In this case, the simple sufficient and complex redundant utterance are equally likely around .5 (because they are both equally informative and we are ignoring costs), and all other utterances are highly unlikely. The interesting question is under which circumstances, if any, the standard color-size asymmetry emerges: redundant *size-color* utterances are more likely than sufficient utterances where the fidelity of the sufficient dimension is lower than the fidelity of the insufficient dimension, for fidelities greater than .5.

Let's focus on a particular example. Assume the context in Figure 1a, where size is sufficient for uniquely singling out the target. If color fidelity is high (e.g., .999, dark blue line) and size fidelity is

[4] This is not entirely true, as has been pointed out by cite: red hair has a very different color than red wine, which in turn has a different color from a red bell pepper. If presented out of context, only the last red is likely to be judged as red (?, ?). For our purposes, it suffices that one can give a color judgment but not a size judgment for an object presented in isolation.

[5] Here we show the results for $\lambda = 30$ and no utterance cost (i.e., $\beta_c = 0$). For a visualization of model behavior under varying λs, see Appendix A.

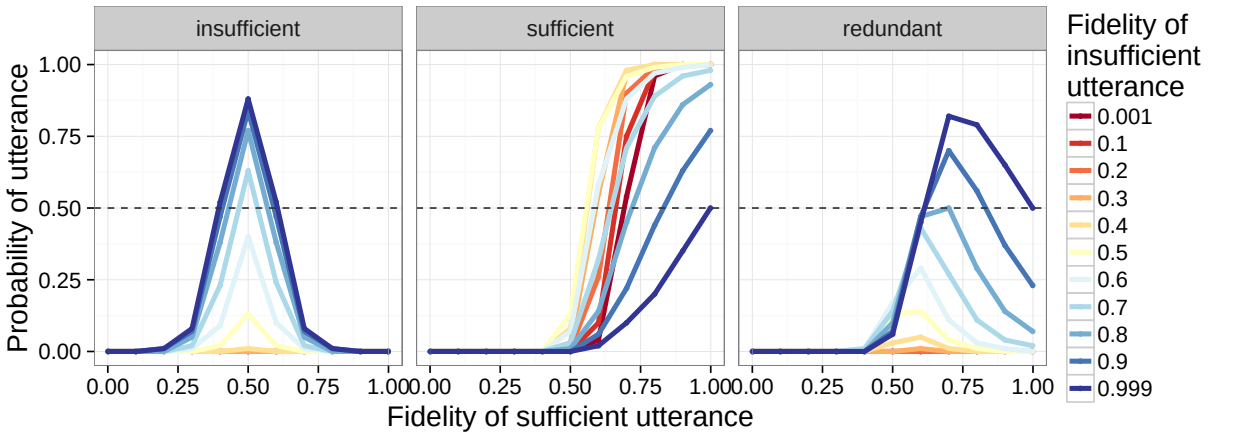


Figure 3: Probability of producing insufficient, sufficient, and redundant utterance in contexts as depicted in Figure 1a and Figure 1b, as a function of fidelity of sufficient and insufficient utterance type (for $\lambda = 30$ and $\beta_c = 0$).

relatively high, but not as high as color fidelity (e.g., .8 on x-axis), the probability of the redundant *size-color* utterance *small blue thumbtack* is $\approx .8$ and the probability of the simple *size* utterance *small thumbtack* is $\approx .2$. If we assume the same fidelity values for color and size in the color sufficient context in Figure 1b, the probability of the redundant *size-color* utterance is $\approx .05$ and the probability of the simple *color* utterance *blue thumbtack* is $\approx .95$. Thus, when size adjectives are noisier than color adjectives, the model produces overinformative referring expressions with color, but not with size – precisely the pattern observed in the literature. Indeed, these particular values are very similar to those found by Gatt et al. (2011). Note also that no difference in adjective *cost* is necessary for obtaining the overinformativeness asymmetry. However, assuming a greater cost for size than for color does further increase the observed asymmetry. We defer a discussion of costs to Section 3.1, where we infer the best parameter values (size and color cost and fidelity) given data from a reference game experiment.

A final observation regarding the probability of producing the insufficient utterance (e.g., *blue thumbtack* in the size sufficient contexts in Figure 1a). The probability of producing the insufficient utterance is very high where its fidelity is high and the fidelity of the sufficient utterance is intermediate. This is because intermediate fidelity values lead to an utterance being randomly interpreted correctly or incorrectly; that is, *small thumbtack* with fidelity(u_{size}) $= .5$ will apply equally to big and small objects in the context. The effect of this is that the literal listener returns a uniform distribution over all three objects in context upon observing *small thumbtack*, adding no information. In contrast, a literal listener that observes *blue thumbtack* assigns equal probability to the target and the color competitor, but lower probability to the distractor. Thus, even though the literal listener cannot distinguish between target and color competitor, the increased probability of correctly choosing the target by chance, due to the reduced probability of choosing the distractor, warrants the use of the insufficient *blue thumbtack* utterance.

To summarize, we have thus far shown that RSA with non-deterministic adjective semantics can give rise to the well-documented color-size asymmetry in the production of overinformative referring expressions when size adjectives are noisier than color adjectives. But this basic asymmetry is only one of many intriguing patterns in the literature on referring expressions, including effects of scene

variation and feature typicality, discussed in the Introduction. We turn to these phenomena next.

2.3 RSA with non-deterministic semantics – scene variation

As discussed above, increased scene variation has been shown to increase overinformativeness, but scene variation can be quantified in many different ways. For concreteness sake we simulate the conditions reported by Koolen et al. (2013), who quantified scene variation as the number of feature dimensions along which pieces of furniture in a scene varied: type (e.g., chair, fan), size (big, small), and color (e.g., red, blue).[6] In particular, we simulate the high and low variation conditions from their Experiments 1 and 2, reproduced in Figure 4.

In both conditions in both experiments, color was not necessary for establishing reference; that is, color mentions were always redundant. The two experiments differed in the dimension necessary for unique reference. In Exp. 1, only type was necessary (*fan* and *couch* in the low and high variation conditions in Figure 4, respectively). In Exp. 2, size and type were necessary (*big chair* and *small chair* in Figure 4, respectively). Koolen et al. (2013) found lower rates of redundant color use in the low variation conditions (4% and 9%) than in the high variation conditions (24% and 18%).

We generated model predictions for precisely these four conditions. Note that by adding the type dimension as a distinguishing dimension, we must allow for an additional type fidelity parameter.

Koolen et al. (2013) counted any mention of color as a redundant mention. In Exp. 1, this includes the simple redundant utterances like *blue couch* as well as complex redundant utterances like *small blue couch*. In Exp. 2, where size was necessary for unique reference, only the complex redundant utterance *small brown chair* was truly redundant. The results of simulating these conditions for $\lambda = 30$, $\beta_c = c(u_{\text{size}}) = c(u_{\text{color}}) = 1$, fidelity$(u_{\text{size}}) = .8$, fidelity$(u_{\text{color}}) = .999$, fidelity$(u_{\text{type}}) = .9$ are shown in Figure 5.[7]

For both experiments, the model retrieves the empirically observed effect of variation on the probability of redundant color mention: when variation is greater, redundant color mention is more likely. Note that the absolute values predicted by the model ($\approx 8\%$ to $\approx 75\%$) are different from the values observed by Koolen et al. (2013) ($\approx 4\%$ to $\approx 24\%$). This need not concern us here: our goal was to investigate whether, using the same parameter values that best fit the few data points from the Gatt et al. (2011) study, the model predicts the qualitative effect of scene variation on redundancy. Indeed it does.

Differences in exact values may stem from various sources. First, the best λ value to assume may differ from experiment to experiment. Second, fidelity values may differ between experiments. Indeed, assuming a lower color fidelity of .9 maintains the qualitative effects but lowers to highest probability of redundancy to .26. Importantly, the basic requirements to yield the empirical scene variation effect are that size, type, and color fidelities follow the following ranking: fidelity(u_{size}) $\leq$ fidelity(u_{type}) $<$ fidelity(u_{color}). If type fidelity is greater than color fidelity, the probability of redundantly mentioning color is close to zero and does not differ between variation conditions. This is because in those cases, color mention reduces, rather than adding, information about the target. Third, the values reported by Koolen et al. (2013) were averaged over many different items – here, we only reported model predictions for the example items they reported.

[6] They also included orientation (left-facing, right-facing) as a dimension along which objects could vary in certain cases. We ignore this dimension here for simplicity's sake – we simply wish to demonstrate that the model does indeed predict increased color redundancy with an increase in number of dimensions along which there is variation.

[7] See Appendix B for a visualization of model predictions under a fuller exploration of parameter combinations.

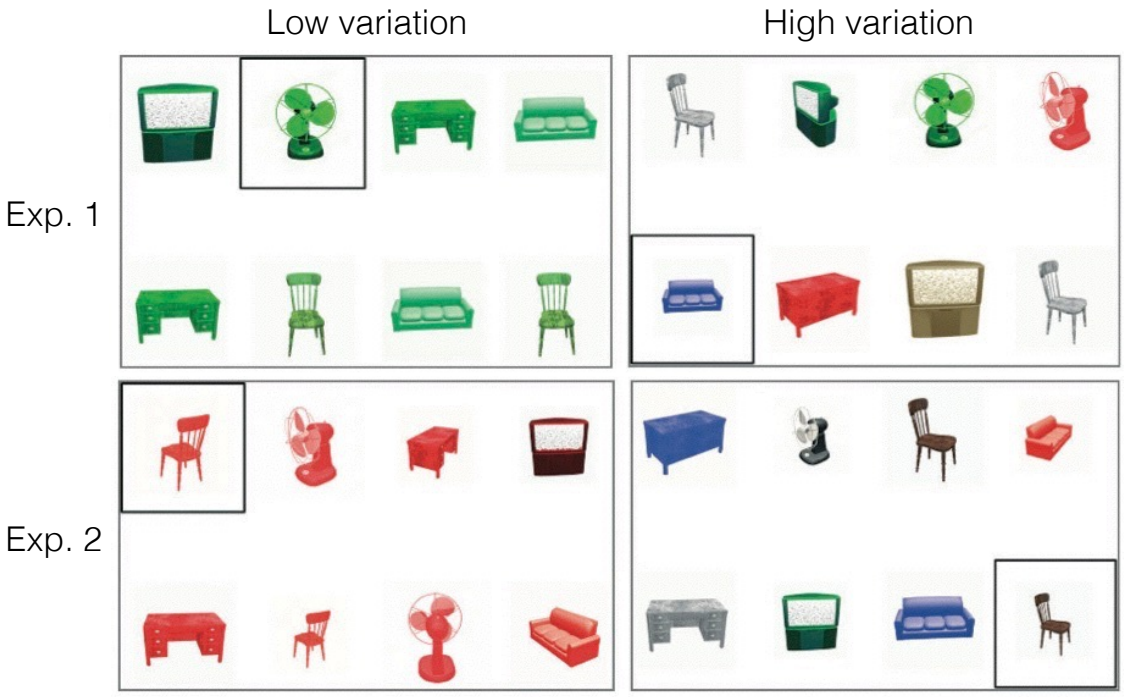


Figure 4: Contexts from Koolen et al.'s low variation (left column) and high variation (right column) conditions in Exp. 1 (top row) and Exp. 2 (bottom row).

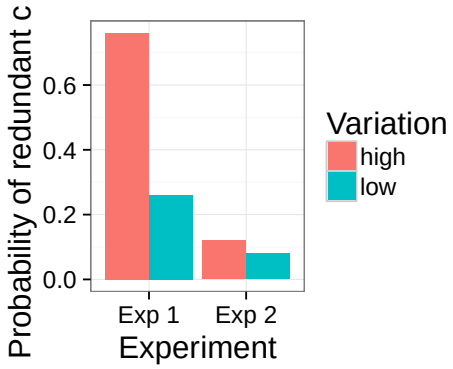


Figure 5: Model predicted probability of redundant color utterance in Koolen conditions for $\lambda = 30$, $\beta_c = c(u_{\text{size}}) = c(u_{\text{color}}) = 1$, fidelity$(u_{\text{size}}) = .8$, fidelity$(u_{\text{color}}) = .999$, fidelity$(u_{\text{type}}) = .9$.

These results are encouraging: RSA not only predicts a systematic color-size asymmetry in propensity to redundantly produce adjectives when size is noisier than color; it also predicts that there should be more redundant color mention as the number of dimensions along which objects in the scene vary increases. However, thus far we have only probed the model for qualitative effects from very few data points previously reported in the literature. Independently evaluating the utility of the model requires testing it on a large dataset. This is what we turn to next.

3 Non-deterministic RSA for modified referring expressions

Adequately assessing the explanatory value of RSA with non-deterministic truth functions requires evaluating how well it does at predicting the probability of various types of utterances in large datasets of naturally produced referring expressions. To this end we proceed in two steps. First we report the results of a web-based interactive reference game in which we systematically manipulate scene variation (in a somewhat different way than Koolen et al. (2013) did). We then perform Bayesian data analysis to generate model predictions, conditioning on the observed production data. This will both allow us a) to assess how likely the model is to generate the actually observed data – i.e., to obtain a measure of model quality – and b) to infer the posterior probability of parameter values – i.e., to understand whether the assumed asymmetries in adjective fidelity and/or cost discussed in the previous section are warranted.

3.1 Experiment 1: scene variation in modified referring expressions

We saw in Section 2.3 that non-deterministic RSA correctly predicts effects of scene variation on redundant adjective use. In particular, we saw that color is more likely to be used redundantly as the number of dimensions along which objects in a scene vary increases. However, we would like to a) go beyond a qualitative investigation of scene variation effects and also b) ask whether redundant size mention is also affected by scene variation. The notion of scene variation we employ is the proportion of distractor items that do not share the value of the insufficient feature with the target, that is, as the number of distractors n_{diff} that differ in the value of the insufficient feature divided by the total number of distractors n_{total}:

$$\text{scenevar} = \frac{n_{\text{diff}}}{n_{\text{total}}}$$

To explain, let's turn again to Figure 1a. Here, the target item is the small blue thumbtack and there are two distractor items: a big blue thumbtack and a big red thumbtack. Thus, for the purpose of establishing unique reference, size is the sufficient dimension and color the insufficient dimension. There is one distractor that differs from the target in color (the big red thumbtack) and there are two distractors in total. That is, scenevar $= \frac{1}{2} = .5$. Scene variation is minimal when all distractors are of the same color as the target, in which case it is 0. Scene variation is maximal when all distractors except for one (in order for the dimension to remain insufficient for establishing reference) are of a different color than the target. That is, scene variation may take on values between 0 and $\frac{n_{\text{total}-1}}{n_{\text{total}}}$, i.e, approaching but never reaching 1.

Using the same parameter values as in the previous two model explorations ($\lambda = 30$, $\beta_c = c(u_{\text{size}}) = c(u_{\text{color}}) = 1$, fidelity$(u_{\text{size}}) = .8$, fidelity$(u_{\text{color}}) = .999$), we generate model predictions for size-sufficient and color-sufficient contexts, varying scene variation by varying number of distractors (2, 3, or 4) and number of distractors that don't share the insufficient feature value. The resulting

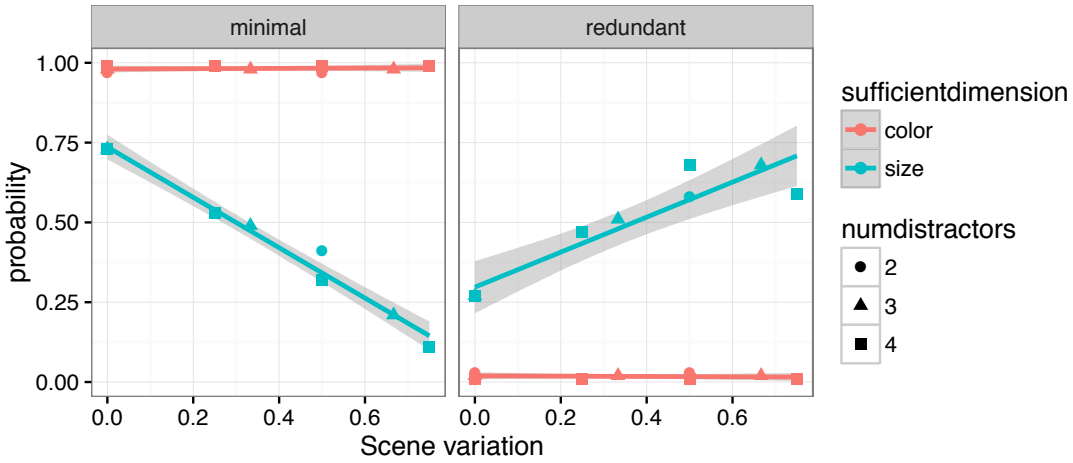


Figure 6: Probability of minimal and redundant utterance as a function of scene variation and sufficient dimension (for $\lambda = 30$, $\beta_c = c(u_{\text{size}}) = c(u_{\text{color}}) = 1$, fidelity$(u_{\text{size}}) = .8$, fidelity$(u_{\text{color}}) = .999$).

