Over ‘overinformativeness’: rationally redundant referring expressions [ndg: consider alternatives: ”rationally redundant referring expressions”, ”A Bayesian approach to redundancy in reference”, ...?]

Judith Degen   Robert X.D. Hawkins   Caroline Graf   Elisa Kreiss
Noah D. Goodman

February 14, 2019

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 supports the view of a language production system geared towards communicative efficiency. [ndg: note to self: review after full read.]

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

> Merrick: “The vaccine will be distributed gratis.”
> Al: Free gratis.
> Merrick: “Free gratis” is a redundancy.
> EB: Does that mean, repeats itself?
> Al: Then leave “gratis” out.
> Merrick: What luck for me, Al, that you have such a keen editorial sense. Free. “Distributed free.” Period.
>
> Deadwood, Season 1, Episode 6 [ndg: while this is amusing... it's not really very relevant?]

1 Introduction

Reference to objects is one of the most basic and prevalent uses of language. In order to refer, speakers must choose from amongst a 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 establish unique 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. [ndg: include basic example of over-informativeness here? eg point to fig 1, and include a bit more explanation either in caption or here. (a speaker often says "little blue pin" when "little pin" is sufficient.)] Providing a unified theory for speakers' systematic patterns of overinformativeness has so far proven elusive.

This paper is concerned with accounting for these systematic patterns in overinformative referring expressions. 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, with any number of adjectives occurring between the determiner and the noun.[1] A model of these referring expressions will allow us to unify two domains in language production that have been typically treated as separate, and that have been discussed for different reasons: the production of (purportedly) overmodified referring expressions on the one hand, which much language production literature has been devoted to (Herrmann & Deutsch, 1976; Pechmann, 1989; Nadig & Sedivy, 2002; Sedivy, 2003; Maes, Arts, & Noordman, 2004; Engelhardt, Bailey, & Ferreira, 2006; 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) and in the developmental literature on generalizing basic level terms (?, ?). In the following, we review some of the key phenomena and puzzles in each of these literatures. We then present a model of referring expression production within the Rational Speech Act framework (M. C. Frank & Goodman, 2012; Goodman & Frank, 2016), 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. Treating speakers as mostly rational agents operating on a relaxed semantics predicts that adding seemingly overinformative modifiers or using nouns that are seemingly too specific can be useful and informative, to the extent that not doing so might allow the listener to go astray (or to invest too much processing effort in inferring the speaker's correct intention). This provides a unified explanation for a number of seemingly disparate phenomena from the modified and nominal referring expression literature.

We spend the remainder of the paper demonstrating how this account applies to various phenomena. In Section 1 we spell out the problem and introduce the key overinformativeness phenomena. In Section 2 we introduce the basic (deterministic semantics) and modified (relaxed semantics) Rational Speech Act framework. In Sections 3 - 5 we evaluate the relaxed semantics RSA model on

[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.

data from interactive online reference game experiments that exhibit the phenomena introduced in Section 1: size and color modifier choice under varying conditions of scene complexity; typicality effects in the choice of color modifier; and choice of nominal level of reference. We wrap up in Section 6 by summarizing our findings and discussing the far-reaching implications of and further challenges for this line of work.

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

How should a cooperative speaker choose between competing 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 is required (for the purposes of the exchange).

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

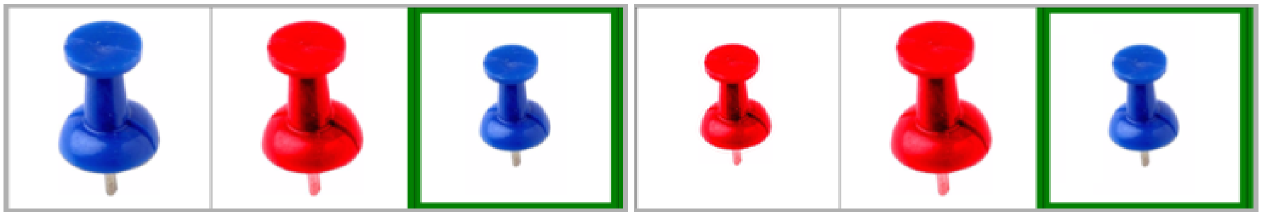
That is, speakers should aim to produce neither under- nor overinformative utterances. While much support has been found for the avoidance of underinformativeness (Brennan & Clark, 1996; R. Brown, 1958; Olson, 1970; Levinson, 1983; Engelhardt et al., 2006; ?, ?), speakers seem remarkably happy to systematically violate Quantity-2. In modified referring expressions, they routinely produce modifiers that are not necessary for uniquely establishing reference (e.g., *the small blue pin* instead of *the small pin* in contexts like Figure 1a (Gatt, van Gompel, Krahmer, & van Deemter, 2011; Gatt, Krahmer, van Deemter, & van Gompel, 2014; Arts et al., 2011; Koolen et al., 2011)). 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 17 (Rosch et al., 1976; Hoffmann & Ziessler, 1983; Tanaka & Taylor, 1991a; Johnson & Mervis, 1997; R. Brown, 1958)).

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 (Engelhardt et al., 2006; Pechmann, 1989). Some have derived a built-in preference for referring at the basic level from considerations of perceptual factors such as shape (Rosch et al., 1976; Rosch, 1973; Murphy & Smith, 1982). Others have argued for salience-driven effects on willingness to overmodify (Gatt et al., 2014; Westerbeek, Koolen, & Maes, 2015). 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



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

Figure 1: Example contexts where (a) size only or (b) color only is sufficient for unique reference. A green border marks the intended referent.

frequently do not include only the minimal modifiers required for establishing reference, but often also include redundant modifiers (Pechmann, 1989; Nadig & Sedivy, 2002; Maes et al., 2004; Engelhardt et al., 2006; 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 others (Sedivy, 2003). Here we review some key patterns of overmodification that have been observed, before spelling out our account of these phenomena in Section 2.