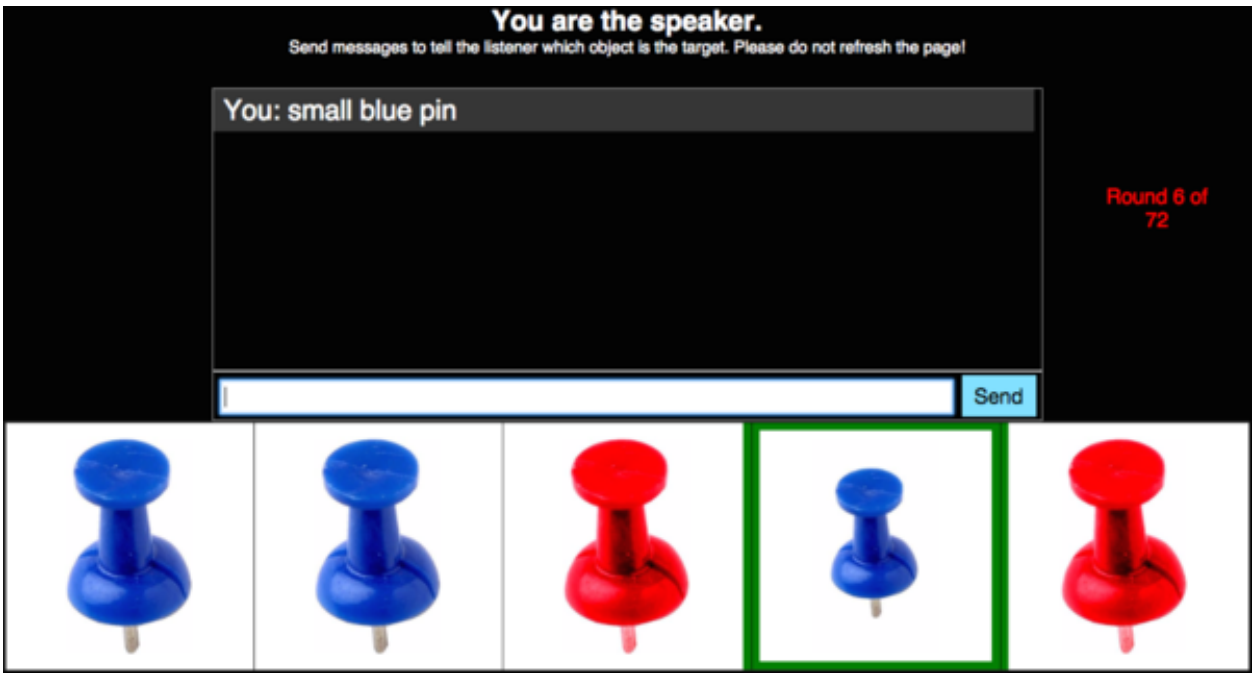
model predictions are shown in Figure 6: the probability of redundant adjective use increases with increasing scene variation when size is sufficient, but not when color is sufficient. This can be explained by noise distributions in the literal listener across contexts: in size-sufficient contexts, as the number of distractors of a different color than the target increases, using the relatively noiseless color term in addition to the more noisy size term reduces uncertainty about the target object. However, the same is not true of the color-sufficient contexts: there is very little uncertainty about the target upon observing the minimal color utterance – adding the size term only introduces more uncertainty about the target, regardless of the amount of scene variation. For slightly lower color fidelities a small increase in redundant size use is also predicted. In general: increased scene variation is predicted to lead to more redundant adjective use for less noisy adjectives.

To test non-deterministic RSA predictions, we conducted an interactive web-based written production study within a reference game setting.[8] Speakers and listeners were shown arrays of objects of that could vary in color and size. Speakers were asked to produce a referring expression to allow the listener to identify a target object. We manipulated the number of distractor objects in the grid, as well as the variation in color and size among distractor objects.

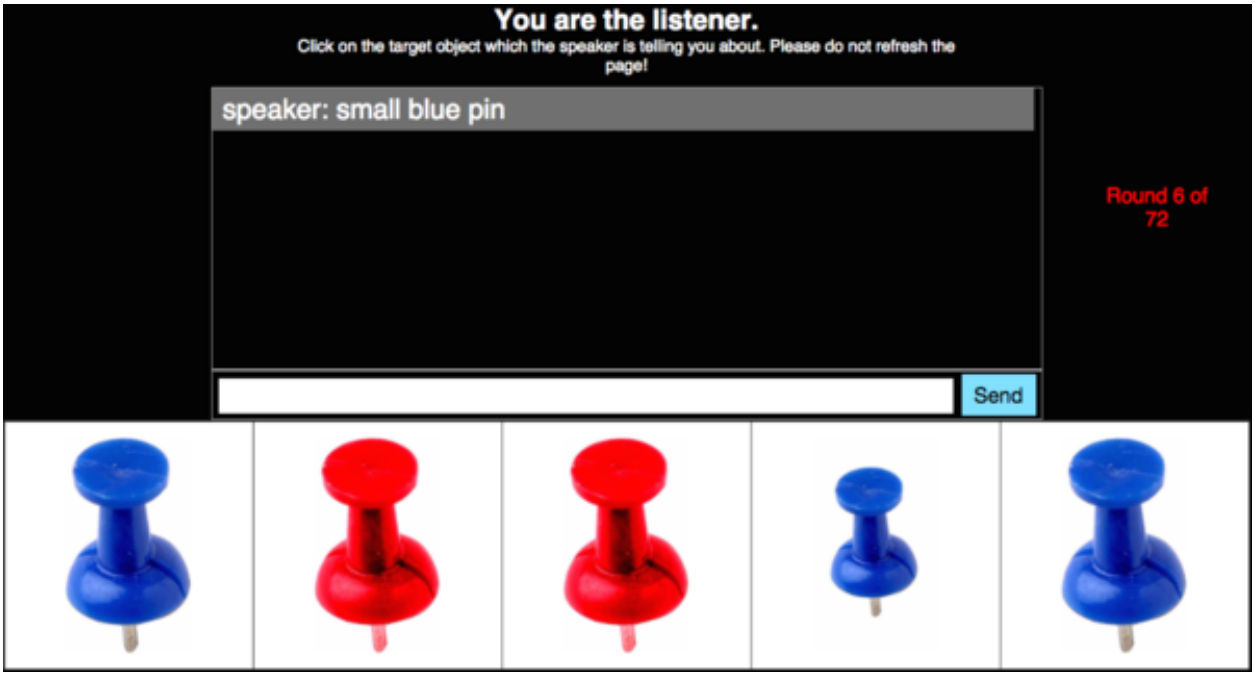
3.1.1 Method

Participants We recruited 58 pairs of participants (116 participants total) over Amazon's Mechanical Turk who were each paid $1.75 for their participation. Data from another 7 pairs who prematurely dropped out of the experiment and who could therefore not be compensated for their work, were also included. Here and in all other experiments reported in this paper, participants' IP address was limited to US addresses and only participants with a past work approval rate of at least 95% were accepted.

[8]See Appendix C for a validation of the general paradigm, in which we qualitatively replicate the findings of Gatt et al. (2011) with a different set of stimuli.



(a) Speakers' perspective



(b) Listeners' perspective.

Figure 7: Example displays from the (a) speaker's and the (b) listener's perspective on a *size-sufficient 4-2* trial.

Procedure Participants were paired up through a real-time multi-player interface (Hawkins, 2015). For each pair, one participant was assigned the speaker role and one the listener role. They initially received written instructions that informed participants that one of them would be the Speaker and the other the Listener. They were further told that they would see some number of objects on each round and that the speaker's task is to communicate one of those objects, marked by a green border, to the listener. They were explicitly told that using locative modifiers would be useless because the order of objects on their partner's screen would be different than on their own screen. Before continuing to the experiment, participants were required to correctly answer a series of questions about the experimental procedure. These questions are listed in Appendix D.

On each trial participants saw an array of objects. The array contained the same objects for both speaker and listener, but the order of objects was randomized and was typically different for speaker and listener. In the speaker's display, one of the objects – henceforth the *target* – was highlighted with a green border. See Figure 7 for an example of the listener's and speaker's view on a particular trial.

The speaker produced a referring expression to communicate the target to the listener by typing in a chat window. After pressing Enter or clicking the 'Send' button, the speaker's message was shown to the listener. The listener then clicked on the object they thought was the target, given the speaker's message. Once the listener clicked an object, a red border appeared around that object in both the listener and the speaker's display for 1 second before advancing to the next trial.

Both speakers and listeners could write in the chat window, allowing listeners to request clarification if necessary. Listeners could only click on an object and advance to the next trial once the speaker had sent a message.

Materials Participants proceeded through 72 trials. Of these, half were critical trials of interest and half were filler trials. On critical trials, we varied the feature that was sufficient to mention for uniquely establishing reference, the total number of objects in the array, and the number of objects that shared the non-sufficient feature with the target.

Objects varied in color and size. On 18 trials, color was a sufficient property for distinguishing the target. On the other 18 trials, size was sufficient. See Figure 7 for an example of a size-sufficient trial from both the speaker's and the listener's perspective.

We further varied the amount of variation in the scene by varying the number of distractor objects in each array (2, 3, or 4) and the number of distractors that did share the non-sufficient feature value with the target. That is, when size was the sufficiently distinguishing property, we varied the number of distractors that shared the same color as the target. This number had to be at least one, since otherwise the non-sufficient property would have been sufficient for uniquely establishing reference, i.e. it would not have been redundant to mention it. Each total number of distractors was crossed with each possible number of distractors that shared the non-sufficient property, leading to the following nine conditions: *2-1, 2-2, 3-1, 3-2, 3-3, 4-1, 4-2, 4-3,* and *4-4*, where the first number indicates the total number and the second number the shared number of distractors. Each condition occurred twice with each sufficient dimension. Objects never differed in type within one array (e.g., all objects are thumbtacks in Figure 7 but always differed in type across trials. Each object type could occur in two different sizes and two different colors. We deliberately chose photo-realistic objects of intuitively fairly typical colors. The 36 different object types and the colors they could occur with are listed in Appendix E.

Fillers were target trials from Exp. 2, a replication of (?, ?). Each filler item contained a three-

object grid. None of the filler objects occurred on target trials. Objects stood in various taxonomic relations to each other and required neither size nor color mention for unique reference. See Section **??** for a description of these materials.

3.1.2 Data pre-processing and exclusion

We collected data from 2171 critical trials. Of these, 33 (1.5%) were excluded because the listener did not select the target.

Because we did not restrict participants' utterances in any way, they produced many different kinds of referential expressions. Testing the model's predictions required, for each trial, either excluding it or classifying the produced utterance as an instance of a *color*-only mention, a *size*-only mention, or a *color-and-size* mention. To this end we conducted the following semi-automatic data pre-processing.

First, 33 trials on which the listener selected the wrong referent were excluded, leading to the elimination of 1.5% of trials. Then, an R script automatically checked whether the speaker's utterance contained a precoded color (i.e. *black, blue, brown, gold, green, orange, pink, purple, red, silver, violet, white, yellow*) or size (i.e. *big, bigger, biggest, huge, large, larger, largest, little, small, smaller, smallest, tiny*) term. In this way, 95.7 % of cases were classified as mentioning size and/or color. However, this did not capture that sometimes, participants produced meaning-equivalent modifications of color/size terms for instance by adding suffixes (e.g., *bluish*), using abbreviations (e.g., *lg* for *large* or *purp* for *purple*), or using non-precoded color labels (e.g., *lime* or *lavender*). Expressions containing a typo (e.g., *pruple* instead of *purple*) could also not be classified automatically. In the next step, one of the authors (CG) therefore manually checked the automatic coding to include these kinds of modifications in the analysis. This caught another 1.5% of trials. Most of the time, participants converged on a convention of mentioning simply the target's size and/or color, e.g., *purple* or *big blue*, without even using an article (e.g., *the*) or mentioning the object's type (e.g., *comb*). Articles were omitted in 93.1 % of cases and object types were omitted in 71.5 % of cases. We did not analyze this any further.

There were 50 cases (2.3%) in which the speaker made reference to the distinguishing dimension in an abstract way, e.g. *different color*, *unique one*, *ripest*, *very girly*, or *guitar closest to viewer*. While interesting as utterance choices,[9] these cases were excluded from the analysis. There were 3 cases that were nonsensical, e.g. *bigger off a shade*, which were also excluded. Finally, there were 6 cases where only the insufficient dimension was mentioned – these were excluded from the analysis reported in the next section, where we are only interested in minimal or redundant utterances, not underinformative ones, but were included in the Bayesian data analysis reported in Section 3.2. After the exclusion, 2079 cases classified as one of *color*, *size*, or *color-and-size* entered into the analysis.

3.1.3 Results

Proportions of redundant *color-and-size* and minimal *color* or *size* utterances are shown in Figure 8 alongside model results (to be explained further in Section 3.2). There are three main questions of interest: first, do we replicate the color/size asymmetry in probability of redundant adjective

[9]Certain participants seemed to have deliberately used this as a strategy even though simply mentioning the distinguishing property would have been shorter in most cases. In all, only 12 participants produced these kinds of utterances: one 18 times, one 8 times, one 6 times, two 3 times, one 2 times, and the remaining six only once each.

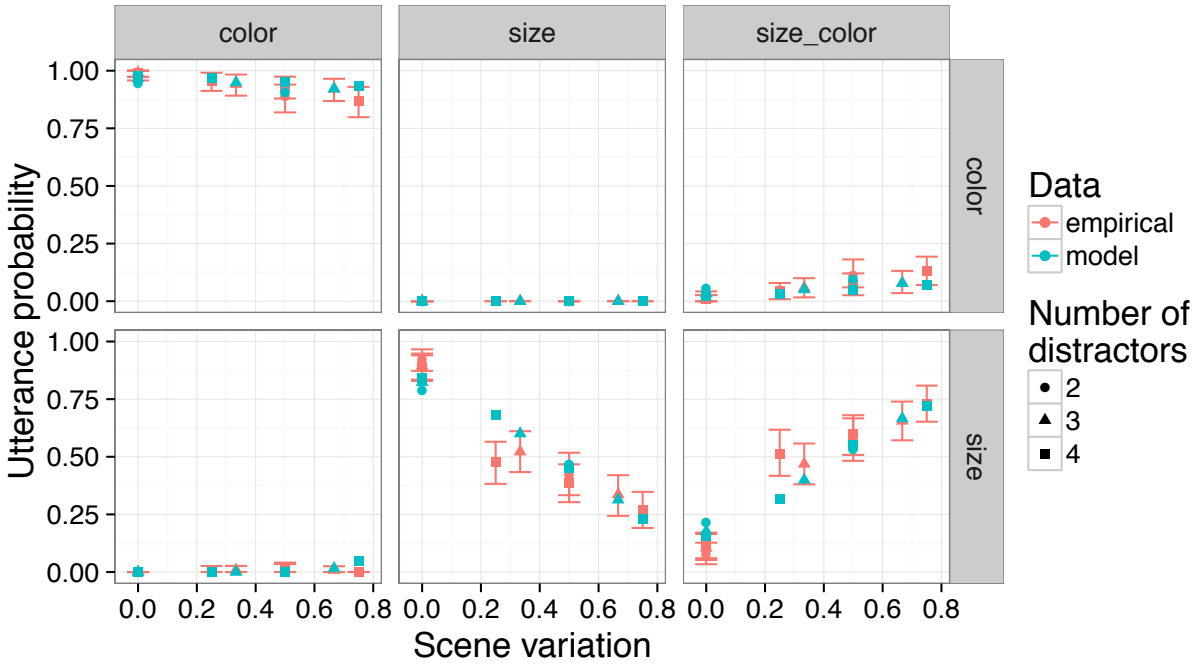


Figure 8: Empirical utterance proportions (red) alongside point-wise maximum a posteriori (MAP) estimates of the model's posterior predictives for utterance probability (blue) as a function of scene variation. Rows indicate the sufficient dimension, columns the produced utterance. Here and in all following plots, error bars indicate 95% bootstrapped confidence intervals.

use? Second, do we replicate the previously established effect of increased redundant color use with increasing scene variation? Third, is there an effect of scene variation on redundant size use and if so, is it smaller compared to that on color use, as is predicted under asymmetric color and size adjective fidelities?

We addressed all of these questions in one fell swoop by conducting a mixed effects logistic regression analysis predicting redundant over minimal adjective use from fixed effects of sufficient property (color vs. size), scene variation (proportion of distractors that does not share the insufficient property value with the target), and the interaction between the two. The model included the maximal random effects structure that allowed the model to converge: by-speaker and by-item random intercepts as well as by-speaker random slopes for scene variation.

We observed a main effect of sufficient property such that speakers were more likely to redundantly use color than size adjectives ($\beta = 3.61$, $SE = .23$, $p < .0001$), replicating the much-documented color-size asymmetry. We further observed a main effect of scene variation such that redundant adjective use increased with increasing scene variation ($\beta = 4.11$, $SE = .49$, $p < .0001$). Finally, we also observed a significant interaction between sufficient property and scene variation ($\beta = 3.03$, $SE = .81$, $p < .0002$). Simple effects analysis revealed that the interaction is driven by the scene variation effect being much smaller in the *color-sufficient* condition ($\beta = 2.59$, $SE = .78$, $p < .0009$) than in the *size-sufficient* condition ($\beta = 5.63$, $SE = .45$, $p < .0001$), as predicted under the assumption that size modifiers are noisier than color modifiers.

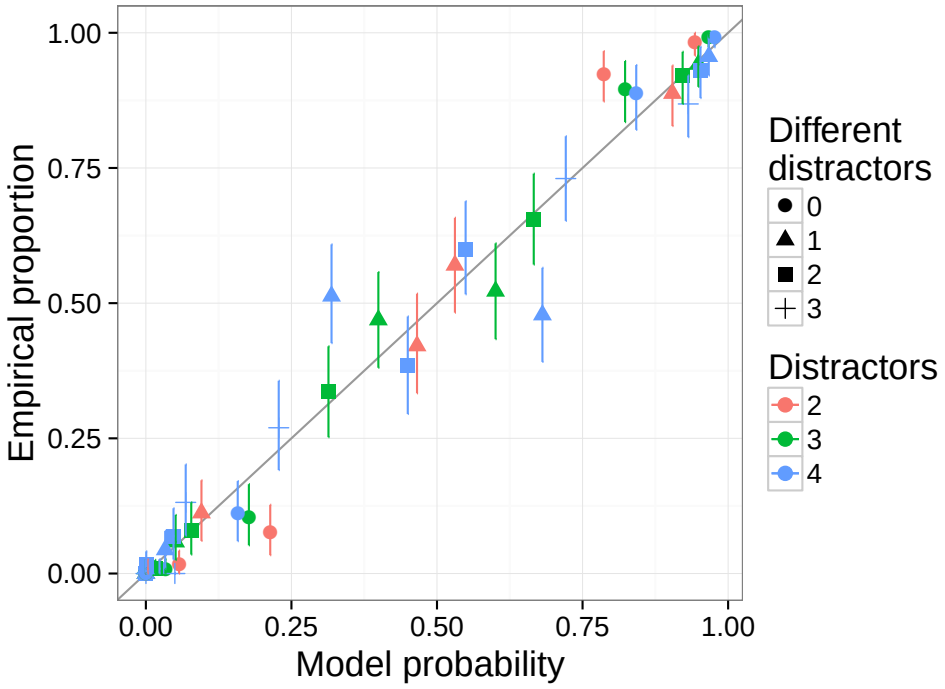


Figure 9: Scatterplot of point-wise maximum a posteriori (MAP) estimates of the model's posterior predictives against empirical proportions (r =.99).

3.2 Model evaluation: scene variation

We performed Bayesian Data Analysis to generate model predictions and infer likely parameter values, conditioning on the observed production data (coded into *size*, *color*, and *size-and-color* utterances as described above) and integrating over the following free parameters: color fidelity f_c, size fidelity f_s, color cost c_{color}, size cost c_{size}, cost weight β_{cost}, and speaker rationality parameter λ. We assumed uniform priors for each parameter: $f_c \sim \mathcal{U}(0,1)$, $f_s \sim \mathcal{U}(0,1)$, $c_{color} \sim \mathcal{U}(0,2)$, $c_{size} \sim \mathcal{U}(0,2)$, $\beta_{cost} \sim \mathcal{U}(0,10)$, $\lambda \sim \mathcal{U}(0,40)$. We implemented both the cognitive and data-analysis models in the probabilistic programming language WebPPL (Goodman & Stuhlmüller, electronic). Inference for the cognitive model was exact, while we used Markov Chain Monte Carlo (MCMC) to infer posteriors for the six free parameters.

Point-wise maximum a posteriori (MAP) estimates of the model's posterior predictives for each combination of utterance, sufficient dimension, number of distractors, and number of different distractors (collapsing across different items) are compared to empirical data in Figure 9. At this level, the model achieves a correlation of $r = .99$. Looking at results additionally on the by-item level yields a correlation of $r = .85$. The model thus does a very good job of capturing the quantitative patterns in the data. This can also be seen in Figure 8, where model predictions are plotted alongside the empirical proportions by condition. The only clear flaw is that the model predicts greater redundant adjective use than empirically observed when there is no scene variation at all.

Parameter posteriors are shown in Figure 10. Crucially, the fidelity of color is inferred to be higher than that of size – there is no overlap between the 95% highest density intervals (HDIs) for

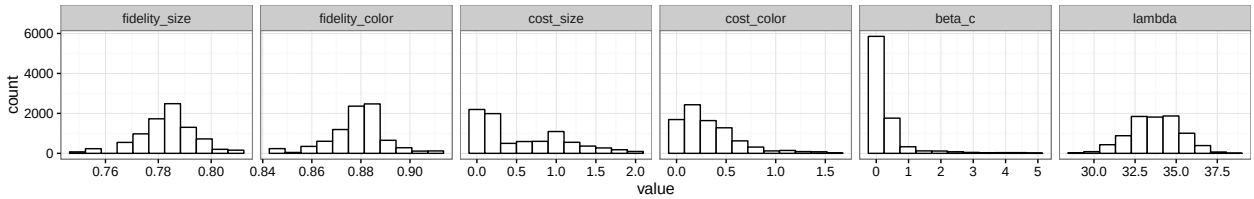


Figure 10: Posterior distribution over model parameters. Maximum a posteriori (MAP) $f_s = 0.79$, 95% highest density interval (HDI) = [0.76,0.80]; MAP $f_c = 0.88$, HDI = [0.86,0.91]; MAP $c_{\text{size}} = .08$, HDI = [0, 1.5]; MAP $c_{\text{color}} = 0.07$, HDI = [0,0.9]; MAP $\beta_c = 0.04$, HDI = [0,1.6]; MAP $\lambda = 34.0$, HDI = [30.8,36.5]

the two parameters. That is, size modifiers are inferred to be noisier than color modifiers. The relatively high inferred λ suggests that this difference in fidelity contributes substantially to the observed color-size asymmetries in redundant adjective use. As for cost, there is a lot of overlap in the inferred cost of size and color modifiers, suggesting that no cost difference is necessary to obtain the color-size asymmetry and the scene variation effects. These results are compatible with previous claims cite cite that part of the explanation for the color-size asymmetry stems from the low cognitive cost involved in producing color modifiers compared to size modifiers. However, the results do suggest that a cost asymmetry is not the driving force behind the asymmetry in redundant adjective use. Note further that the asymmetry cannot be reduced to cost differences: in Section 2.2 we showed that the color-size asymmetry in redundant adjective use requires an asymmetry in modifier fidelity. An asymmetry in cost only serves to further enhance the asymmetry brought about by the fidelity asymmetry, but cannot carry the redundant use asymmetry on its own.

3.3 Discussion

As should be apparent from this section, non-deterministic RSA provides an excellent fit to data of freely produced modified referring expressions. In particular, we have shown that the crucial element in obtaining the much-documented color-size asymmetry in the propensity to overmodify is that the semantic truth functions of size adjectives be noisier than those of color adjectives. Asymmetries in cost of adjectives only serve to further enhance the asymmetries resulting from asymmetries in utterance fidelity. In addition, we showed that asymmetric effects of scene variation on overmodification are also well captured by non-deterministic RSA: scene variation leads to a greater increase in overmodification with less noisy than with more noisy modifiers.

Some readers may have found themselves wondering about the status of the fidelity term: are we claiming that color modifiers have inherently higher fidelity than size modifiers? Is the difference constant? What if the color modifier is a less well known one like *mauve*? The way we have set the model up thus far, there would indeed be no fidelity difference between *red* and *mauve*. Moreover, the model is not equipped to handle potential object-level idiosyncracies in fidelity such as the typicality effects discussed in Section 1.2.3. We defer a fuller discussion of the status of the fidelity term to the General Discussion and turn first to non-deterministic RSA's potential for capturing typicality effects.

Table 4: Hypothetical fidelity values for utterances (rows) as applied to objects (columns).

Utterance	yellow banana	brown banana	blue banana	other
banana	.9	.35	.1	.015
yellow banana	.99	.015	.015	.015
brown banana	.015	.99	.015	.015
blue banana	.015	.015	.99	.015
other	.015	.015	.015	.99

4 Feature typicality

In Section 3 we showed that non-deterministic RSA successfully captures both the basic asymmetry in overmodification with color vs. size as well as effects of scene variation, quantified in various different ways. But in Section 1.2.3 we discussed a further characteristic of speakers' overmodification behavior: speakers are more likely to redundantly produce modifiers that denote atypical rather than typical object features, i.e., they are more likely to refer to a blue banana as a *blue banana* rather than as a *banana*, and they are more likely to refer to a yellow banana as a *banana* than as a *yellow banana* (Sedivy, 2003; Westerbeek et al., 2015). Non-deterministic RSA as we have set it up thus far does not capture this asymmetry: it knows that a particular modifier is a color modifier with a particular fidelity; it does not know anything about the typicality of the denoted properties for the referent.

We would like to warn and disillusion the reader upfront: we will not solve the problem of how to get overmodification behavior from the typicality of features compositionally. This is a problem for all theories of modification (?, ?). However, we would like to offer a proof of concept showing that, if the non-determinism in the RSA semantics is not at the adjective type (color, size) level, but instead at the level of combinations of referring expressions and objects, the model produces precisely the sorts of typicality effects reported in the literature.

Let us elaborate. Where before we took a fidelity to be a number between 0 and 1 indicating how likely a type of modifier (size, color) was to correctly apply to an object, we now treat it as indicating how good an instance of a particular referring expression the object in question is. For example, take the banana case: assume three contexts of objects with yellow, brown, and blue objects. Assume further that one of the objects is a banana, and the only difference between the three contexts is whether the banana is blue, brown, or yellow. In every context there is another object of the same color as the banana, so color is redundant, while there are no other bananas, so object category mention is sufficient for reference. Assume further the fidelity values shown in Table 4. These values should be read as follows: a yellow banana is a very good or typical instance of a*banana* – *banana* applied to yellow bananas has a high fidelity of .9. In contrast, brown bananas are less typical instances of *banana*s (.35), and blue bananas are highly atypical *banana*s (.1) but still better than objects of an other non-banana type (.015). Going along the diagonal, you can see that we assume for each remaining utterance that its fidelity is very high (.99) when applied to an object in its extension and very low otherwise (.015).

With $\lambda = 12$ and $\beta_c = 5$ (that is, both informativeness and utterance cost receive a substantial weight), the resulting speaker probabilities for the (minimal) *banana* are .99, .37, and .05, respectively, to refer to the yellow banana, the brown banana, and the blue banana. In contrast, the

resulting speaker probabilities for the redundant *yellow banana*, *brown banana*, and *blue banana* are .01, .63, and .95, respectively. That is, redundant color mention increases with decreasing fidelity of the simple *banana* utterance.

So far we have shown that non-deterministic RSA can capture typicality effects in principle if we assume that fidelity does not operate at the adjective type level but instead captures the typicality of an object for the alternative (minimal and redundant) referring expressions. If an object is more typical for the redundant expression than for the minimal expression, then the bigger the difference in typicality, the greater the relative informativeness of the redundant expression, and the greater the probability of it being produced.

We can now ask whether taking into account this more fine-grained notion of non-deterministic semantics plays a role in the dataset collected in Exp. 1. A note upfront: the stimuli for Exp. 1 were specifically designed to be realistic objects; that is, very low typicality values or even a large degree of variation in typicality would be surprising. Nevertheless, it is plausible that typicality differences exist. If they do, there are two interesting questions to ask: first, do we replicate the typicality effects reported in the literature – that is, are less typical objects more likely to lead to redundant adjective use than more typical objects? Second, does including empirically elicited typicality values at the object-utterance level further improve the quality of the RSA model? We address the first question in Section 4.1 and the second question in Section 4.2.

4.1 Experiment 1a: Typicality effects in Exp. 1

To assess whether we replicate the color typicality effects previously reported in the literature (Sedivy, 2003; Westerbeek et al., 2015; Rubio-Fernandez, 2016), we elicited color typicality norms for each of our items and then included typicality as an additional predictor of redundant adjective use in the regression analysis reported previously.

4.1.1 Methods

Participants We recruited 60 participants over Amazon's Mechanical Turk who were each paid $0.25 for their participation.

Procedure and materials On each trial, participants saw one of the big versions of the items used in Exp. 1 and were asked to answer the question "How typical is this for an X?" on a continuous slider with endpoints labeled "very atypical" to "very typical." X was a referring expression consisting of either only the correct noun (e.g., *stapler*) or the noun modified by the correct color (e.g., *red stapler*). Figure 11 shows an example of a modified trial.

Each participant saw each of the 36 objects once. An object was randomly displayed in one of the two colors it occurred with in Exp. 1 and was randomly displayed with either the correct modified utterance or the correct unmodified utterance, in order to obtain roughly equal numbers of object-utterance combinations.

Importantly, we only elicited typicality norms for unmodified utterances and utterances with color modifiers, but not utterances with size modifiers. This was because it is virtually impossible to obtain size typicality norms for objects presented in isolation, due to the inherently relational nature of size adjectives. Consequently, we only test for the effect of typicality on *size-sufficient* trials.

Figure 11: A modified example trial from the typicality elicitation experiment.

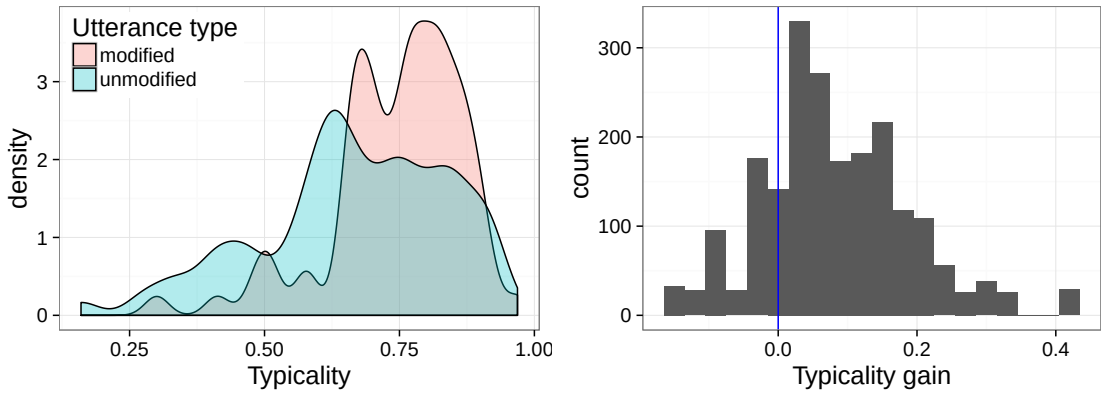


Figure 12: Typicality densities for modified and unmodified utterances (left) and histogram of typicality gains (differences between modified and unmodified typicalities, right).

4.1.2 Results and discussion

We coded the slider endpoints as 0 ("very atypical") and 1 ("very typical"), essentially treating each response as a typicality value between 0 and 1. For each combination of object, color, and utterance (modified/unmodified), we computed that item's mean. Mean typicalities were generally lower for unmodified than for modified utterances: mean typicality for unmodified utterances was .67 (sd=.17, mode=.76) and for modified utterances .75 (sd=.12, mode=.81). This can also be seen on the left in Figure 12. Note that, as expected given how the stimuli were constructed, typicality was generally skewed towards the high end, even for unmodified utterances. This means that there was not much variation in the difference in typicality between modified and unmodified utterances. We will refer to this difference as *typicality gain*, reflecting the overall gain in typicality via color modification over the unmodified baseline. As can be seen on the right in Figure 12, in most cases typicality gain was close to zero.

This makes the typicality analysis difficult: if typicality gain is close to zero for most cases

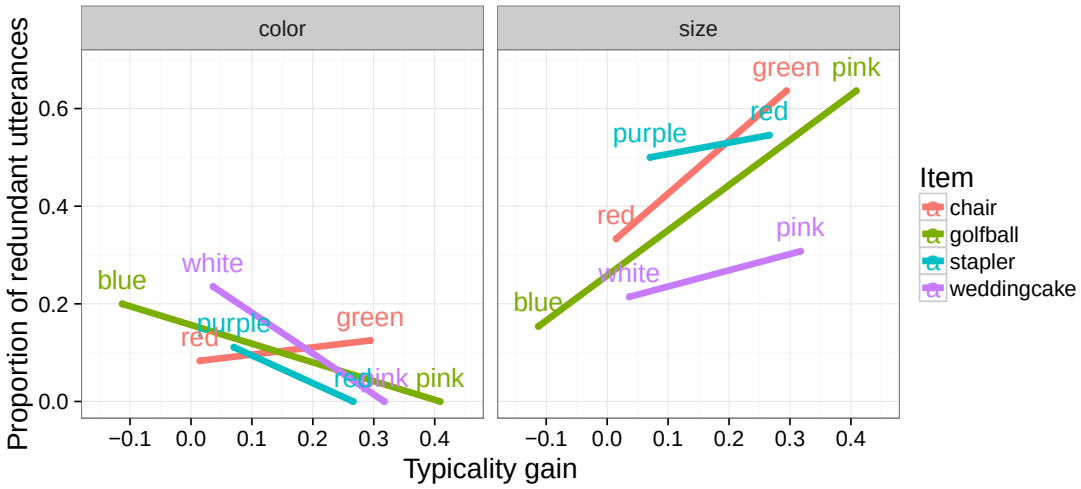


Figure 13: Utterance probability for four items as a function of difference in typicality between modified and unmodified utterance (x-axis) and sufficient dimension (columns).

(and, taking into account confidence intervals, effectively zero), it is hard to evaluate the effect of typicality on redundant adjective use. In order to maximize power, we therefore conducted the analysis only on those items for which for at least one color the confidence intervals for the modified and unmodified utterances did not overlap. There were only four such cases: *(pink) golfball*, *(pink) wedding cake*, *(green) chair*, and *(red) stapler*, for a total of 231 data points.

Predictions differ for size-sufficient and color-sufficient trials. Given the typicality effects reported in the literature and the predictions of non-deterministic RSA, we expect greater redundant color use on size-sufficient trials with *increasing* typicality gain. The predictions for redundant size use on color-sufficient trials are unclear from the previous literature. Non-deterministic RSA, however, predicts greater redundant size use with *decreasing* typicality gain: small color typicality gains reflect the relatively low out-of-context utility of color. In these cases, it may be useful to redundantly use a size modifier even if that modifier is noisy. If borne out, these predictions should surface in an interaction between sufficient property and typicality gain. Visual inspection of the empirical proportions of redundant adjective use in Figure 13 suggests that this pattern is indeed borne out.

In order to investigate the effect of typicality gain on redundant adjective use, we conducted a mixed effects logistic regression analysis predicting redundant over minimal adjective use from fixed effects of scene variation, sufficient dimension, the interaction of scene variation and sufficient property, and the interaction of typicality gain and sufficient property. This is the same model as reported in Section 3.1.3, with the only difference that the interaction between sufficient property and typicality gain was added. All predictors were centered before entering the analysis. The model contained the maximal random effects structure that allowed it to converge: by-participant and by-item (where item was a color-object combination) random intercepts.

The model summary is shown in Table 5. We replicate the effects of sufficient property and scene variation observed earlier on this smaller dataset. Crucially, we observe a significant interaction between sufficient property and typicality gain.[10] Simple effects analysis reveals that this interaction

[10]Conducting the same analysis on the entire dataset (i.e., using all of the noisy typicality estimates, replicated the

Table 5: Model coefficients, standard errors, and p-values. Significant p-values are bolded.