1.2.1 Asymmetry in redundant use of color and size adjectives

In Figure 1a, singling out the object highlighted by the green border requires only mentioning its size (*the small pin*). But it is now well-documented that speakers routinely include redundant color adjectives (*the small blue pin*) which are not necessary for uniquely singling out the intended referent in these kinds of contexts (Pechmann, 1989; Belke & Meyer, 2002; Gatt et al., 2011). However, the same is not true for size: in contexts like Figure 1b, where color is sufficient for unique reference (*the blue pin*), speakers overmodify much more rarely. Though there is quite a bit of variation in proportions of overmodification, this asymmetry in the propensity for overmodifying with color but not size has been documented repeatedly (Pechmann, 1989; Sedivy, 2003; Gatt et al., 2011; Rubio-Fernandez, 2016; Westerbeek et al., 2015; Koolen, Goudbeek, & Krahmer, 2013).

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 pin*, which violate the preference for size adjectives to occur before color adjectives in English (Bloomfield, 1933; Sproat & Shih, 1991). While Pechmann did observe such violations in his dataset, most cases of overmodification did not constitute such violations, and he himself concluded 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). We return to this idea in Section 2 and the General Discussion.

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. This is equivalent to the very simple interpretation of Grice laid out above, and consequently these models demonstrated the same inability to capture the color-size asymmetry: they only produced 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.

1.2.2 Scene variation

So far we have portrayed speakers' propensity to overmodify with color as a fixed quantity (though varying by experiment). However, this propensity 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 (Davies & Katsos, 2013; Koolen et al., 2013). How exactly scene variation is quantified differs across experiments. 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. Similarly, Davies and Katsos (2013) found that listeners judge overmodified referring expressions in low-variation scenes of four objects as less natural than in high-variation scenes of 4 potentially compositional 'objects-on-objects' (e.g., a button on a sock). And finally, Gatt, Krahmer, Van Deemter, and van Gompel (2017), while not reporting differences in overmodification behavior, did find that when size and color are jointly disambiguating, speech onset times

(a) Typical color, type sufficient. (b) Atypical color, type sufficient.

Figure 2: Example contexts where type (*banana*) is sufficient for unique reference and color is (a) typical or (b) atypical. A green border marks the intended referent.

for non-redundant *color-and-size* utterances increased as the number of distractors in the display increased.

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).

Above, we have considered three different ways of quantifying scene variation: the number of dimensions along which objects differ, whether objects are 'simple' or 'compositional', and the number of distractors present in a scene. A model of referring expression generation should ideally capture all of these types of variation in a unified way.

1.2.3 Feature typicality

Modifier type and amount of scene variation are not the only factors determining overmodification. Overmodification with color has been shown to be systematically related to the typicality of the color for the object. Building on work by Sedivy (2003), Westerbeek et al. (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 in the absence of other bananas as *the yellow banana* (see Figure 2a), 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* (see Figure 2b). 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. Similarly, Arts et al. (2011) showed that overspecified expressions result in faster referent identification. 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 for the time being and will return to this issue in the General Discussion.

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 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* (being a dalmatian entails being a 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 that differ in word length.

Nevertheless, cognitive cost affects the choice of reference level: in particular, speakers prefer more frequent words over less frequent ones (Oldfield & Wingfield, 1965), and they prefer shorter ones over longer ones (Degen, Franke, & Jäger, 2013; Rohde, Seyfarth, Clark, Jäger, & Kaufmann, 2012). 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, 1991b). That is, in the absence of other constraints, even when a superordinate level term would be sufficient for establishing reference (as in Figure 3b), speakers prefer to say *the dog* rather than *the animal*.

Contextual informativeness is another factor that has been shown to affect speakers' nominal production choices (e.g., Brennan & Clark, 1996). For instance, in a context like Figure 3a, speakers should use the subordinate level term *dalmatian* to refer to the target marked with a green border, because a higher-level term (*dog*, *animal*) would be contextually underinformative. However, there are nevertheless cases of contexts where either the superordinate *animal* or the basic level *dog* term would be sufficient for unique reference, as in Figure 3b, in which speakers nevertheless 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, compared to the other objects in the context. 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 many factors, including an utterance's informativeness, its cost relative to alternative utterances, and the typicality of an object or its features. Critically, these factors are all in play at once, potentially interacting in rich and complex ways. In the next section, we provide an explicit computational account of these different factors and how they interact, with a focus on cases where speakers appear to be overinformative – either by adding more modifiers or by referring at a more specific

(a) Subordinate level term necessary. (b) Superordinate level term sufficient.

Figure 3: Example contexts in which different levels of reference are necessary for establishing unique reference to the target marked with a green border. (a) subordinate (*dalmatian*) necessary; (b) superordinate (*animal*) sufficient, but basic (*dog*) or subordinate (*dalmatian*) possible.

Table 1: List of effects a theory of referring expression production should account for and paper section(s) in which they are treated.

Section	Effect	Description
2 & 3	Color/size asymmetry	More redundant use of color than size [2]
2 & 3	Scene variation	More redundant use of color with increasing scene variation [3]
4	Color typicality	More redundant use of color with decreasing color typicality [4]
5	Basic level preference	Preference for basic level term when superordinate sufficient [5]
5	Subordinate level use	Unnecessary use of subordinate level term [6]

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 1.

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, which we then use to capture these disparate phenomena in one model.

2 Modeling speakers' choice of referring expression

Here we propose a computational model of referring expression production that accounts for the phenomena introduced above. The model is formulated within the Rational Speech Act (RSA) framework (M. C. Frank & Goodman, 2012; Goodman & Frank, 2016).[7] It provides a principled

[2] Reported by many (e.g., Pechmann, 1989; Engelhardt et al., 2006; Gatt et al., 2011; Rubio-Fernandez, 2016)

[3] Multiple replications reported (e.g., Davies & Katsos, 2013; Koolen et al., 2013)

[4] Multiple replications reported (e.g. Sedivy, 2003; Westerbeek et al., 2015; Rubio-Fernandez, 2016)

[5] Originally reported by Rosch et al. (1976), dozens of replications.