	Coef β	SE(β)	p
Intercept	−1.85	0.34	**<.0001**
Scene variation	4.29	1.16	**<.001**
Sufficient property	2.72	0.60	**<.0001**
Scene variation : Sufficient property	0.88	2.12	<0.68
Sufficient property : Typicality gain	9.43	2.68	**<.001**

is due to a positive effect of typicality gain on redundant adjective use in the size-sufficient condition ($\beta = 4.47$, $SE = 1.65$, $p < .007$) but a negative effect of typicality gain on redundant adjective use in the color-sufficient condition ($\beta = -5.77$, $SE = 2.49$, $p < .03$).

An important point is of note: the typicality elicitation procedure we employed here is somewhat different from that employed by Westerbeek et al. (2015), who asked their participants "How typical is this color for this object?" We did this for conceptual reasons: the values that go into the semantics of the RSA model are most easily conceptualized as the typicality of an object as an instance of an utterance. While the typicality of a feature for an object type no doubt plays into how good of an instance of the utterance the object is, deriving our typicalities from the statistical properties of the subjective distributions of features over objects is beyond the scope of this paper. However, in a separate experiment we did ask participants the Westerbeek question. The correlation between mean typicality ratings from the Westerbeek version and the unmodified "How typical is this for X" version was .75. The correlation between the Westerbeek version and the modified version was .64. The correlation between the Westerbeek version and typicality gain was -.52.

For comparison, including typicality means obtained via the Westerbeek question as a predictor instead of typicality gain on the four high-powered items replicated the significant interaction between typicality and sufficient property ($\beta = -6.77$, $SE = 1.88$, $p < .0003$). Simple effects analysis revealed that the interaction is again due to a difference in slope in the two sufficient property conditions: in the size-sufficient condition, color is less likely to be mentioned with increasing color typicality ($\beta = -3.66$, $SE = 1.18$, $p < .002$), whereas in the color-sufficient condition, size is more likely to be mentioned with increasing color typicality ($\beta = 3.09$, $SE = 1.45$, $p < .04$).[11]

We thus overall find moderate evidence for typicality effects in our dataset. Typicality effects are strong for those items that clearly display typicality differences between the modified and unmodified utterance, but much weaker for the remaining items. That the evidence for typicality effects is relatively scarce is no surprise: the stimuli were specifically designed to minimize effects of typicality. However, the fact that both ways of quantifying typicality predicted redundant adjective use in the expected direction suggests that with more power or with stimuli that exhibit greater typicality variation, these effects may show up more clearly.

In the next section we evaluate whether the fit of non-deterministic RSA to the data is improved by using the empirically elicited typicalities as the values in the non-deterministic semantics.

scene variation and sufficient property effects. The interaction of typicality gain and sufficient property went in the same direction numerically, but failed to reach significance ($\beta = 1.52$, $SE = 1.45$, $p < .29$).

[11] Again, conducting this analysis on the entire dataset yielded only a marginal interaction of sufficient property and color typicality in the right direction ($\beta = -1.10$, $SE = .64$, $p < .09$).

4.2 Model evaluation: color typicality

[jd: insert actual numbers below! figure out what's going wrong in the bda that the returned color fidelities are lower than the returned size fidelities]

In order to evaluate the effect of utterance-object level typicality we proceed in two steps: first, we present the results of performing Bayesian data analysis in the same way as reported in Section 3.2, with the only difference that instead of fixed utterance type level fidelities we include the more fine-grained fidelity values corresponding to the typicality norms. We will focus more closely on the four items shown in Figure 13 in order to demonstrate the effect of including typicality in the model once you know, actually put a summary sentence here instead of "in order to see how the model bla". In a second step we then address the pressing question of whether more fine-grained typicalities add any predictive value.

4.2.1 Model evaluation: empirical typicalities

In order to generate model predictions and infer likely parameter values for the dataset reported above, we repeated the same Bayesian data analysis procedure as described above, with one difference: instead of using the utterance-type level fidelities we fed the model the empirically elicited typicality norms. This was slightly less trivial than it sounds for two reasons. First, we only elicited typicality norms for object-utterance pairs for which the utterance was either just the simple object category noun (e.g., *chair*) or the color-modified noun (e.g., *green chair*); that is, we did not elicit size typicality norms (for reasons described in the previous section). Second, we did not elicit typicality norms for utterance-object pairs where the object would not be in the deterministic semantics' extension (e.g., *green chair* used to refer to a red chair). In order to fill in the typicality gaps, so to speak, we assigned fidelity values to utterances as follows: object-utterance pairs where the object is in the extension of the utterance were assigned the empirically elicited typicality for that pair (e.g., *chair* typicality for any chair, *red chair* typicality for red chairs, see Table 6). If the object was not in the utterance extension, it received a fidelity of $1 - f_c$, where f_c is an utterance type level fidelity parameter for color (as in the basic non-deterministic model reported above). For objects in the utterance extension where the utterance additionally contained a size modifier, the empirical fidelity was multiplied by an utterance type level size fidelity f_s. See Table 6 for an overview of fidelity values for one item (red/green chair).

One final point of note: empirically elicited typicality values were rescaled to range from 0.5 to 1 (instead of from 0 to 1). This was done because a fidelity value of .5, when there are only two potential feature values (e.g., two colors in the scene, red and green), leads to a random choice between green and red items; that is, the modifier contains no information. When the fidelity value is below .5 in these two-feature value scenarios, the modifier contextually acquires the meaning of the other feature dimension, e.g., *green* is more likely to pick out red than green objects. The typicality values we elicited were not degree of membership values (which is what the model expects). Rather, they are more comparable to distance from prototype values (see ?, ?, for a discussion of the difference). By rescaling the empirical typicality values to fall above .5, we guaranteed that the utterance would have at least an above chance of meaning what it would mean under a deterministic semantics. Table 7 exemplifies the effect of rescaling the raw typicality values for the red/green chair item.

Table 6: Fidelity values (rescaled) for one example item (chair) that occurred in two colors (red, green). Rows indicate different utterances, columns indicate different objects. See Table 7 for the raw and rescaled empirical typicality values for the red/green chair item.

Utterance	small red chair	small green chair	big red chair	big green chair
chair	.83	.67	.83	.67
red chair	.84	$1 - f_c$	.84	$1 - f_c$
green chair	$1 - f_c$	.85	$1 - f_c$	.85
small chair	$f_s \cdot .83$	$f_s \cdot .67$	$(1 - f_s) \cdot .83$	$(1 - f_s) \cdot .67$
big chair	$(1 - f_s) \cdot .83$	$(1 - f_s) \cdot .67$	$f_s \cdot .83$	$f_s \cdot .67$
small red chair	$f_s \cdot .84$	$f_s \cdot (1 - f_c)$	$(1 - f_s) \cdot .84$	$(1 - f_s) \cdot (1 - f_c)$
big red chair	$(1 - f_s) \cdot .84$	$(1 - f_s) \cdot (1 - f_c)$	$f_s \cdot .84$	$f_s \cdot (1 - f_c)$
small green chair	$f_s \cdot (1 - f_c)$	$f_s \cdot .85$	$(1 - f_s) \cdot (1 - f_c)$	$(1 - f_s) \cdot .85$
big green chair	$(1 - f_s) \cdot (1 - f_c)$	$.(1 - f_s) \cdot .85$	$f_s \cdot (1 - f_c)$	$f_s \cdot .85$

Table 7: Raw and rescaled typicalities for the red and green chair items.

	Raw		Rescaled	
Utterance	red chair	green chair	red chair	green chair
chair	.68	.41	.83	.67
red chair	.69	NA	.84	NA
green chair	NA	.70	NA	.85

Figure 14: Posterior distribution over model parameters. Maximum a posteriori (MAP) $f_s = 0.79$, 95% highest density interval (HDI) = [0.76,0.80]; MAP $f_c = 0.88$, HDI = [0.86,0.91]; MAP $c_{size} = .08$, HDI = [0, 1.5]; MAP $c_{color} = 0.07$, HDI = [0,0.9]; MAP $\beta_c = 0.04$, HDI = [0,1.6]; MAP $\lambda = 34.0$, HDI = [30.8,36.5]

Figure 15: Posterior distribution over model parameters. Maximum a posteriori (MAP) $f_s = 0.79$, 95% highest density interval (HDI) = [0.76,0.80]; MAP $f_c = 0.88$, HDI = [0.86,0.91]; MAP $c_{size} = .08$, HDI = [0, 1.5]; MAP $c_{color} = 0.07$, HDI = [0,0.9]; MAP $\beta_c = 0.04$, HDI = [0,1.6]; MAP $\lambda = 34.0$, HDI = [30.8,36.5]

Results Including typicality yielded similar item-wise model-data correlations as the basic model XXX. Posteriors over parameters are shown in Figure **??**. Discuss similarities/diffs to basic model.

Posterior predictives for the cases with greatest typicality gain – *chair*, *golfball*, *weddingcake*, and *stapler* – are shown in Figure **??** alongside the posterior predictives from the basic model (with utterance type level fidelities). In the basic non-deterministic model, probability of redundant utterances is similar for items of different colors. In the model that includes empirically elicited typicalities, the probability of redundant utterances is greater where typicality gain is greater; that is, for cases where the unmodified utterance has low typicality and the modified utterance high typicality, analogous to the *blue banana* case.

4.2.2 Model evaluation: interpolation analysis

Because the correlations between model-predicted utterance probability posterior predictives and empirical proportions are very similar across the basic and empirical typicality model, the question arises whether utterance-level typicalities add any predictive value whatsoever. In order to address this question we present a second BDA analysis in which we introduce an additional parameter in the model that functions as a weight on fidelity type: if the weight is 0, only the utterance-type level fidelities are used; if the weight is 1 only the empirical typicalities are used. Therefore, if the BDA returns posterior values for fidelity type weight greater than 0, empirical typicalities are justified.

Results Posteriors over parameters are shown in Figure **??**. insert figure The MAP estimate for fidelity type weight is XXX (HDI = [X,X]), suggesting that utterance-level typicality adds predictive value.

4.3 Discussion

main points: a) that non-deterministic RSA with utterance-level typicalities as fidelity values captures the color typicality effects reported in the literature qualitatively (yellow/blue banana); b) that even in our dataset, where items were designed to not exhibit great typicality effects, we find evidence of typicality effects on utterance probability, replicating previous studies; and c) BDA shows that including empirically elicited typicality norms in the model adds predictive value. This suggests that speakers are tracking typicality at a very fine-grained level.

[jd: This is all well and good, but to what extent is non-deterministic RSA just a model of modifier choice in modified referring expressions? Put differently, does non-deterministic RSA provide a good account of content selection in referring expressions more generally? To answer this question we move beyond modified referring expressions and turn to simple nominal referring expressions.]

In the next section we turn to extending non-deterministic RSA beyond the choice of modifier.

5 Evaluating non-deterministic RSA for nominal choice

In this section we investigate whether non-deterministic RSA can account for referring expression production beyond the choice of modifier. To do so, we begin by presenting a second production experiment. This experiment investigates speakers' choice of level of reference in nominal referring expression (*dalmation*, *dog*, or *animal*). As discussed in Section 5, multiple factors have been shown to play a role in the choice of nominal referring expression, including an expression's contextual informativeness, its cognitive cost (short and frequent terms are preferred over long and infrequent ones) cite cite, and its typicality (an utterance is more likely to be used if the object is a good example of it) is that true? cite. yes, caroline put ref in cogsci talk. We then evaluate non-deterministic RSA on the nominal choice dataset by conducting the same type of Bayesian data analysis as reported in the previous section.

5.1 Experiment 2: level of reference in nominal referring expressions

Exp 2 employed the same procedure as Exp. 1, but each display consisted of three objects.[12] We manipulated the contextual informativeness of each level of reference – subordinate (*dalmatian*), basic (*dog*), and superordinate (*animal*) – by manipulating the distractor items.

5.1.1 Method

Participants We recruited 58 pairs of participants (116 participants total) over Amazon's Mechanical Turk who were each paid $1.75 for their participation.

Procedure and materials The procedure was identical to that of Exp. 1. Participants proceeded through 72 trials. Of these, half were critical trials of interest and half were filler trials (the critical trials from Exp. 1). On critical trials, we varied the level of reference that was sufficient to mention for uniquely establishing reference.

Stimuli were selected from nine distinct domains, each corresponding to distinct basic level categories such as *dog*. For each domain, we selected four subcategories to form our target set (e.g. *dalmatian*, *pug*, *German Shepherd* and *husky*). See Table 8 for a full list of domains and their associated target items. Each domain also contained an additional item which belonged to the same basic level category as the target (e.g., *greyhound*) and items which belonged to the same supercategory but not the same basic level (e.g., *elephant* or *squirrel*). The latter items were used as distractors.

Each trial consisted of a display of three images, one of which was designated as the target object. Each pair of participants saw each target exactly once, for a total of 36 trials per pair.

[12] Exp. 2 constitutes a replication of Graf, Degen, Hawkins, and Goodman (2016).

Table 8: List of domains and associated superordinate category, target stimuli, and mean length (standard deviation) in characters of actually produced subordinate level utterances in Exp. 2.

Domain	Super	Targets	Mean sub length (sd)
bear	animal	black bear	9.9 (.14)
		polar bear	8.8 (.35)
		panda bear	5.5 (.2)
		grizzly bear	9 (.98)
bird	animal	eagle	4.9 (.1)
		parrot	6.1 (.13)
		pigeon	5.9 (.22)
		hummingbird	10.1 (.5)
candy	snack	MnMs	4.4 (.49)
		skittles	6.9 (.43)
		gummy bears	8.5 (.47)
		jelly beans	9.3 (.44)
car	vehicle	SUV	3 (0)
		minivan	5.7 (.27)
		sports car	9.8 (.23)
		convertible	11.1 (.2)
dog	animal	pug	3 (.08)
		husky	4.7 (.22)
		dalmatian	8.8 (.18)
		German Shepherd	13.1 (.82)
fish	animal	catfish	6.6 (.4)
		goldfish	7.9 (.22)
		swordfish	8 (.43)
		clownfish	9.1 (.38)
flower	plant	rose	4 (0)
		tulip	4.4 (.18)
		daisy	5.9 (.55)
		sunflower	9 (.11)
shirt	clothing	T-shirt	6.4 (.48)
		polo shirt	6.7 (.79)
		dress shirt	11 (0)
		Hawaii shirt	12.6 (.46)
table	furniture	picnic table	9.7 (.58)
		dining table	12 (0)
		coffee table	9.1 (.95)
		bedside table	8.3 (.68)

These target items were randomly assigned distractor items which were selected from four different context conditions, corresponding to different communicative pressures (see Figure 2). We refer to these conditions with pairs of numerals specifying which levels of the taxonomy are present in the distractors: (a) item12 contexts contain one distractor of the same basic level and one distractor of the same superlevel (e.g., target: *dalmatian*, distractor 1: *greyhound* (also a dog), distractor 2: *squirrel* (also an animal)); (b) item22 contexts contain two distractors of the same superlevel but different basic level as the target (e.g., target: *husky*, distractors: *hamster* and *elephant*); (c) item23 contexts contain one distractor of the same superlevel and one unrelated item (e.g., target: *pug*, distractor 1: *cow*, distractor 2: *table*); and (d) item33 contexts contain two unrelated items (e.g., target: *German Shepherd*, distractors: *shirt* and *cookie*).

This context manipulation served as a manipulation of utterance informativeness: any target could be referred to at the sub (*dalmatian*), basic (*dog*) or super (*animal*) level. However, the level of reference necessary for uniquely referring differed across contexts: in item12 contexts, the sub level was necessary. In item22 and item23 contexts, the basic level was necessary (though the sub level was also possible). In item33 contexts all three utterances were possible.

5.1.2 Data pre-processing and exclusion

We collected 2187 referential expressions. To determine the level of reference for each trial, we followed the following procedure. First, 41 trials on which the listener selected the wrong referent were excluded, leading to the elimination of 1.9% of trials. Then, speakers' and listeners' messages were parsed automatically; the referential expression used by the speaker was extracted for each trial and checked for whether it contained the current target's correct sub, basic or super level term using a simple grep search. In this way, 72.1% of trials were labelled as mentioning a pre-coded level of reference. In the next step, remaining utterances were checked manually to determine whether they contained a correct level of reference term which was not detected by the grep search due to typos or grammatical modification of the expression. In this way, meaning-equivalent alternatives such as *doggie* for *dog*, or reduced forms such as *gummi*, *gummies* and *bears* for *gummy bears* were counted as containing the corresponding level of reference term. This covered another 15.1% of trials. A total of 12.8% of correct trials were excluded because the utterance consisted only of an attribute of the superclass (*the living thing* for *animal*), of the basic level (*can fly* for *bird*), of the subcategory (*barks* for *dog*) or of the particular instance (*the thing facing left*) rather than a category noun. These kinds of attributes were also mentioned in addition to the noun on trials which were included in the analysis for 8.9% of sub level terms, 19.1% of basic level terms, and 66.7% of super level terms. On 1.2% of trials two different levels of reference were mentioned; in this case the more specific level of reference was counted as being mentioned in this trial. After all exclusion and pre-processing, 1870 cases classified as one of *sub*, *basic*, or *super* entered into the analysis.

5.1.3 Results and discussion

Proportions of sub, basic, and super level utterances are shown in the top row of Figure 16. Overall, super level mentions are highly dispreferred ($< 2\%$), so we focus in this section only on predictors of sub over basic level mentions. The clearest pattern of note is that sub level mentions are only preferred in the most constrained context that necessitates the sub level mention for unique reference (item12, e.g. target: dalmatian, distractor: greyhound). Nevertheless, even in these contexts there

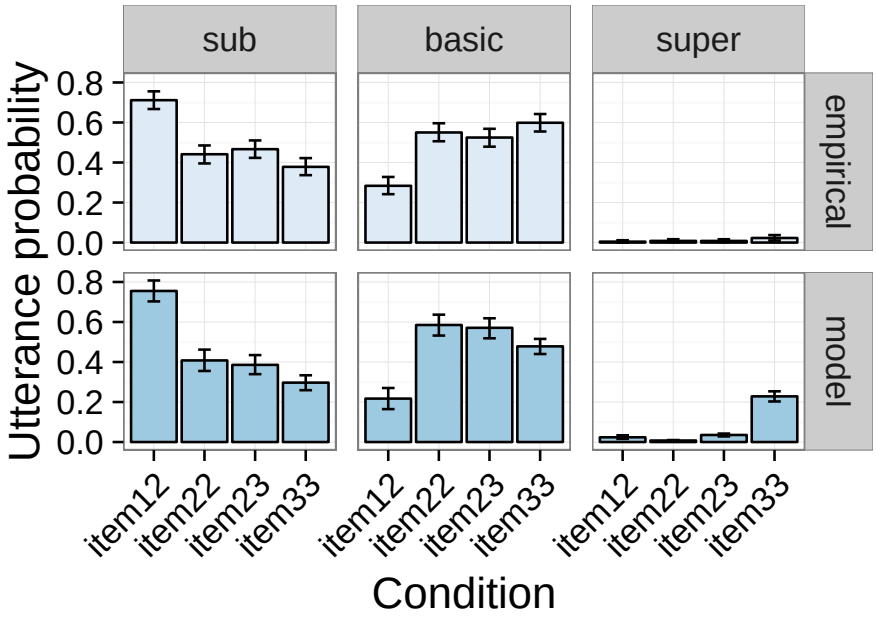


Figure 16: Utterance probabilities across different conditions. Columns indicate utterances, rows indicate data type (empirical proportion, MAP estimates of posterior predictives for full model with cost and non-deterministic semantics.

is a non-negligible proportion of basic level mentions (28%). In the remaining contexts, where the sub and basic level are equally informative, there is a clear preference for the basic level.

What explains these preferences? In order to test for effects of informativeness, length, frequency, and typicality on nominal choice we conducted a mixed effects logistic regression predicting sub over basic level mention from centered predictors for the factors of interest and the maximal random effects structure that allowed the model to converge (random by-speaker and by-target intercepts).

Frequency was coded as the difference between the sub and the basic level's log frequency, as extracted from the Google Books Ngram English corpus ranging from 1960 to 2008.

Length was coded as the ratio of the sub to the basic level's length. We used the mean empirical lengths in characters of the utterances participants produced. For example, the minivan, when referred to at the subcategory level, was sometimes called "minivan" and sometimes "van" leading to a mean empirical length of 5.71. This is the value that was used, rather than 7, the length of "minivan". That is, a higher frequency difference indicates a *lower* cost for the sub level term compared to the basic level, while a higher length ratio reflects a *higher* cost for the sub level term compared to the basic level.[13]

Typicality was coded as the ratio of the target's sub to basic level label typicality.[14] That is, the higher the ratio, the more typical the object was for the sub level label compared to the basic level; or in other words, a higher ratio indicates that the object was relatively atypical for the basic label compared to the sub label. For instance, the panda was relatively atypical for its basic level

[13]We replicate the well-documented negative correlation between length and log frequency ($r = -.49$ in our dataset).

[14]Typicalities were elicited in a separate norming study that was identical in procedure to that of Exp. 1a. See Appendix F for details about the study.

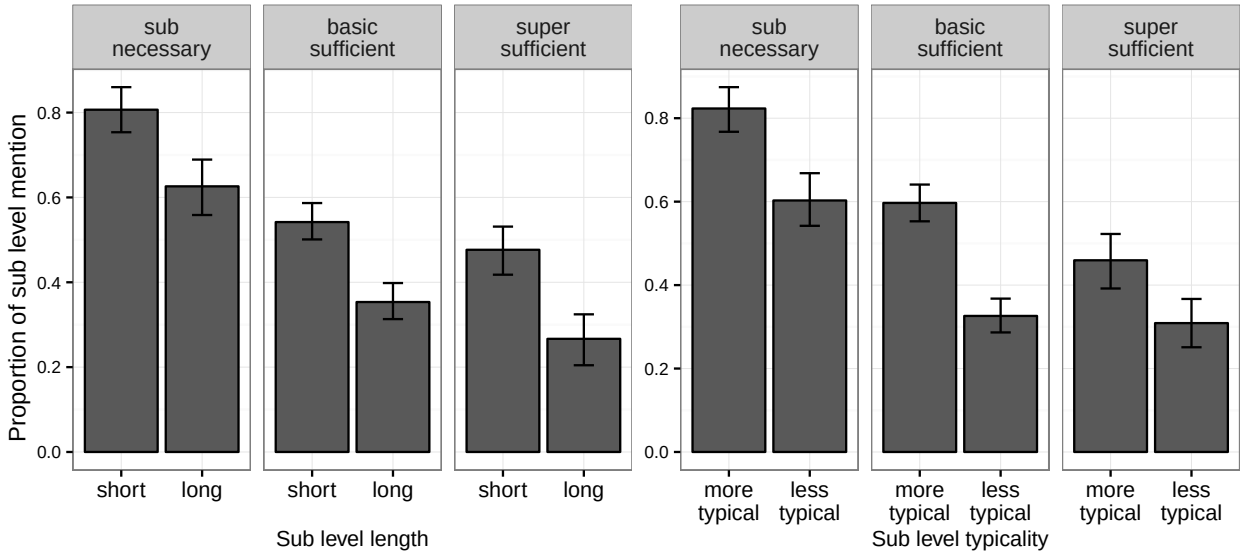


Figure 17: Proportion of sub level (over sub and basic level) terms across conditions. Left: when the sub length is relatively short (.67,1.82] or long [1.82,4.3) compared to the basic level term. Right: when the target object was relatively more [1.06,1.91) or less (.88,1.06] typical for the sub compared to the basic level term. Intervals were generated by splitting data into groups of roughly equal numbers of observations.

"bear" (mean rating 0.75) compared to the sub level term "panda bear" (mean rating 0.98), which resulted in a relatively *high* typicality ratio.

Informativeness condition was coded as a three-level factor: *sub necessary*, *basic sufficient*, and *super sufficient*, where item22 and item23 were collapsed into *basic sufficient*. Condition was Helmert-coded: two contrasts over the three condition levels were included in the model, comparing each level against the mean of the remaining levels (in order: *sub necessary*, *basic sufficient*, *super sufficient*). This allowed us to determine whether the probabilities of type mention for neighboring conditions were significantly different from each other, as suggested by Figure 16.

The log odds of mentioning the sub level term were greater in the *sub necessary* condition than in either of the other two conditions ($\beta = 2.05$, $SE = .17$, $p < .0001$), and greater in the *basic sufficient* condition than in the *super sufficient* condition ($\beta = .54$, $SE = .15$, $p < .001$), suggesting that the contextual informativeness of the sub level mention has a gradient effect on utterance choice.[15] There was also a main effect of typicality, such that the sub level term was preferred for objects that were more typical for the sub level compared to the basic level description ($\beta = .4.84$, $SE = 1.32$, $p < .001$, see Figure 17). In addition, there was a main effect of length, such that as the length of the sub level term increased compared to the basic level term ("chihuahua"/"dog" vs. "pug"/"dog"), the sub level term was dispreferred ("chihuahua" is dispreferred compared to "pug", $\beta = -.95$, $SE = .27$, $p < .001$, see Figure 17). The main effect of frequency did not reach significance ($\beta = .07$, $SE = .10$, $p < .51$).

Unsurprisingly, there was also significant by-participant and by-domain variation in sub level

[15]Importantly, model comparison between the reported model and one that subsumes basic and super under the same factor level revealed that the three-level condition variable is justified ($\chi^2(1) = 12.82$, $p < .0004$), suggesting that participants don't simply revert to the basic level unless contextually forced not to.

term mention. For instance, mentioning the sub over the basic level term was preferred more in some domains (e.g. in the "candy" domain) than in others. Likewise, some domains had a greater preference for basic level terms (e.g. the "shirt" domain). Using the super term also ranged from hardly being observable (e.g. the "flower" domain) to being used more frequently (e.g. in the "table" and "car" domain).

We thus replicate the well-documented preference to refer to objects at the basic level, which is partly modulated by contextual informativeness and partly a result of the basic level term's cognitive cost and typicality compared to its sub level competitor.

Perhaps surprisingly given the previous literature, we did not observe an effect of frequency on sub level term mention. This may have a number of reasons. For instance, the modality of the experiment may have mattered here: the current study was a written production study, while most studies that have identified frequency as a factor governing production choices are spoken production studies (cite cite). It may be that the cognitive cost of typing longer words may be disproportionately higher than that of producing longer words in speech, thus obscuring a potential effect of frequency.

5.2 Non-deterministic RSA for nominal choice

Here we show that non-deterministic RSA as presented in Section 2.2 can be straightforwardly extended to modeling the choice of taxonomic level of reference. We include three modifications, while leaving the general framework as is. The first modification concerns the utterance alternatives. The second concerns the elicited typicality values and the resulting fidelity values. The third concerns the cost function. We briefly elaborate on each in turn.

Utterance alternatives. Whereas the modifier choice model treats all individual features and feature combinations represented in the display as utterance alternatives, the nominal choice model considers only the three different levels of reference to the target as alternatives, e.g., *dalmatian*, *dog*, *animal*. That is, assuming a German Shepherd as a distractor, *German Shepherd* is *not* considered an alternative. This has consequences for the assumed fidelity values, which we turn to next. [jd: we should probably discuss this in the GD? ie, if we also assumed distractor labels as alternatives, we would have to do the rescaling – would results be different? or the othe rway round: if we assume in modifier choice only the target's features are available as alternatives, would results be different?]

Fidelity values. Just as we did for capturing color typicality effects in Section 4.2, we elicited empirical typicality values for object-utterance combinations.[16] For each display, we know the typicality of each object in the display as an instance of the three potential target utterances (capturing, for instance, that the word "dog" describes a dalmatian better than a grizzly bear, but it also describes a grizzly bear better than a tennis ball). This allows us to use the typicality values as fidelity values directly, without rescaling as was necessary in the modifier choice model.

Cost function. Recall the pragmatic speaker's utility function from Section 2.2, where the weighted informativeness term $\lambda \ln P_{L_0}(o|u)$ traded off against the weighted utterance cost $\beta_c c(u)$. In the modifier model we assumed a constant cost for each added modifier. Because all utterance

[16] See Appendix F for details of typicality elicitation experiment.

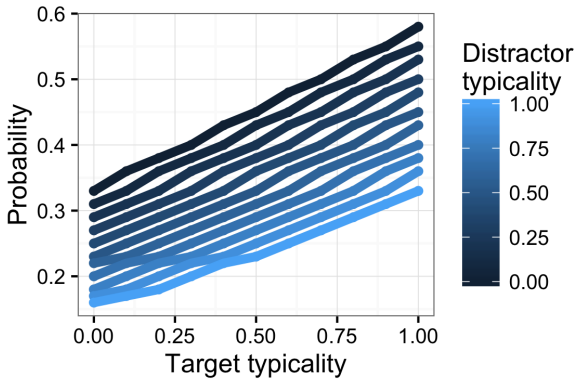


Figure 18: Literal listener probability of choosing the target under different typicalities of the target (x-axis) or the distractors (color) for the observed utterance. For simplicity we assume equal typicality of both distractors. The remaining probability mass for each case is thus uniformly distributed over both distractors.

alternatives in the nominal choice model have word length 1, we update the cost function to be composed of each utterance's length $\hat{c}_l$ and frequency $\hat{c}_f$ (as described in the previous section), weighted by free parameters β_f and β_l :

$$P_{S_1}(u|o) \propto e^{\lambda \ln P_{L_0}(o|u) + \beta_f \hat{c}_f + \beta_l \hat{c}_l} \tag{3}$$

To understand the qualitative behavior of the model, we briefly delve into two aspects of the model: first, the effect of typicality on the literal listener (and, in consequence via the pressure to be informative) the speaker. And second, the effect of cost (utterance length and frequency) on the speaker.

5.2.1 Typicality effects

Literal listener behavior. The literal listener's probability of choosing the target under different typicalities for the observed utterance are shown in Figure 18. In general: as the target's typicality as an instance of the utterance increases and the distractors' typicality decreases, the probability of the literal listener choosing the target increases. Subordinate level terms tend to fall in the upper right quadrant of this graph. Basic level terms in the *sub necessary* conditions tend to fall in the lower right quadrant, while basic level terms in the *basic sufficient* conditions tend to fall in the upper right quadrant as well.

Pragmatic speaker behavior. To understand the effect of typicality on the speaker's behavior it is useful to think about the problem of deciding which taxonomic level to refer at in terms of typicality gain, as we did in Section 4 for the choice between modified and unmodified expression. There, we found that relatively large target (compared to distractor) typicality gains in going from unmodified to modified expressions compared resulted in greater probability of overmodification. Here we observe the same effect in going from a higher (less specific) to a lower (more specific) taxonomic level. This can be seen in Figure 19, which shows the probability of each utterance (sub, basic, or super) as a function of absolute target typicality as well as target typicality gain. Target

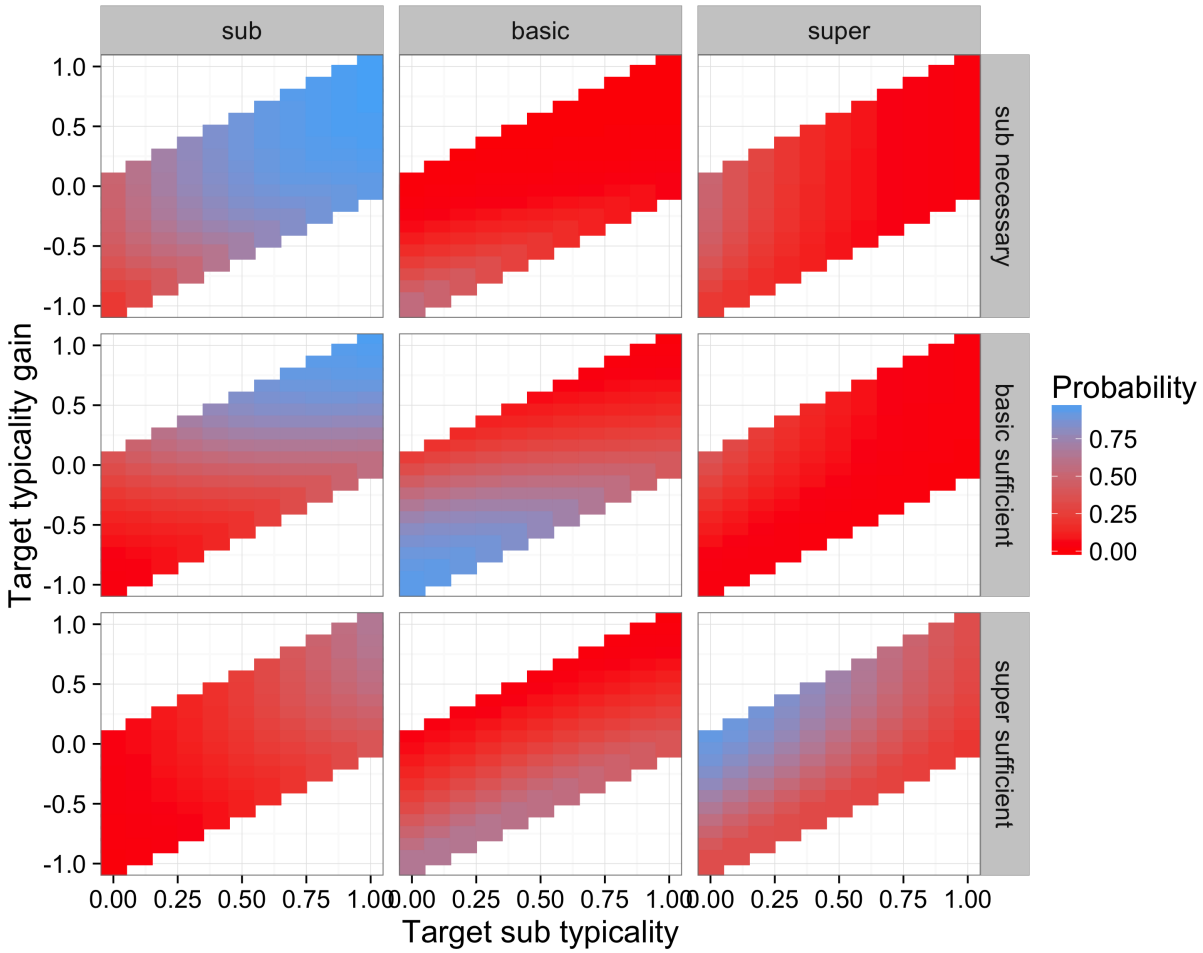


Figure 19: Pragmatic speaker probability of choosing each utterance (sub, basic, super) under varying absolute target sub typicalities (x-axis) and target typicality gains (y-axis), assuming equal typicality values for both distractors. Rows indicate different simulated conditions.

typicality gain is the difference between the target's sub level typicality and the target's basic level typicality. Probabilities are shown for contexts with three items, always assuming $\lambda = 7$, but manipulating distractor typicality to simulate conditions analogous to our experimental conditions *sub necessary*, *basic sufficient*, and *super sufficient*. Simulated distractor typicalities for sub, basic, and super level reference are shown in Table 9.

The blue areas in the graph indicate highest-probability regions. For example, as expected in the *sub necessary* condition, the sub level term is the most likely one. However, in certain cases the basic level term also receives non-zero probability, notably when the target is a better instance of the basic than the sub level term, or (not pictured) when the typicality of the distractor as an instance of the basic level term is very low (e.g., the typicality of the koala bear as an instance of "bear" was only 0.50). Indeed, the grizzly (with high typicality for basic level "bear", .97) is referred to as "bear" rather than "grizzly bear" in 85% of *sub necessary* conditions when the koala is the distractor.

Table 9: Simulated distractor typicality (fidelity) values for sub, basic, and super level utterances in simulated conditions. In contrast to the actual experimental conditions, we assume equal typicality values for both distractors.