[6] Reported by Jolicoeur et al. (1984)

[7] All RSA models and Bayesian data analyses reported in this paper were implemented in the probabilistic programming language WebPPL (Goodman & Stuhlmüller, electronic) and can be viewed at `https://github.com/thegricean/RE_production`. All experimental materials and analysis scripts are available in the same repository.

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 Section 6. We proceed by first presenting the general framework in Section 2.1, and show why the most basic model, as formulated by M. C. Frank & Goodman, 2012, does not produce the phenomena outlined above due to its strong focus on speakers maximizing the informativeness of 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.

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) does not generate overinformative referring expressions for two reasons. The first is trivial: the model was only applied to one-word utterances. This is easily remedied by simply including two-word utterances in the set of utterance alternatives. [jd: ndg, do you like this formulation better than the previous one?] The second is more substantial: even when allowing two-word (or n-word) utterances, the speaker's utility function does not allow for producing redundant referring expressions as long as additional words contribute non-negative costs to the overall utterance cost. To see this, and as a basis for the innovation introduced in Section 2.2 it is useful to reiterate the basic form of the model.

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:

$$P_{L_0}(o|u) \propto \mathcal{L}(u, o). \tag{1}$$

The literal listener L_0 observes an utterance u from the set of utterances U, consisting of single adjectives denoting features available in the context of a set of objects O, and returns a distribution over objects $o \in O$. Here, $\mathcal{L}(u, o)$ is the lexicon that encodes deterministic lexical meanings such that:

$$\mathcal{L}(u, o) = \begin{cases} 1 & \text{if } u \text{ is true of } o \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Thus, $P_{L_0}(o|u)$ returns a uniform distribution over all contextually available o in the extension of u. For example, in the size-sufficient context shown in Figure 1a, $U = \{big, small, blue, red\}$ and $O = \{o_{\text{big\_blue}}, o_{\text{big\_red}}, o_{\text{small\_blue}}\}$. Upon observing *blue*, the literal listener therefore assigns equal probability to $o_{\text{big\_blue}}$ and $o_{\text{small\_blue}}$. Values of $P_{L_0}(o|u)$ for each u are shown on the left in Table 2.

The pragmatic speaker in turn produces an utterance with probability proportional to the utility of that utterance:

$$P_{S_1}(u|o) \propto e^{U(u,o)} \tag{3}$$

The speaker's utility $U(u, o)$ is a function of both the utterance's *informativeness* with respect to the literal listener $P_{L_0}(o|u)$ and the utterance's *cost* $c(u)$:

$$U(u, o) = \beta_i \ln P_{L_0}(o|u) - \beta_c c(u) \quad (4)$$

Two free parameters, β_i and β_c enter the computation, weighting the respective contributions of informativeness and utterance cost, respectively.[8] In order to understand the effect of β_i, it is useful to explore its effect when utterances are cost-free. In this case, as β_i approaches infinity, the speaker increasingly only chooses utterances that maximize informativeness; if β_i is 0, informativeness is disregarded and the speaker chooses randomly from the set of all available utterances; if β_i is 1, the speaker probability-matches, i.e., chooses utterances proportional to their informativeness (equivalent to Luce's choice rule, Luce, 1959). Applied to the example in Table 2, if the speaker wants to refer to $o_{\text{small\_blue}}$ they have two semantically possible utterances, *small* and *blue*, where *small* is twice as informative as *blue*. They produce *small* with probability 1 when $\beta_i \to \infty$, probability 2/3 when $\beta_i = 1$ and probability 1/4 when $\beta_i = 0$.[9]

Conversely, disregarding informativeness and focusing only on cost, 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 does not 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} that consist of a two adjective sequence $u_{\text{size}} \in \{\textit{big}, \textit{small}\}$ and $u_{\text{color}} \in \{\textit{blue}, \textit{red}\}$, such that the meaning of a complex two-word utterance is defined as

$$\mathcal{L}(u_{\text{complex}}, o) = \mathcal{L}(u_{\text{size}}, o) \times \mathcal{L}(u_{\text{color}}, o). \quad (5)$$

The resulting renormalized literal listener distributions for our example size-sufficient context in Figure 1a are shown in the middle columns in Table 2.

Unfortunately, simply including complex utterances in the set of alternatives does not solve the problem. 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 as each other, for referring to the small blue object. Because they are equally contextually informative, the only way for the complex utterance to be chosen with greater probability than the simple utterance is if it was the *cheaper* one. While this would achieve the desired mathematical effect, the cognitive plausibility of complex utterances being cheaper than simple utterances is highly dubious. Even if it wasn't dubious: as mentioned previously, proportions of overinformative referring expressions are variable across experiments. The only way to achieve

[8] M. C. Frank and Goodman (2012) fixed $\beta_i = 1$ and did not include cost in their formulation, because they assumed equal costs for all utterances. Subsequent work has demonstrated the importance of taking into account utterance cost in modeling interpretation phenomena like cost-based quantity implicatures (Degen, Franke, & Jäger, 2013) and M-implicature (Bergen, Levy, & Goodman, 2016). We include it here because of the importance that cost has played in explanations of overinformative referring expressions, where it typically surfaces as the idea that speakers have different overall preferences for mentioning color vs. size modifiers (Dale & Reiter, 1995; Koolen et al., 2011; Gatt et al., 2013). At this point we remain agnostic about the factors that contribute to an utterance's cost $c(u)$. In later sections we allow cost to be a function of properties (e.g. color & size) mentioned in the utterance, or of an utterance's empirical length and corpus frequency; our policy for these cases is to introduce free cost parameters for each linear component of the cost function.

[9] Note that instead of a β_i parameter weighting informativeness *inside* the utility function, other recent formulations have used an α parameter modulating the entire utility function, i.e. $P_{S_1}(u|o) \propto \exp \alpha U(u, o)$. These parameterizations are equivalent. In the present work, where informativeness and cost both play important roles, we chose the 'flattened' linear combination with independent weights for simplicity.

Table 2: Row-wise literal listener distributions $P_{L_0}(o|u)$ for each utterance u in the size-sufficient context depicted in Figure 1a, allowing only simple one-word utterances (left) or one- and two-word utterances (middle, right) under a deterministic semantics (left, middle) or under a continuous semantics (right) with $x_{\text{size}} = .8$, $x_{\text{color}} = .99$. Bolded numbers indicate crucial comparisons between literal listener probabilities in correctly selecting the intended referent $o_{\text{small\_blue}}$ in response to observing the sufficient *small* and the redundant *small blue* utterances.

	deterministic (simple)			deterministic (complex)			non-deterministic		
	$o_{\text{big\_blue}}$	$o_{\text{big\_red}}$	$o_{\text{small\_blue}}$	$o_{\text{big\_blue}}$	$o_{\text{big\_red}}$	$o_{\text{small\_blue}}$	$o_{\text{big\_blue}}$	$o_{\text{big\_red}}$	$o_{\text{small\_blue}}$
big	.5	.5	0	.5	.5	0	.44	.44	.11
small	0	0	1	0	0	**1**	.17	.17	**.67**
blue	.5	0	.5	.5	0	.5	.50	.01	.50
red	0	1	0	0	1	0	.01	.99	.01
big blue	NA	NA	NA	1	0	0	.79	.01	.20
big red	NA	NA	NA	0	1	0	.01	.99	.00
small blue	NA	NA	NA	0	0	**1**	.20	.00	**.80**

that variability under the basic model is to assume that the costs of utterances vary from task to task. This also seems to us an implausible assumption. Thus we must look elsewhere to account for overinformativeness. We propose that the place to look is the computation of informativeness itself.

2.2 RSA with continuous semantics – emergent color-size asymmetry

Here we introduce the crucial innovation: rather than assuming a deterministic truth-conditional semantics that returns true (1) or false (0) for any combination of expression and object, we relax to a continuous semantics that returns real values in the interval $[0, 1]$. Formally, the only change is in the values that the lexicon can return:

$$\mathcal{L}(u, o) \in [0, 1] \subset \mathbb{R} \tag{6}$$

That is, rather than assuming that an object is unambiguously big (or not) or unambiguously blue (or not), this continuous semantics captures that objects count as big or blue to varying degrees (similar to approaches in fuzzy logic and prototype theory, Zadeh, 1965; Rosch, 1973).

An alternative, but equivalent, interpretation of the continuous semantics is as a form of noise or non-determinism in an otherwise deterministic truth-conditional semantics. In our working example below, we take the noisy truth-conditional perspective. [ndg: showing this is kind of subtle... seems like we should put a brief argument here?] [jd: ndg can you do this? not sure abt the amount of detail necessary here. i wonder if already talking in depth about the connection before going through an example will be helpful or distracting? another option is to not say anything about the connection here and instead put it in the discussion?]

The semantic value of utterance-object combinations can vary in complex ways, to be determined by lexical and compositional semantic rules. To see the basic effect of switching to a continuous semantics, and to see how far we can get in capturing overinformativeness patterns with this change, let us explore a simple semantic theory in which all colors are treated the same, all sizes are as well, and the two compose via a product rule. That is, when a size adjective would be 'true' of an object under a deterministic semantics, we take $\mathcal{L}(u, o) = x_{\text{size}}$, a constant; when it is 'false' of the object, $\mathcal{L}(u, o) = 1 - x_{\text{size}}$. Similarly for color adjectives. This results in two free model parameters, x_{size} and x_{color}, that can take on different values, capturing that size and color adjectives may apply more or less well/reliably to objects. Together with the product composition rule, Eq. 5, this fully specifies a relaxed semantic function for our reference domain.

Now consider the RSA literal listener, Eq. 1, who uses these relaxed semantic values. Given an utterance, the listener simply normalizes over potential referents. As an example, the resulting renormalized literal listener distributions for the size-sufficient example context in Figure 1a are shown for values $x_{\text{size}} = .8$ and $x_{\text{color}} = .99$[10] on the right in Table 2. Recall that in this context, the speaker intends for the listener to select the small blue pin. To see which would be the best utterance to produce for this purpose, we compare the literal listener probabilities in the $o_{\text{small\_blue}}$ column. The two best utterances under both the deterministic and the continuous semantics are bolded in the table: under the deterministic semantics, the two best utterances are *small* and *small blue*, with no difference in listener probability. In contrast, under the continuous semantics *small* has a smaller literal listener probability (.67) of retrieving the intended referent than the redundant *small blue* (.80). Consequently, the pragmatic speaker will be more likely to produce *small blue* than *small*, though the precise probabilities depend on the cost and informativeness parameters β_c and β_i.

Crucially, the reverse is not the case when color is the distinguishing dimension. Imagine the speaker in the same context wanted to communicate the big red pin. The two best utterances for this purpose are *red* (.99) and *big red* (.99). In contrast to the results for the small blue pin, these utterances do not differ in their capacity to direct the literal listener to the intended referent. The reason for this is that we defined color to be almost noiseless, with the result that the literal listener distributions in response to utterances containing color terms are more similar to those obtained via a deterministic semantics than the distributions obtained in response to utterances containing size terms. The reader is encouraged to verify this by comparing the row-wise distributions under the deterministic and continuous semantics in Table 2.