		Condition		
		sub necessary	basic sufficient	super sufficient
Utterance	sub	0	0	0
	basic	.8	.1	0
	super	.8	.8	0

In the *basic sufficient* conditions, sub level reference is nevertheless strongly predicted when target sub typicality gain is positive (i.e., when the target is a much better instance of the sub than of the basic level term). An example of such a case is the panda bear, who received a sub level typicality of .98 and a basic level typicality of only .75. Indeed, even when basic level reference was sufficient, the panda was referred to as the "panda" 81% of the time.

These patterns mirror the typicality effects obtained via the mixed effects regression.

5.2.2 Cost effects

The additional effect of cost on nominal choice is straightforward: the costlier an utterance (relative to its alternatives), the less likely it is to be used. This pattern, too, is one observed in the mixed effects regression. For instance, the (short, less costly) pug is almost three times as likely as the (long, more costly) German Shepherd to be referred to by its subordinate level term in the *basic sufficient* and *super sufficient* conditions, where subordinate level reference is unnecessary.

In Section 5.2 we showed that non-deterministic RSA captures the right kinds of qualitative effects as observed in the mixed effects regression. In the next section we evaluate how well the model captures nominal choice preferences quantitatively.

5.3 Model evaluation: nominal choice

In order to evaluate non-deterministic RSA for nominal choice, we repeated the same Bayesian data analysis as reported in Section 3.2 and Section 4.2 to generate model predictions and infer likely parameter values. We did so by conditioning on the observed production data (coded into *sub*, *basic*, and *super* level mentions as described above) and integrating over the three free parameters $\lambda \sim \mathcal{U}(0, 20)$, $\beta_f \sim \mathcal{U}(0, 5)$, $\beta_l \sim \mathcal{U}(0, 5)$.

Point-wise maximum a posteriori (MAP) estimates of the model's posterior predictives for each combination of utterance and informativeness condition (collapsing across different items) are compared to empirical data in Figure 16. The model clearly captures the preference towards sub level mentions in the *sub necessary* conditions and the basic level preference in all other conditions. It also captures the further decrease in sub level mentions in the *super sufficient condition*. However, it does overpredict super level mentions, though not as badly as models that either assume a deterministic semantics or that ignore utterance cost.[17] At this level, the model achieves a correlation of

[17]The reader is referred to Appendix G for a comparison of the models containing a) only informativeness with deterministic semantics; b) only informativeness with non-deterministic semantics; c) informativeness with deterministic

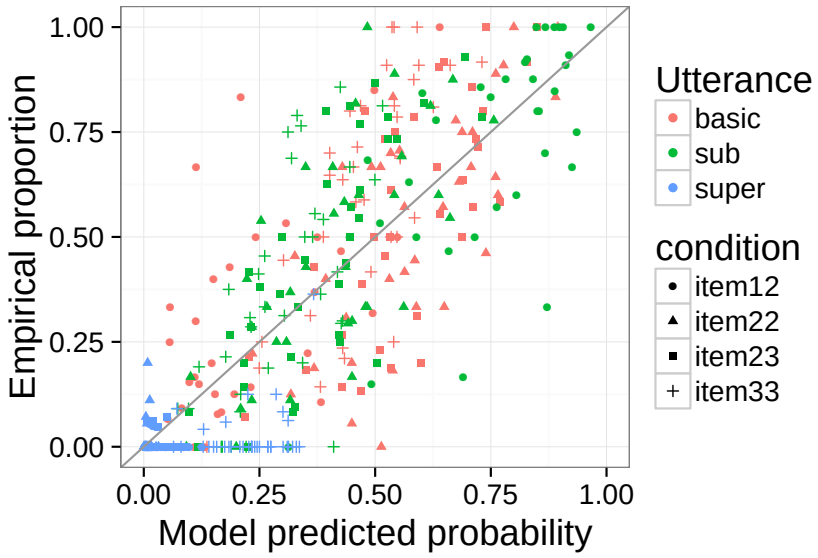


Figure 20: Scatterplot of by-target empirical utterance proportions against model posterior predictive MAP estimates. Gray line indicates perfect correlation line.

$r = .94$. Computing correlations additionally on the by-target level yields a correlation of $r = .84$ (see also the scatterplot in Figure 20).

Parameter posteriors are shown in Figure 21. Both informativeness and length receive significant weight. In contrast, the effect of frequency appears to be much weaker with a MAP of .1 and the HDIs overlapping with 0. This mirrors the null effect of frequency found in the regression analysis. However, a large number of cases also received a non-zero frequency weight.

In order to ascertain whether typicality as incorporated in the non-deterministic semantics was indeed contributing to the explanatory power of the model, we ran an additional Bayesian data analysis with an added typicality weight parameter $\beta_t \in [0, 1]$. This parameter interpolated between empirical typicality values (when $\beta_t = 1$) and deterministic (i.e., 0 or 1) a priori values based on the true taxonomy (when $\beta_t = 0$). We found a MAP estimate for β_t of .95, HDI = [0.82,.99], strongly indicating that it is useful to incorporate empirical typicality values and thus providing further support for the value of non-deterministic truth functions in modeling referential expressions.

6 General Discussion

6.1 Summary

How do speakers choose a referring expression? Here we have shown that they do so by trading off various factors: the contextual informativeness of the referring expression on the one hand, and the cognitive cost of the expression on the other. Importantly, computing contextual informativeness with respect to a *non-deterministic* underlying semantics was crucial for capturing various aspects of speakers' referring behavior. First, the non-deterministic semantics allowed us to capture the basic well-documented asymmetry for speakers to be more likely to redundantly use color adjectives

semantics and cost; d) informativeness with non-deterministic semantics and cost (the current model).

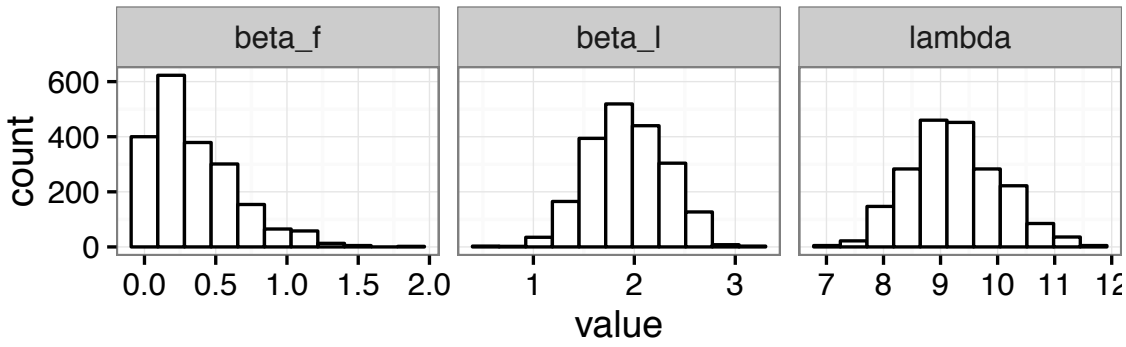


Figure 21: Posterior distribution over model parameters. Maximum a posteriori (MAP) $\beta_f = 0.10$, 95% highest density interval (HDI) = [0.002,0.95]; MAP $\beta_l = 1.85$, HDI = [1.23,2.65]; MAP $\lambda = 9.19$, HDI = [7.72,10.80].

rather than size adjectives. In addition, it predicted an interaction between sufficient dimension and scene variation on the probability of redundancy, which was very clearly borne out in the data: increased scene variation resulted in a much greater increase in redundant color than in redundant size adjective use. Finally, the non-determinism in the semantics gave rise to well-documented effects of typicality in both modifier choice and noun choice. A modifier was more likely to be mentioned redundantly when the object was a substantially less good instance of the unmodified than of the modified expression. Analogously, a noun at a taxonomically lower level than necessary for establishing reference was more likely to be mentioned when the object was a substantially less good instance of the higher than of the lower level.

We have thus shown that with one key innovation – a non-deterministic semantics – one can retain the assumption that speakers rationally trade off informativeness and cost of utterances in language production. Rather than being wastefully overinformative, adding redundant modifiers or referring at a lower taxonomic level than strictly necessary *is* in fact informative when the prima facie sufficiently informative expression is substantially noisier than its redundant/overly specific counterpart. This innovation thus not only provides a unified explanation for a number of key patterns within the overinformative referring expression literature that have thus far eluded a unified explanation; it also extends to the domain of nominal choice.

In the following we discuss a number of intriguing questions this work raises and avenues for future research that it suggests.

6.2 'Overinformativeness'

This work challenges the traditional notion of overinformativeness in the linguistic and psychological literature (Engelhardt, Bailey, & Ferreira, 2006b; ?, ?). The reason that redundant referring expressions became interesting for psycholinguists to study is because they seem to constitute a clear violation of rational theories of language production. For example, Grice's Quantity-2 maxim, which asks of speakers to "not make [their] contribution more informative than is required" (Grice, 1975), appears violated by any redundant referring expression – if size is the only feature that distinguishes the target object from the rest, the mention of color seems more informative than required.

This conception of (over-)informativeness assumes that all modifiers are born equal – i.e., that

there are no a priori differences in the utility of mentioning different properties of an object. Under this conception of modifiers, there are hard lines between modifiers that are and aren't informative in a context. However, what we have shown here is that under a non-deterministic semantics, a modifier that would be regarded as overinformative under the traditional conception may nevertheless add some information about the referent. In particular, the more visual variation there is in the scene and the less noisy the redundant modifier is compared to the modifier that selects the dimension that uniquely singles out the target, the more information it adds about the referent, and the more likely it therefore is to be mentioned. This work thus challenges the traditional notion of utterance overinformativeness by providing an alternative that nicely captures the quantitative variation observed in speakers' production in a principled way while still assuming that speakers are aiming to be informative.

What, then, would count as an overinformative utterance under non-deterministic RSA? The answer is simple: the less expected the use of a redundant modifier is (given knowledge of utterance noise and scene variation), the more the use of that modifier will be considered overinformative.

6.3 Comprehension

While the account proposed in this paper is not directly concerned with predicting listeners' behavior in interpreting referring expressions, it can be extended to do so relatively straightforwardly. RSA models typically assume that listeners, in interpreting utterances, are doing so by reasoning about their model of the speaker. In this paper we have provided precisely such a model of the speaker. In what way should the predicted speaker probabilities enter into comprehension? Here we can make a direct connection to surprisal theory in sentence processing (?, ?), where it has been shown that the effort involved in processing a sentence is a function of how surprising that sentence is under the listener's language model. While in these studies surprisal is usually estimated from syntactically parsed corpora, here we are providing a speaker model from which we can derive estimates of *pragmatic surprisal*. Generally, the more likely a redundant utterance is, the easier it should be to process in context. We have shown that redundant expressions are more likely than minimal expressions when the distinguishing dimension is relatively noisy and scene variation is relatively high. In situations like these, one would thus expect the redundant expression to be easier to process than in cases where the redundant expression is relatively less likely.

Is there evidence that listeners do behave in accordance with this prediction? While we have not run processing studies ourselves, we can look into the literature. Indeed, there is evidence that in situations where the redundant modifier does provide some information about the referent, listeners are faster to respond and select the intended referent when they observe a redundant referring expression than when they observe a minimal one (Arts et al., 2011; Paraboni et al., 2007). However, there is also evidence that redundancy sometimes incurs a processing cost: both Engelhardt, Demiral, and Ferreira (2011) and Davies and Katsos (2013) (Exp. 2) found that listeners were slower to identify the target referent in response to redundant compared to minimal utterances. It is useful to examine the stimuli they used. In the Engelhardt et al study, there was only one distractor that varied in type, i.e., type was sufficient for establishing reference. This distractor varied either in size or in color. Thus, scene variation was very low and overinformative expressions therefore likely surprising. Interestingly, the incurred cost was greater for redundant size than for redundant color modifiers, in line with the RSA predictions that color should be generally more likely to be used redundantly than size. In the Davies et al study, the 'overinformative' conditions contained displays of four objects which differed in type. Stimuli were selected via a production

Table 10: Fidelity across models alongside the effects from Table 2 that each model captures.

Exp.	Model	Fidelity level	How obtained	Effect(s)
1	basic non-deterministic	modifier type	inferred	color/size asymmetry & scene variation
1	typicality (modified/unmodified)	utterance-object	elicited (nominal, color) and inferred (size)	color/size asymmetry, scene variation, & color typicality
2	typicality (level of reference)	utterance-object	elicited	basic level preference & subordinate level mention

pre-test: only those objects that in isolation were not referred to with a modifier were selected for the study. That is, stimuli were selected precisely on the basis that redundant modifier use would be unlikely.

While the online processing of redundant referring expressions is yet to be systematically explored under the non-deterministic RSA account, this cursory overview of the patterns reported in the existing literature suggests that pragmatic surprisal (i.e., negative log-transformed speaker probabilities) may be a plausible linking function from model predictions to processing times.

6.4 Fidelity

The model crucially relies on a non-deterministic semantics to capture the effects we have reported in this paper. But what is the nature of this non-determinism? What does it represent? For the purpose of Exp. 1 (modifier choise), fidelity initially applied at the modifier *type* level. The semantics of modifiers was underlyingly truth-conditional and the fidelity term captured the probability that a modifier's truth conditions would accidentally be inverted. This model included only two fidelity terms, one for size and one for color. We then extended the notion of fidelity to apply at the level of utterance-object combinations (e.g., *golf ball* vs. *pink golf ball* as applied to a pink golf ball) to account for color typicality effects. In this instantiation of the model, fidelity differed for every utterance-object combination and captured how good of an instance of an utterance an object was. Similarly, in Exp. 2 (nominal choice) fidelity differed for every utterance-object combination (e.g., *dog* vs. *dalmatian* as applied to a dalmatian). This is summarized in Table 10.

What we have said nothing about thus far is where these numbers come from; in particular, which aspects of our experience – linguistic, perceptual, conceptual, communicative – they represent. We will offer some speculative remarks and directions for future research here.

First, it is possible that the numbers represent the difficulty associated with verifying whether the property denoted by the utterance holds of the object. This difficulty may be perceptual – for example, it may be relatively easier to visually determine of an object whether it is red than whether it it is big. Similarly, at the object-utterance level, it may be easier to determine of a yellow banana than of a blue banana whether it exhibits banana-hood, in consequence yielding a lower typicality value for a blue banana than for a yellow banana as an instance of *banana*. It may also be conceptual – for example, it may be easier to determine whether a box belongs to John than whether XXX.

Another possibility is that the numbers represent aspects of agents' prior beliefs (world knowledge) about the correlations between features of objects. For example, conditioning on bananahood

holding of objects and asking for the relative probabilities of various colors obtaining in that set will yield a high number for yellow and a low one for blue.[18]

Another hypothesis is that the numbers capture the past probability of communicative success in using a particular utterance (e.g., *banana*) to refer to an object with a particular set of features (e.g., blue bananas vs. yellow bananas). However, this probability is likely itself not independent of the first two possibilities discussed.

Finally, it is also possible that the numbers are simply an irreducible part of the lexical entry of each utterance-object pair. This seems unlikely, given that this would require a separate number for each utterance and object token. It also suggests that the numbers should not be updated in response to further exposure of objects. For example, if the numbers were a fixed component of the lexical entry *banana*, then even being exposed to a large number of blue bananas should not change the value. This seems unlikely but deserves to be investigated further.

6.5 Audience design

One question which has plagued the literature on language production is that of whether, and to what degree, speakers actually tailor their utterances to their audience (Clark & Murphy, 1982; Horton & Keysar, 1996; Brown-Schmidt & Heller, 2014). This is also known as the question of *audience design*. With regards to redundant referring expressions, the question is whether speakers produce redundant expressions because they can't help it (i.e., due to internal production pressures) or specifically because it is helpful for their interlocutor (i.e., due to considerations of audience design).

Non-deterministic RSA seems to make a claim about this issue (hm, I'm not sure about this): the non-determinism is located in the literal listener component, with respect to which speakers are trying to be informative. That is, it would seem that speakers produce referring expressions that are tailored to their listeners. However, this is misleading. The ontological status of the literal listener is as a "dummy component" that allows the pragmatic recursion to get off the ground. "Actual" pragmatic listeners are, in line with previous work, more likely fall into the class of L_1 listeners; listeners who reason about the speaker's intended meaning via Bayesian inference (M. C. Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013). Thus, the RSA model as formulated here remains agnostic about whether the speaker's (over)informativeness should be considered as geared towards listeners or simply a production-internal process.

6.6 Other factors affecting redundancy

Non-deterministic RSA as presented in this paper straightforwardly accounts for effects of typicality, cost, and scene variation on redundancy in referring expressions. However, other factors have been identified as contributing to redundancy. For example, Rubio-Fernandez (2016) has shown that colors are mentioned more often redundantly for clothes than for geometrical shapes. Her explanation: knowing an object's color is generally more useful for clothing than it is for shapes. While she doesn't provide a detailed explanation for why this is the case, it is plausible that agents' knowledge of *goals* may be relevant here. For example, knowing the color of clothing is relevant to the goal of deciding what to wear or buy. In contrast, knowing the color of geometrical shapes is rarely relevant to any everyday goal agents might have. While the RSA model as implemented

[18] Though these probabilities cannot directly match up with the elicited typicality values, given that probabilities will have to sum up to 1, while typicality values were not normalized.

here does not accommodate an agent's goals, it can be extended to do so via projection functions, as has been done for capturing figurative language use (e.g., Kao, Wu, Bergen, & Goodman, 2014) or question-answer behavior (Hawkins, Stuhlmüller, Degen, & Goodman, 2015). This should be explored further in future research.