To gain a wider understanding of the effects of assuming continuous meanings in contexts like that depicted in Figure 1a, we visualize the results of varying x_{size} and x_{color} in Figure 4. To orient the reader to the graph: the deterministic semantics of utterances is approximated where the semantic values of both size and color utterances are close to 1 (.999, top right-most point in graph). In this case, the simple sufficient (*small pin*) and complex redundant utterance (*small blue pin*) are equally likely around .5, because they are both equally informative and utterances are assumed to have 0 cost. All other utterances are highly unlikely. The interesting question is under which circumstances, if any, the standard color-size asymmetry emerges. This is the yellow/orange/red space in the 'small blue' facet, characterized by values of x_{size} that are lower than x_{color}, with high values for x_{color}. That is, redundant utterances are more likely than sufficient utterances when the redundant dimension (in this case color) is less noisy than the sufficient dimension (in this case size) and overall is close to noiseless.

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 (Pechmann, 1989; Gatt et al., 2011). Note also that no difference in adjective *cost* is necessary

[10] These values were chosen for the demonstration because they are the ones that result in the best approximation of the proportion of redundant referring expressions reported in Gatt et al. (2013, 80% in size-sufficient contexts; 8% in color-sufficient contexts).

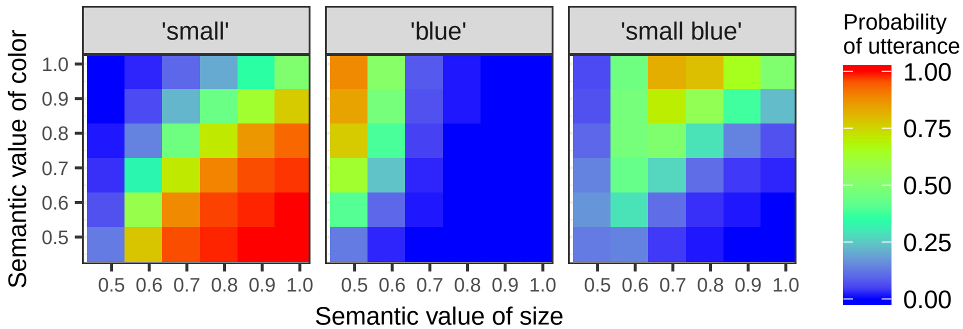


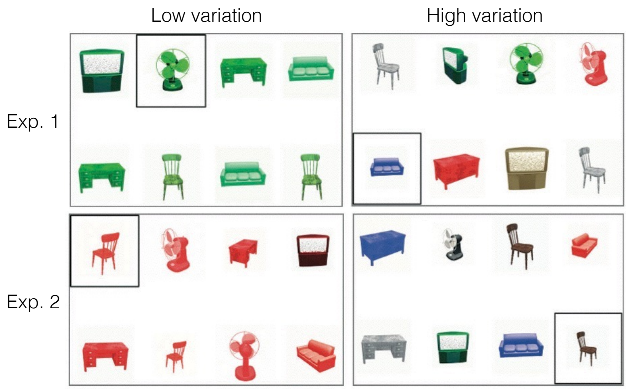
Figure 4: Probability of producing sufficient *small pin*, insufficient *blue pin*, and redundant *small blue pin* in contexts as depicted in Figure 1a, as a function of semantic value of color and size utterances (for $\beta_i = 30$ and $\beta_c = 0$). For a visualization of model behavior under varying αs, see Appendix A.

for obtaining the overinformativeness asymmetry, though 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 for both the costs and the semantic values of size and color, given data from a reference game experiment.

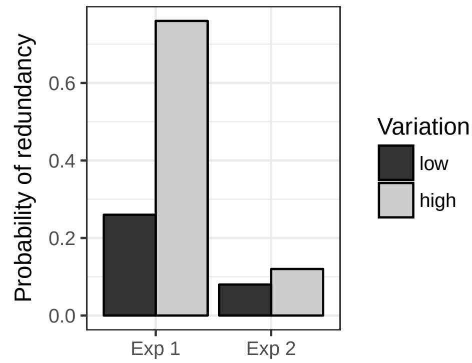
We defer a complete discussion of the important potential psychological and linguistic interpretation of these continuous semantic values to the General Discussion in Section 6. However, it is worth reflecting on why size adjectives may be inherently noisier than color adjectives. Color adjectives are typically treated as *absolute adjectives* while size adjectives are inherently *relative* (Kennedy & McNally, 2005). 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.[11] In addition, color as a property has been claimed to be inherently salient in a way that size is not (Arts et al., 2011; Gatt et al., 2013). Finally, we have shown in recent work that color adjectives are rated as less subjective than size adjectives (Scontras, Degen, & Goodman, 2017). All of these suggest that the use of size adjectives may be more likely to vary across people and contexts than color.

To summarize, we have thus far shown that RSA with continuous adjective semantics can give rise to the well-documented color-size asymmetry in the production of overinformative referring expressions when color adjectives are closer to deterministic truth-functions than size adjectives. The crucial mechanism is this: when modifiers are relaxed, adding additional, 'stricter' modifiers adds information. From this perspective, these redundant modifiers are not *over*informative; they are rationally redundant, or sufficiently informative given the needs of the listener. We spend the remainder of the paper demonstrating the far-reaching effects of assuming relaxed semantic values.

[11]This is not entirely true, as has been repeatedly pointed out (e.g., Cohen & Murphy, 1984): 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.



(a) 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).



(b) Predicted probability of redundant color utterance in Koolen conditions for $\beta_i = 30$, $\beta_c = c(u_{\text{size}}) = c(u_{\text{color}}) = 1$, $x_{\text{size}} = .8$, $x_{\text{color}} = .999$, $x_{\text{type}} = .9$.

Figure 5: Koolen et al. contexts and RSA model predictions.

2.3 RSA with continuous semantics – scene variation

As discussed in Section 1, increased scene variation has been shown to increase the probability of referring expressions that are overmodified with color. Here we simulate the experimental conditions reported by Koolen et al. (2013) and explore continuous semantics RSA's predictions for these situations. Koolen et al. (2013) 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).[12] Here, we simulate the high and low variation conditions from their Experiments 1 and 2, reproduced in Figure 5a.

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 5a, respectively). In Exp. 2, size and type were necessary (*big chair* and *small chair* in Figure 5a, 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 semantic value x_{type}, which encodes how noisy nouns are.

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 $\beta_i = 30$, $\beta_c = c(u_{\text{size}}) = c(u_{\text{color}}) = 1$, $x_{\text{size}} = .8$, $x_{\text{color}} = .999$, and $x_{\text{type}} = .9$ are shown in Figure 5b, under the assumption that the cost of a two-word utterance $c(u)$ is the sum of the costs of the one-word utterances:[13][ndg: how were these values chosen? if they we fit to data, say

[12] 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 the sake of simplicity.

[13] These values were chosen to maintain constant the semantic values used in the previous example; include a

so.][jd: is the now included footnote good enough?]

$$c(u) = \beta_{c(\text{color})}\mathbb{1}_{color}(u) + \beta_{c(\text{size})}\mathbb{1}_{size}(u) \tag{7}$$

For both experiments, the model retrieves the empirically observed qualitative effect of variation on the probability of redundant color mention: when variation is greater, redundant color mention is more likely. Indeed, the qualitative empirical scene variation effect is predicted by the model anytime the semantic values for size, type, and color are ordered as: $x_{\text{size}} \leq x_{\text{type}} < x_{\text{color}}$. If x_{type} is greater than x_{color}, 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 adds, information about the target.

RSA with a continuous semantics thus captures the qualitative effects of color-size asymmetry and scene variation in production of redundant expressions. A quantitative test of the model will require larger data sets. In Sections 3, 4, and 5 we quantitatively evaluate continuous semantics RSA on datasets capturing the phenomena described in the Introduction (for a summary see Table 1): modifier type and scene variation effects on modified referring expressions, typicality effects on color mention, and the choice of taxonomic level of reference in nominal choice, respectively.

3 Modified referring expressions: size and color modifiers under different scene variation conditions

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 occurring 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 allows us to both a) assess how likely the model is to generate the actually observed data – i.e., to obtain a measure of model quality – and b) infer the posterior probability of parameter values – i.e., to understand whether the assumed asymmetries in the adjectives' semantic values 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 continuous semantics RSA correctly predicts qualitative 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}:

constant but non-zero cost of utterances (so two-word utterances are dispreferred over one-word utterances); and a relatively high β_i that makes the probabilities more extreme than they would otherwise be, to demonstrate the qualitative pattern of redundancy.

$$\text{scene variation} = \frac{n_{\text{diff}}}{n_{\text{total}}}$$

To explain, let's turn again to Figure 1a. Here, the target item is the small blue pin and there are two distractor items: a big blue pin and a big red pin. 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 pin) 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.[14]

Using the same parameter values as in the previous two model explorations ($\beta_i = 30$, $\beta_c = c(u_{\text{size}}) = c(u_{\text{color}}) = 1$, $x_{\text{size}} = .8$, $x_{\text{color}} = .999$), 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 model predictions are shown in Figure 6: the probability of redundant adjective use increases with increasing scene variation when size is sufficient (and color redundant), but not when color is sufficient (and size redundant). 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 more and more. 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. Note that this is highly dependent on the actual semantic value of color, with slightly lower semantic values for color, the model predicts small increases in redundant size use. This will be important for the interpretation of the empirical results. In general: increased scene variation is predicted to lead to a greater increase in redundant adjective use for less noisy adjectives.

To test continuous semantics RSA predictions, we conducted an interactive web-based written production study within a reference game setting. Speakers and listeners were shown arrays of objects that varied 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

[14] Some readers might find this unintuitive: shouldn't scene variation be maximal when there is an equal number of same and different colors? Or when the different colors are also all different from one another? As discussed in the Introduction, there are many ways of quantifying (different aspects of) scene variation. Here we explore just one such measure; it is an interesting question whether RSA accounts equally well for different ways of quantifying scene variation. Fortunately, it is very straightforward to implement such different measures by manipulating features of distractor items and exploring the model's behavior in these contexts.

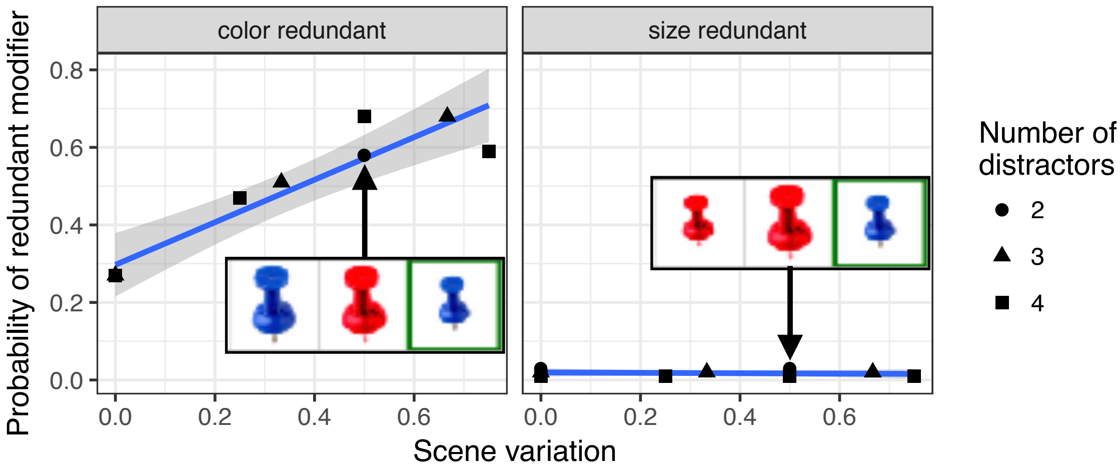


Figure 6: Probability of redundant utterance (*small blue pin*) as a function of scene variation when size is sufficient (and color redundant, left) and when color is sufficient (and size redundant, right), for $\beta_i = 30$, $\beta_c = c(u_{\text{size}}) = c(u_{\text{color}}) = 1$, $x_{\text{size}} = .8$, $x_{\text{color}} = .999$. Linear smoothers overlaid.