[jd: a note on incrementality? eg, pechmann says incrementality is to blame for redundancy: we retrieve words when we can, and colors are easier to retrieve, so we throw them out there regardless of whether or not they're redundant. The problem with this is that this makes a prediction about the order of adjectives; in particular, the preferred order should be reversed. Pechmann does find some instances of this, but not very many. But there are other ways incrementality could play a role. For example, throwing out the color word may help when the noun is hard to retrieve. This predicts that in languages with post-nominal adjectives, where you can't use this as a delay strategy for holding off on planning the noun, there should be less color redundancy; indeed, Rubio-Fernandez 2016 shows this for Spanish. The dynamic nature of language processing plays a role in other ways, too: it allows us to update our beliefs about individual speakers' use of modifiers and generate better expectations about upcoming input. For example, Pogue et al 2016 have shown that listeners, after being exposed to consistently overinformative speakers, stop drawing early contrastive inferences based on modifier use.]

6.7 Extensions to other language production phenomena

In this paper, we have focused on providing an account of content selection (Gatt et al., 2013) in modified referring expressions on the one hand (i.e., when to mention an object's size or color) and in nominal referring expressions on the other (i.e., at which taxonomic level to refer to an object). Future work should investigate whether these models can be merged to jointly account for the choice of content expressed in modifiers and in nouns. Further, in order to scale up to more naturalistic conversational domains it will be necessary to consider richer language models. Recall that we treated different color names (e.g., *pink* and *purple*) as simply a color mention. Similarly, we treated different nouns that clearly referred at the same level (e.g., *grizzly* and *grizzly bear*) as simple sub level mentions. For the purpose of predicting not only content selection but also utterance choice, a richer inventory of utterance alternatives will need to be explored. An interesting question is how this approach can be extended to other referring expressions mentioned in the Introduction, e.g., names, pronouns, or referring expressions with post-nominal modification.

However, future research should also investigate the very intriguing potential for this approach to be extended to any language production phenomenon that involves content selection. For example, there is a large literature on optional instrument mentions. Brown and Dell (1987) showed that atypical instruments are more likely to be mentioned than typical ones – if a stabbing occurred with an icepick, speakers prefer "The man was stabbed with an ice pick" rather than "The man was stabbed". If instead a stabbing occurred with a knife, "The man was stabbed" is preferred over "The man was stabbed with a knife"). This is very much parallel to the case of atypical color mention. While Brown and Dell (1987)'s account of the effect is that speakers do or don't mention instruments for speaker-internal ego-centric reasons, later evidence suggests an explanation that is rather more driven by audience design considerations. Lockridge and Brennan (2002) replicated the original finding in a story retelling scenario while also manipulating whether or not addressees saw pictures of the actions. Without pictures, speakers produced even more mentions of atypical objects (presumably to prevent addressees from forming a faulty mental model of the situation), suggesting that the typicality effect is in fact an audience design effect.

More generally, the approach should extend to any content selection phenomenon that affords a choice between a more or less specific chunk of linguistic signal. Whenever the chunk adds sufficient information, it should be included. This is related to surprisal theories of production like Uniform Information Density (UID, Jaeger, 2006; Levy & Jaeger, 2007; A. Frank & Jaeger, 2008; Jaeger, 2010), where it has been found that speakers are more likely to omit linguistic signal if the underlying meaning or syntactic structure is highly predictable. Importantly, UID diverges from ours in that ours is (thus far) an account of *content selection*, while UID is an account of the choice between meaning-equivalent alternative *utterances*.

6.8 Conclusion

In conclusion, we have provided an account of redundant referring expressions that challenges the traditional notion of overinformativeness, unifies multiple language production literatures, and has the potential for many further extensions. For the time being, we take this work to suggest that, rather than being wastefully overinformative, speakers are rationally redundant.

[jd: What else needs to be included in GD?]

A Effects of fidelity on utterance probabilities

Here we visualize the effect of fidelity on the probability of producing the simple insufficient, simple sufficient, or complex redundant referring expression to refer to the target in contexts like Figure 1a and Figure 1b, under varying λ values, in Figure 22. This constitutes a generalization of Figure 3, which is duplicated in row 6.

B Model exploration for Koolen scene variation contexts

In Figure **??** we visualize model predicted probability of redundantly using color under varying λ values (columns), color fidelity values (rows), and size fidelity values (x-axis), for the high and low variation conditions in their Exp. 1 (where type was sufficient for reference) and Exp. 2 (where type and size was necessary for reference). The assumed type fidelity is .9.

C Validation of interactive web-based written production paradigm

make sure to discuss why overall we have lower overspecification rates – probably because of color typicality!! we had pretty typical colors in our stimuli

D Pre-experiment quiz

Before continuing to the main experiment, each participant had to correctly respond "True" or "False" to the following statements. Correct answers are given in parentheses after the statement.

- The speaker can click on an object. (False)
- The listener wants to click on the object that the speaker is telling them about. (True)

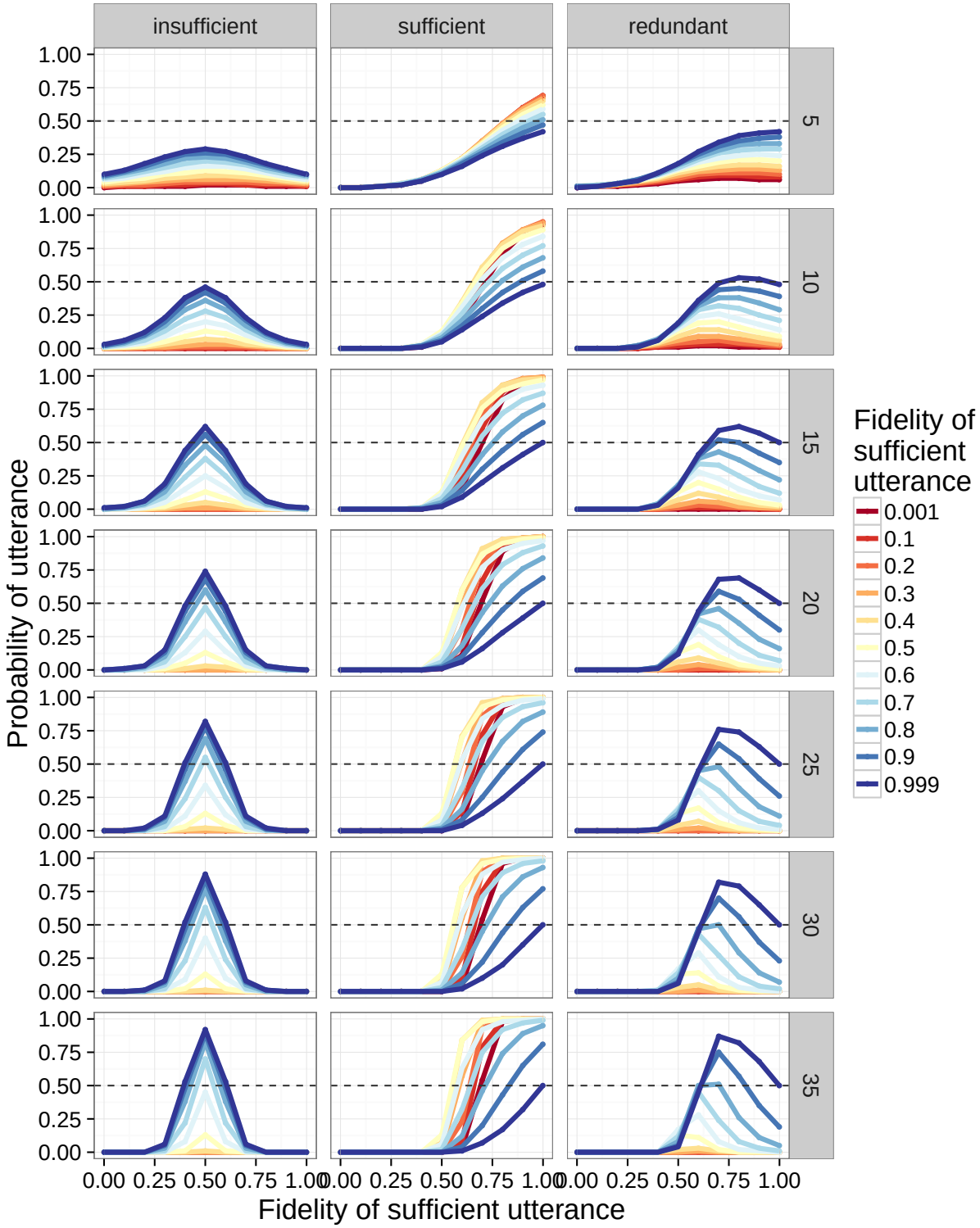


Figure 22: Utterance probability as a function of sufficient and insufficient utterance fidelities (x-axis, colors) and varying λ (rows).

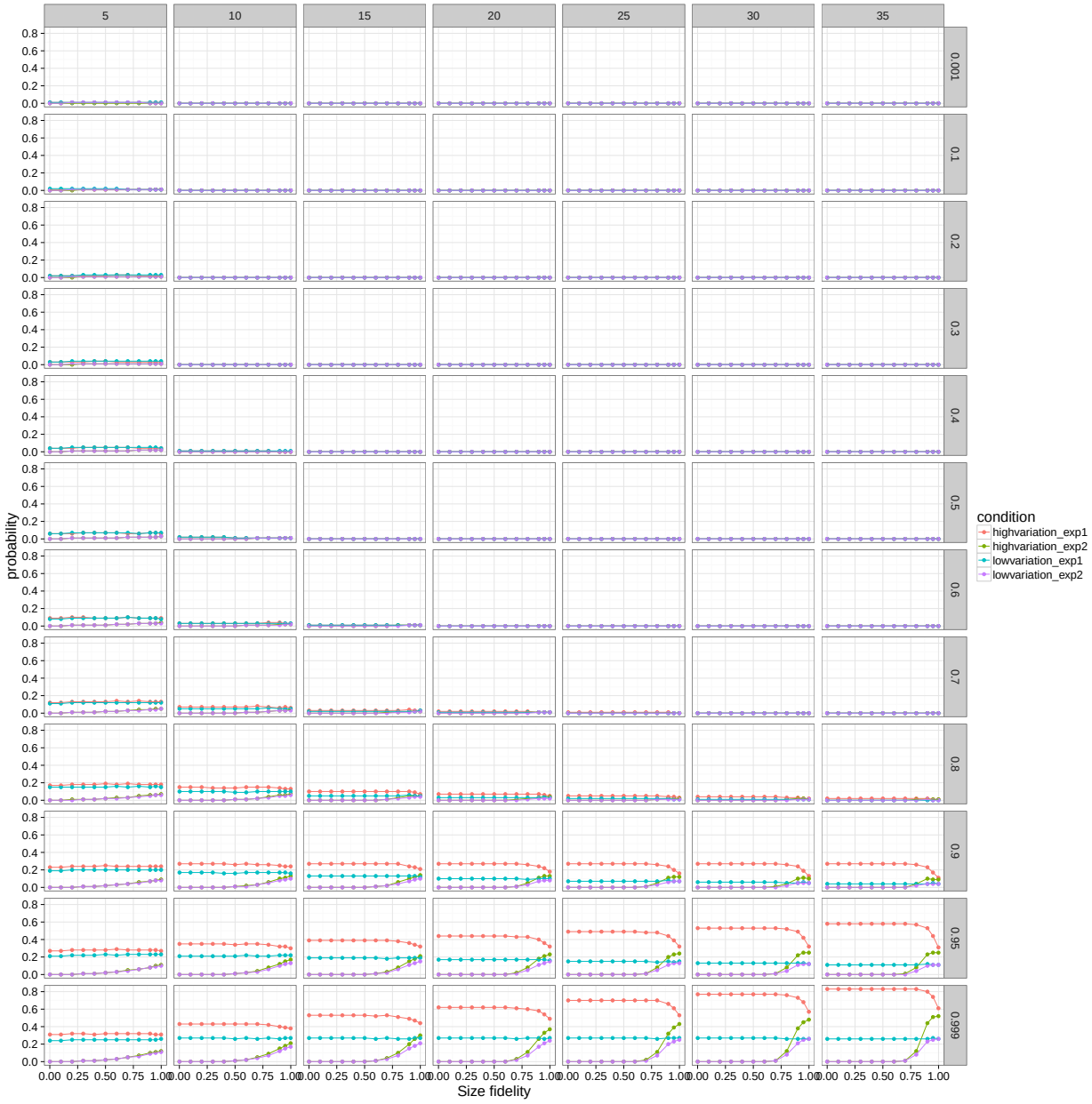


Figure 23: Probability of redundant color mention as a function of size fidelity (x-axis), color fidelity (rows), and varying λ (columns).

- The target is the object which has the red circle around it. (False)
- Only the speaker can send messages. (False)
- There are a total of 72 rounds. (True)
- The locations of the three objects are the same for the speaker and the listener. (False)

E Item types

The following table lists all 36 object types from Exp. XXX and the colors they appeared in:

Object	Colors	Object	Colors
avocado	black, green	balloon	pink, yellow
belt	black, brown	bike	purple, red
billiard ball	orange, purple	binder	blue, green
book	black, blue	bracelet	green, purple
bucket	pink, red	butterfly	blue, purple
candle	blue, red	cap	blue, orange
chair	green, red	coat hanger	orange, purple
comb	black, blue	cushion	blue, orange
flower	purple, red	frame	green, pink
golf ball	blue, pink	guitar	blue, green
hair dryer	pink, purple	jacket	brown, green
napkin	orange, yellow	ornament	blue, purple
pepper	green, red	phone	pink, white
rock	green, purple	rug	blue, purple
shoe	white, yellow	stapler	purple, red
thumb tack	blue, red	tea cup	pink, white
toothbrush	blue, red	turtle	black, brown
wedding cake	pink, white	yarn	purple, red

F Experiment 2a: typicality norms for Experiment 2

Analogous to the color typicality norms elicited for utterances in Exp. 1, we elicited typicality norms for utterances in Exp. 2. The elicited typicalities were used in the Bayesian Data Analysis reported in Section 5.3.

F.0.1 Methods

Participants We recruited 240 participants over Amazon's Mechanical Turk who were each paid $0.50 for their participation.

Procedure and materials On each trial, participants saw one of the images used in Exp. 2 and were asked to answer the question "How typical is this for an X?" on a continuous slider with endpoints labeled "very atypical" to "very typical." X was a nominal referring expression. In

contrast to Exp. 1a, where we only elicited typicality norms for utterance-object pairs where the object was in the extension of the utterance under a deterministic semantics (e.g., here *dalmatian*, *dog*, or *animal* for a dalmatian), in this norming study we also elicited norms for utterance-object pairs where that was not clearly the case (e.g., *a bear* for a bison, *a car* for an ambulance, or *a snack* for a lobster). However, we did not test all utterance-object combinations, which would have led to an explosion of conditions. Instead, we tested each target object with its three utterances (e.g., the dalamtian was paired with *dalmatian*, *dog*, and *animal*; the pug was paired with *pug*, *dog*, and *animal*, etc.). That yielded a total of 108 combinations – four targets in nine domains with three utterances each. We further tested each distractor item that shared the target's superclass category (*dist-samesuper*, e.g., cows share the superclass category animal with dogs) on both the basic level and the super level term (e.g., *dog* for cow and *animal* for cow), for a total of 469 combinations. Finally, we also tested each distractor of a different super category than the target on the target's super level term (*dist-diffsuper*, e.g., *animal* for socks). This yielded another 168 combinations. Overall, we obtained typicality norms for 745 object-utterance combinations. All other object-utterance combinations were assumed to have typicality 0.

Each participant rated 45 items: 7 targets, 10 dist-diffsuper, and 28 dist-samesuper cases. These were randomly sampled from the overall pool of items in each category.

F.0.2 Results and discussion

Each combination was rated at least 5 times and at most 27 times. We coded the slider endpoints as 0 ("very atypical") and 1 ("very typical"). In order to evaluate the model, we used each object-utterance combination's typicality mean as input.

Typicality ratings by item type (target, dist-samesuper, dist-diffsuper) and utterance type (sub, basic, super) are visualized in Figure 24. As expected, typicality was close to 0 for dist-diffsuper cases and for sub/basic terms used with dist-samesuper cases. However, even for these cases, there was some variation.

For targets, typicality of the object for the utterance decreased with increasing reference level, mirroring the typicality ratings obtained for Exp. 1 – a particular object is a better instance of the more specific term than of the more general term for that object.

G Nominal choice model comparison

[jd: This isn't model comparison in the technical sense, just a side-by-side look at the different models. Leave it in or throw out?]

Here we report correlations, MAP estimates of posterior predictives collapsed across targets and items, and scatterplots of posterior predictive MAP estimates on the by-target level for the model containing a) only informativeness with deterministic semantics; b) informativeness with deterministic semantics and cost; c) only informativeness with non-deterministic semantics; d) informativeness with non-deterministic semantics and cost (the model reported in the main text). Table 11 shows correlations. Figure 25 shows the collapsed patterns for utterance choice. Figure 26 shows the scatterplots.

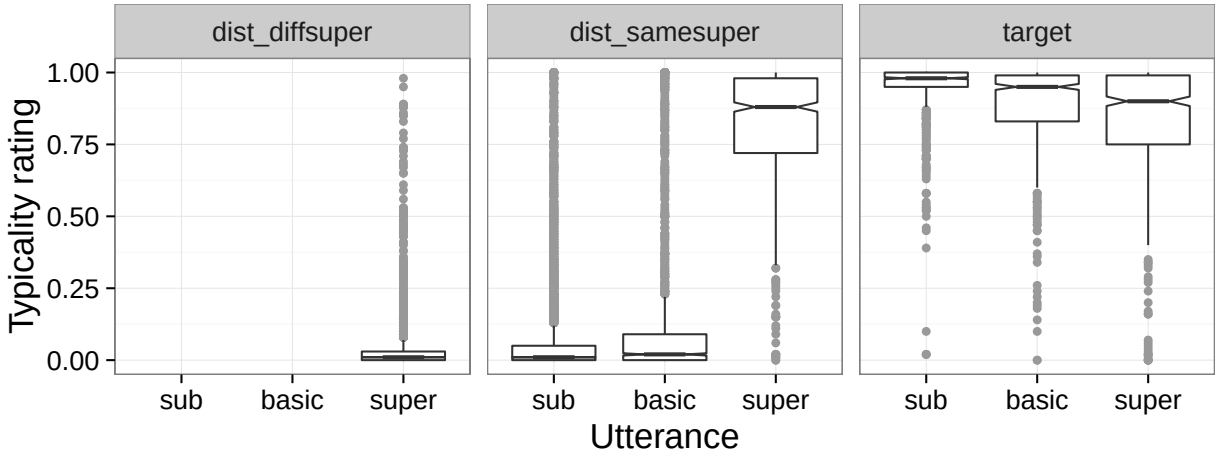


Figure 24: Boxplots of typicality ratings. The lower and upper hinges correspond to the first and third quartiles (the 25th and 75th percentiles). Upper and lower whiskers extend from the respective hinge to the highest and lowest values that are within 1.5 times the inter-quartile range of the hinge. Outliers are indicated as gray dots.