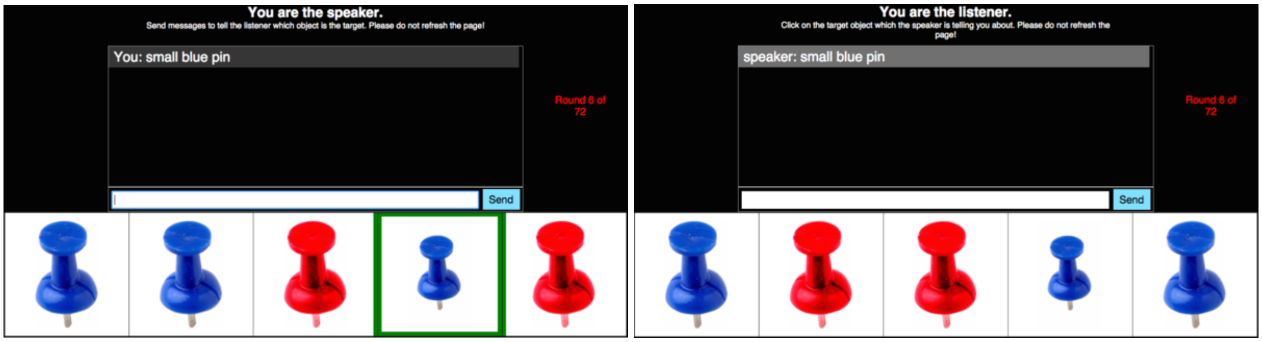
work, were also included. Here and in all other experiments reported in this paper, participants' IP address was limited to US addresses and only participants with a past work approval rate of at least 95% were accepted.

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 (like *left* or *right*) 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 B.

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 into an unrestricted 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 on 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. That is, both participants received feedback about the speaker's intended referent and the listener's inference.

Both speakers and listeners could write in the chat window, allowing listeners to request clari-



(a) Speaker's perspective. (b) Listener's perspective.

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

fication if necessary. Listeners could only click on an object and advance to the next trial once the speaker 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 insufficient feature with the target.

Objects varied in color and size. On 18 trials, color was sufficient for establishing reference. On the other 18 trials, size was sufficient. Figure 7 shows an example of a size-sufficient trial. 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 redundant feature value with the target. That is, when size was sufficient, 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 redundant property would have been sufficient for uniquely establishing reference, i.e. mentioning it would not have been redundant. Each total number of distractors was crossed with each possible number of distractors that shared the redundant 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 pins 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 C.

Fillers were target trials from Exp. 2, a replication of Graf, Degen, Hawkins, and Goodman (2016). 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 5 for a description of these materials.

3.1.2 Data pre-processing and exclusion

We collected data from 2177 critical trials. Because we did not restrict participants' utterances in any way, they produced many different kinds of referring expressions. Testing the model's predictions required, for each trial, classifying the produced utterance as an instance of a *color*-only mention, a *size*-only mention, or a *color-and-size* mention (or excluding the trial if no classification was possible). To this end we conducted the following semi-automatic data pre-processing.

An R script first 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 covered another 1.9% of trials. Most of the time, participants converged on a convention of producing only the target's size and/or color, e.g., *purple* or *big blue*, but not an article (e.g., *the*) or the noun corresponding to the object's type (e.g., *comb*). Articles were omitted in 88.6 % of cases and nouns were omitted in 71.6 % 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,[15] these cases were excluded from the analysis. There were 3 cases that were nonsensical, e.g. *bigger off a shade*, which were also excluded. In 6 cases 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. Finally, we excluded six trials where the speaker did not produce any utterances, and 33 trials on which the listener selected the wrong referent, leading to the elimination of 1.5% of trials. After the exclusion, 2076 cases classified as one of *color*, *size*, or *color-and-size* entered the analysis. After the exclusion, 2076 cases classified as one of *color*, *size*, or *color-and-size* entered 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 predictions (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 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 semantic values for color and size adjectives?

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.[16] The model included

[15]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.

[16]All mixed effects analyses reported in this paper were conducted with the `lme4` package (Bates, Mächler, Bolker,