Table 11: Correlations (r and R^2) of posterior predictive MAPs of four different models (see main text) with empirical proportions of sub, basic, and super level choices.

		Model			
Semantics		deterministic	deterministic	non-deterministic	non-deterministic
Cost		no	yes	no	yes
r	collapsed	.85	.88	.86	.94
	by-target	.63	.71	.71	.84
R^2	collapsed	.72	.77	.74	.89
	by-target	.40	.51	.51	.70

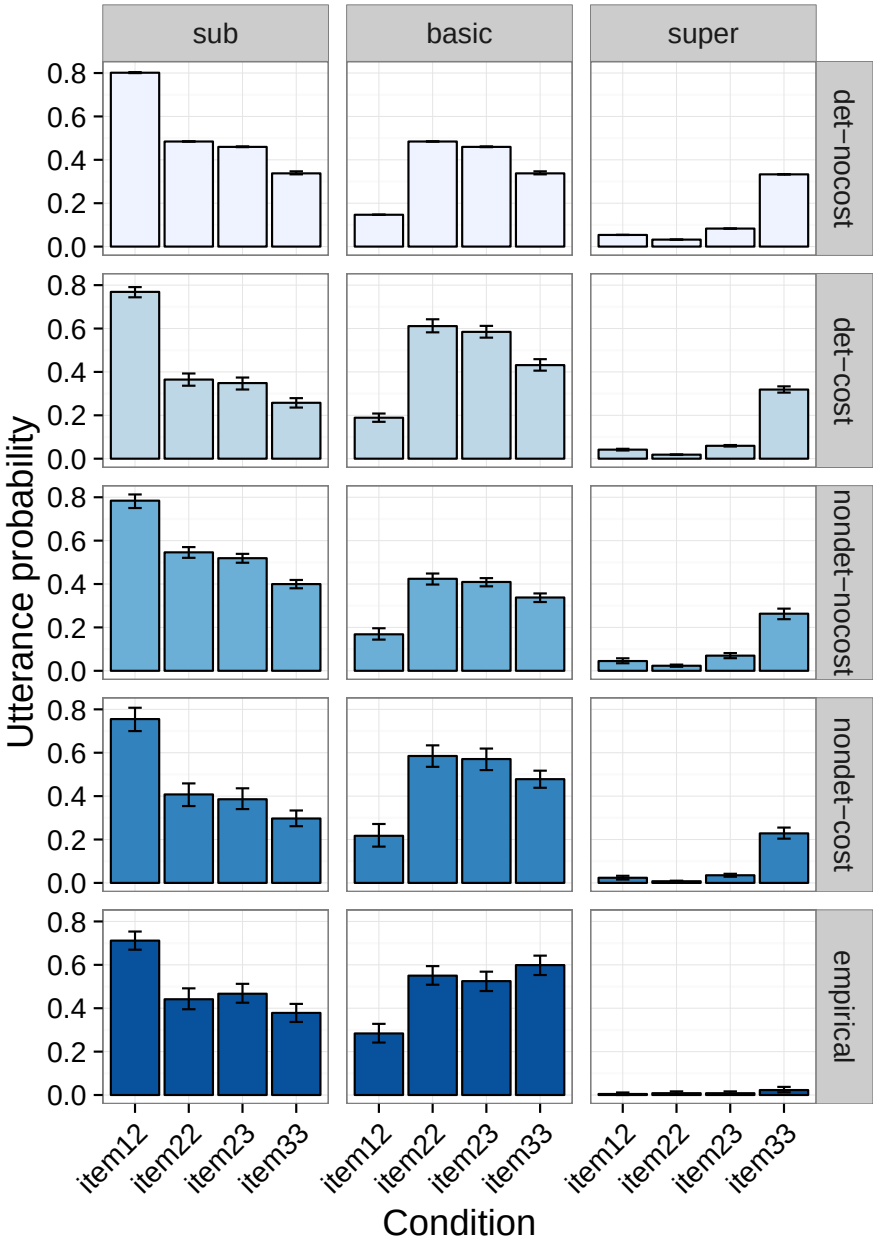


Figure 25: Utterance probabilities across different conditions. Columns indicate utterances, rows indicate data type (empirical proportion, MAP estimates of posterior predictives for the four different models).

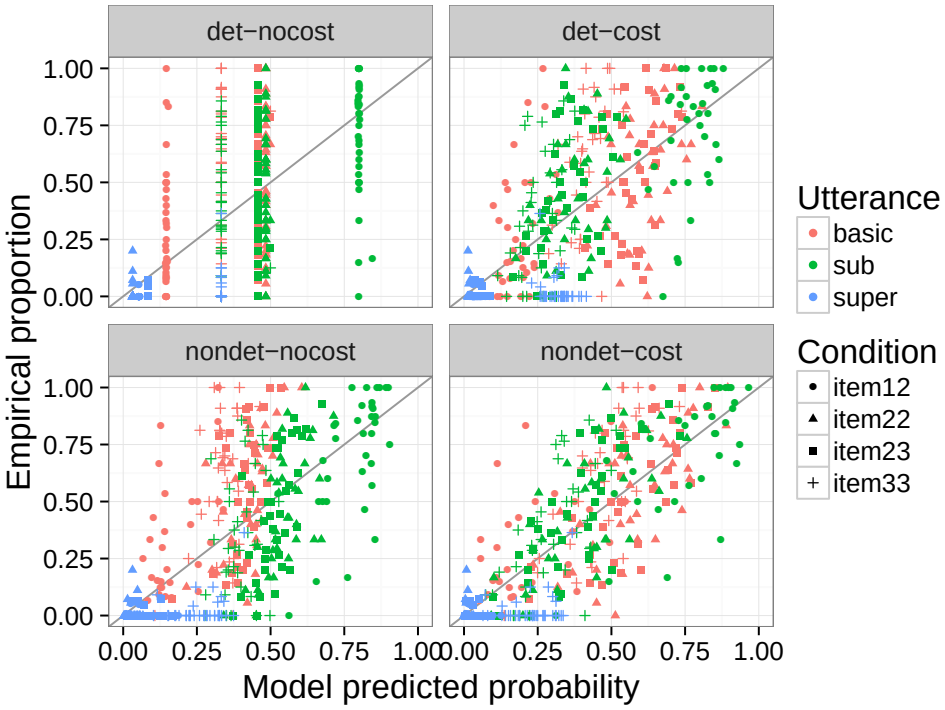


Figure 26: Scatterplot of by-target empirical utterance proportions against model posterior predictive MAP estimates for the four different models. Gray line indicates perfect correlation line.

H Gatt replication

report Gatt et al 2011 replication

References

Arts, A., Maes, A., Noordman, L., & Jansen, C. (2011). Overspecification facilitates object identification. *Journal of Pragmatics*, *43*(1), 361–374. Retrieved from `http://dx.doi.org/10.1016/j.pragma.2010.07.013` doi: 10.1016/j.pragma.2010.07.013

Baumann, P., Clark, B., & Kaufmann, S. (2014). Overspecification and the Cost of Pragmatic Reasoning about Referring Expressions. In *Proceedings of the 36th annual conference of the cognitive science society* (pp. 1898–1903). Austin, TX: Cognitive Science Society.

Belke, E., & Meyer, A. S. (2002). Tracking the time course of multidimensional stimulus discrimination: Analyses of viewing patterns and processing times during same-different decisions. *European Journal of Cognitive Psychology*, *14*(2), 237–266. doi: 10.1080/09541440143000050

Brown, P., & Dell, G. (1987). Adapting Production to Comprehension : Mention of Instruments. *Cognitive Psychology*, *472*, 441–472.

Brown-Schmidt, S., & Heller, D. (2014). What language processing can tell us about perspective taking: A reply to Bezuidenhout (2013). *Journal of Pragmatics*, *60*, 279–284. Retrieved from `http://dx.doi.org/10.1016/j.pragma.2013.09.003` doi: 10.1016/j.pragma.2013.09.003

Clark, H. H., & Murphy, G. L. (1982). Audience Design in Meaning and Reference. *Advances in Psychology*, *9*(C), 287–299. doi: 10.1016/S0166-4115(09)60059-5

Dale, R. (1989). Cooking up referring expressions. *Proceedings of the 27th annual meeting on Association for Computational Linguistics (ACL'89)*, 68–75. Retrieved from `http://portal.acm.org/citation.cfm?doid=981623.981632` doi: 10.3115/981623.981632

Dale, R., & Reiter, E. (1995). Computational Interpretations of the Gricean Maxims in the Generation of Referring Expressions . *Cognitive Science*, *19*, 233 – 263.

Davies, C., & Katsos, N. (2013). Are speakers and listeners 'only moderately Gricean'? An empirical response to Engelhardt et al. (2006). *Journal of Pragmatics*, *49*(1), 78–106. Retrieved from `http://dx.doi.org/10.1016/j.pragma.2013.01.004` doi: 10.1016/j.pragma.2013.01.004

Degen, J., Franke, M., & Jäger, G. (2013). Cost-Based Pragmatic Inference about Referential Expressions. In *Proceedings of the 35th annual conference of the cognitive science society.*

Engelhardt, P. E., Bailey, K., & Ferreira, F. (2006a, may). Do speakers and listeners observe the Gricean Maxim of Quantity? *Journal of Memory and Language*, *54*(4), 554–573. Retrieved from `http://linkinghub.elsevier.com/retrieve/pii/S0749596X05001518` doi: 10.1016/j.jml.2005.12.009

Engelhardt, P. E., Bailey, K., & Ferreira, F. (2006b, may). Do speakers and listeners observe the Gricean Maxim of Quantity? *Journal of Memory and Language*, *54*(4), 554–573. Retrieved from `http://linkinghub.elsevier.com/retrieve/pii/S0749596X05001518` doi: 10.1016/j.jml.2005.12.009

Engelhardt, P. E., Demiral, S. B., & Ferreira, F. (2011). Over-specified referring expressions impair comprehension: An ERP study. *Brain and Cognition*, *77*(2), 304–314. doi: 10.1016/j.bandc.2011.07.004

Frank, A., & Jaeger, T. F. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. In *The 30th annual meeting of the cognitive science society.*

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*, 998.

Gatt, A., van Gompel, R. P. G., Krahmer, E., & van Deemter, K. (2011). Non-deterministic attribute selection in reference production. In *Proceedings of the workshop on production of referring expressions: Bridging the gap between empirical, computational and psycholinguistic approaches to reference (pre-cogsci11).* Boston. Retrieved from `E:$\backslash$Disser$\backslash$Bibliography$\backslash$gatt2011non.pdf`

Gatt, A., van Gompel, R. P. G., van Deemter, K., & Krahmer, E. (2013). Are we Bayesian referring expression generators ? In *Proceedings of the workshop on production of referring expressions: Bridging the gap between computational and cognitive approaches to reference (pre-cogsci'13).*

Goodman, N. D., & Stuhlmüller, A. (2013, jan). Knowledge and implicature: modeling language understanding as social cognition. *Topics in Cognitive Science*, *5*(1), 173–84. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/23335578` doi: 10.1111/tops.12007

Goodman, N. D., & Stuhlmüller, A. (electronic). *The design and implementation of probabilistic programming languages.* Retrieved 2015/1/16, from `http://dippl.org`

Graf, C., Degen, J., Hawkins, R. X. D., & Goodman, N. D. (2016). Animal , dog , or dalmatian ? Level of abstraction in nominal referring expressions. In A. Papafragou, D. Grodner, D. Mirman, & J. Trueswell (Eds.), *Proceedings of the 38th annual conference of the cognitive science society* (pp. 2261–2266). Austin, TX: Cognitive Science Society.

Grice, H. P. (1975). Logic and Conversation. *Syntax and Semantics*, *3*, 41–58. Retrieved from `http://books.google.com/books?hl=en{\&}amp;lr={\&}amp;id=hQCzOmaGeVYC{\&}amp;oi=fnd{\&}amp;pg=PA121{\&}amp;dq=Logic+and+conversation{\&}amp;ots=j7aijUymwm{\&}amp;sig=iV1rz1eEm4ns6bQ6CevIURXFVO4`

Hawkins, R. X. D. (2015). Conducting real-time multiplayer experiments on the web. *Behavior Research Methods*, *47*(4), 966-976.

Hawkins, R. X. D., Stuhlmüller, A., Degen, J., & Goodman, N. D. (2015). Why do you ask ? Good questions provoke informative answers . In *Proceedings of the 37th annual conference of the cognitive science society.*

Herrmann, T., & Deutsch, W. (1976). *Psychologie der Objektbenennung.* Huber.

Horton, W., & Keysar, B. (1996). When do speakers take into account common ground? *Cognition*, *59*, 91–117.

Huettig, F., & Altmann, G. T. M. (2011). Looking at anything that is green when hearing "frog": how object surface colour and stored object colour knowledge influence language-mediated overt attention. *Quarterly journal of experimental psychology (2006)*, *64*(1), 122–145. doi: 10.1080/17470218.2010.481474

Jaeger, T. F. (2006). *Redundancy and Reduction in Spontaneous Speech* (Unpublished doctoral dissertation). Stanford University.

Jaeger, T. F. (2010). Redundancy and reduction: speakers manage syntactic information density. *Cognitive Psychology*, *61*(1), 23–62.

Jolicoeur, P., Gluck, M. A., & Kosslyn, S. M. (1984). Pictures and names: Making the connection. *Cognitive Psychology*, *16*(2), 243–275.

Kao, J., Wu, J., Bergen, L., & Goodman, N. D. (2014). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(33), 12002–12007.

Koolen, R., Gatt, A., Goudbeek, M., & Krahmer, E. (2011). Factors causing overspecification

in definite descriptions. *Journal of Pragmatics*, *43*(13), 3231–3250. Retrieved from http://dx.doi.org/10.1016/j.pragma.2011.06.008 doi: 10.1016/j.pragma.2011.06.008

Koolen, R., Goudbeek, M., & Krahmer, E. (2013). The effect of scene variation on the redundant use of color in definite reference. *Cognitive Science*, *37*(2), 395–411. doi: 10.1111/cogs.12019

Levy, R., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In B. Schlökopf, J. Platt, & T. Hoffman (Eds.), *Advances in neural information processing systems* (Vol. 19, pp. 849–856). Cambridge, MA: MIT Press. Retrieved from http://scholar.google.com/scholar?hl=en{\&}btnG=Search{\&}q=intitle:Speakers+optimize+information+density+through+syntactic+reduction{\#}0

Lockridge, C. B., & Brennan, S. E. (2002, sep). Addressees' needs influence speakers' early syntactic choices. *Psychonomic bulletin & review*, *9*(3), 550–7. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12412896

Maes, A., Arts, A., & Noordman, L. (2004). Reference Management in Instructive Discourse. *Discourse Processes: A Multidisciplinary Journal*, *37*(2), 117–144. Retrieved from http://proxy.lib.uiowa.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true{\&}db=eric{\&}AN=EJ682763{\&}site=ehost-live$\backslash$nhttp://www.leaonline.com doi: 10.1207/s15326950dp3702{_}3

Mitchell, M. (2013). Typicality and object reference. *Proceedings of the 35th ...*, 3062–3067. Retrieved from http://csjarchive.cogsci.rpi.edu/Proceedings/2013/papers/0547/paper0547.pdf

Nadig, A. S., & Sedivy, J. C. (2002, jul). Evidence of Perspective-Taking Constraints in Children's On-Line Reference Resolution. *Psychological Science*, *13*(4), 329–336. Retrieved from http://pss.sagepub.com/lookup/doi/10.1111/j.0956-7976.2002.00460.x doi: 10.1111/j.0956-7976.2002.00460.x

Paraboni, I., van Deemter, K., & Masthoff, J. (2007). Generating Referring Expressions: Making Referents Easy to Identify. *Computational Linguistics*, *33*(2), 229–254. doi: 10.1162/coli.2007.33.2.229

Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics*, *27*(1), 89–110. doi: 10.1515/ling.1989.27.1.89

Rosch, E. (1973, may). Natural categories. *Cognitive Psychology*, *4*(3), 328–350. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/0010028573900170 doi: 10.1016/0010-0285(73)90017-0

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*(3), 382–439. doi: 10.1016/0010-0285(76)90013-X

Rubio-Fernandez, P. (2016). How redundant are redundant color adjectives? An efficiency-based analysis of color overspecification. *Frontiers in Psychology*, *7*(153). doi: 10.3389/fpsyg.2016.00153

Sedivy, J. C. (2003, jan). Pragmatic versus form-based accounts of referential contrast: evidence for effects of informativity expectations. *Journal of psycholinguistic research*, *32*(1), 3–23. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12647560

Tanaka, J. W., & Taylor, M. (1991). Object categories and expertise: Is the basic-level in the eye of the beholder? *Cognitive Psychology*, *23*, 457–482. doi: 10.1016/0010-0285(91)90016-H

Westerbeek, H., Koolen, R., & Maes, A. (2015). Stored object knowledge and the production of referring expressions: the case of color typicality. *Frontiers in Psychology*, *6*(July), 1–12. Retrieved from http://journal.frontiersin.org/Article/10.3389/fpsyg.2015.00935/

abstract doi: 10.3389/fpsyg.2015.00935