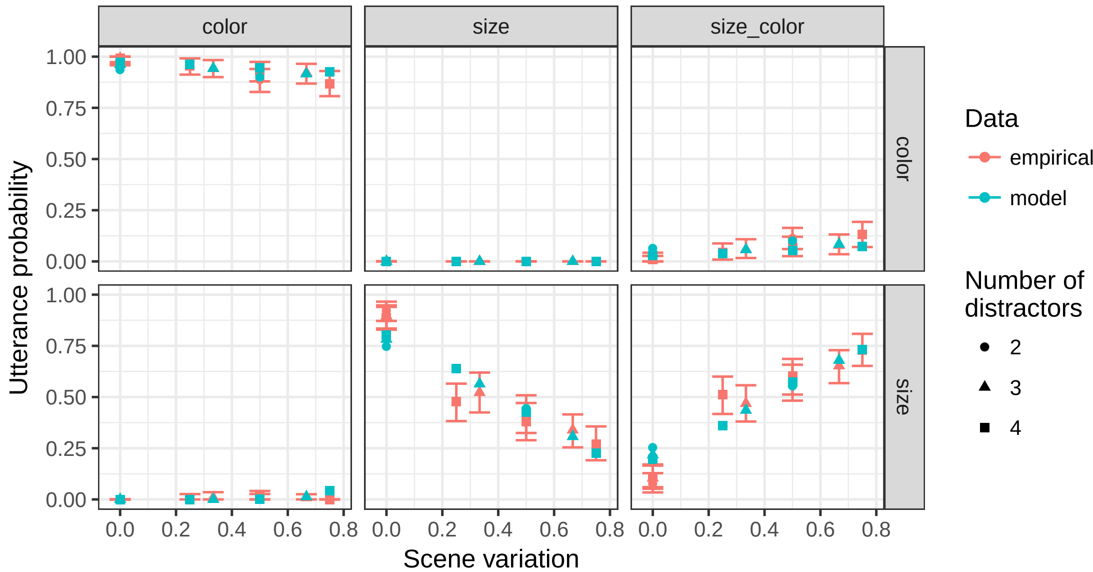


Figure 8: Empirical utterance proportions (red) alongside point-wise maximum a posteriori (MAP) estimates of the RSA model's posterior predictives for utterance probability (blue) as a function of scene variation. Rows indicate the sufficient dimension, columns the produced utterance. Here and in all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

the maximal random effects structure that allowed the model to converge: by-speaker and by-item random intercepts.[17]

We observed a main effect of sufficient property, such that speakers were more likely to redundantly use color than size adjectives ($\beta = 3.54$, $SE = .22$, $p < .0001$), 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 ($\beta = 4.62$, $SE = .38$, $p < .0001$). Finally, we also observed a significant interaction between sufficient property and scene variation ($\beta = 2.26$, $SE = .74$, $p < .003$). Simple effects analysis revealed that the interaction was driven by the scene variation effect being smaller in the *color-sufficient* condition ($\beta = 3.49$, $SE = .65$, $p < .0001$) than in the *size-sufficient* condition ($\beta = 5.75$, $SE = .38$, $p < .0001$), as predicted if size modifiers are noisier than color modifiers.

3.2 Model evaluation

In order to evaluate RSA with non-deterministic truth functions, we asked how well it captures the empirical data. To this end we conducted a Bayesian data analysis. This allowed us to

& Walker, 2015) in R (R Core Team, 2017).

[17] In order to address convergence issues with `lmer` when specifying the full random effects structure – i.e., by-speaker and by-item random intercepts and slopes for all fixed effects and their interactions – we ran a Bayesian binomial mixed effects model with weakly informative priors using the **brms** package (Bürkner, 2017) that included the same fixed effects structure as the lmer model and the full random effects structure. The results were qualitatively identical, yielding evidence for main effects of redundant feature (posterior mean $\beta = 5.91$, 95% CI = [4.15,8.10], $p(\beta > 0) = .98$), scene variation (posterior mean $\beta = 6.18$, 95% CI = [4.30,8.24], $p(\beta > 0) = 1$), and their interaction (posterior mean $\beta = 3.31$, 95% CI = [-0.54,7.23], $p(\beta > 0) = .96$).

simultaneously generate model predictions and infer likely parameter values, by conditioning on the observed production data (coded into *size*, *color*, and *size-and-color* utterances as described above) and integrating over the five free parameters. To allow for differential costs for size and color, we introduce separate cost weights ($\beta_{c(\text{size})}, \beta_{c(\text{color})}$) applying to size and color mentions, respectively, in addition to semantic values for color and size (x_{color}, x_{size}) and an informativeness parameter β_i. We assumed uniform priors for each parameter: $x_{\text{color}}, x_{\text{size}} \sim \mathcal{U}(0,1)$, $\beta_{c(\text{size})}, \beta_{c(\text{color})} \sim \mathcal{U}(0,40)$, $\beta_i \sim \mathcal{U}(0,40)$. Inference for the cognitive model was exact. We used Markov Chain Monte Carlo (MCMC) with a burn-in of 10000 and lag of 10 to draw 2000 samples from the joint posteriors on the five 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 8. 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.

Posteriors over parameters are shown in Figure 9. Crucially, the semantic value of color is inferred to be higher than that of size – there is no overlap between the 95% highest density intervals (HDIs) for the two parameters. That is, size modifiers are inferred to be noisier than color modifiers. The high inferred β_i (MAP $\beta_i = 31.4$, HDI = [30.7,34.5]) suggests that this difference in semantic value contributes substantially to the observed color-size asymmetries in redundant adjective use and that speakers are maximizing quite strongly. As for cost, there is a lot of overlap in the inferred weights of size and color modifiers, which are both skewed very close to zero, suggesting that a cost difference (or indeed any cost at all) is neither necessary to obtain the color-size asymmetry and the scene variation effects, nor justified by the data. Recall further that we already showed in Section 2.2 that the color-size asymmetry in redundant adjective use requires an asymmetry in semantic value and cannot be reduced to cost differences. An asymmetry in cost only serves to further enhance the asymmetry brought about by the asymmetry in semantic value, but cannot carry the redundant use asymmetry on its own.

3.3 Discussion

In this section, we reported the results of a dataset of freely collected referring expressions that replicated the well-documented color-size asymmetry in redundant adjective, the effect of scene variation on redundant color use, and showed a novel effect of scene variation on redundant size use. We also showed that continuous semantics RSA provides an excellent fit to these data. In particular, the crucial element in obtaining the color-size asymmetry in overmodification is that size adjectives be noisier than color adjectives, captured in RSA via a lower semantic value for size compared to color. The effect is that color adjectives are more informative than size adjectives when controlling for the number of distractors each would rule out under a deterministic semantics. Asymmetries in the cost of the adjectives only serve to further enhance the modification asymmetry resulting from the asymmetry in semantic value. In addition, we showed that asymmetric effects of scene variation on overmodification straightforwardly fall out of continuous semantics RSA: scene variation leads to a greater increase in overmodification with less noisy than with more noisy modifiers because the less noisy modifiers (colors) on average provide more information about the target.

These results raise interesting questions regarding the status of the inferred semantic values:

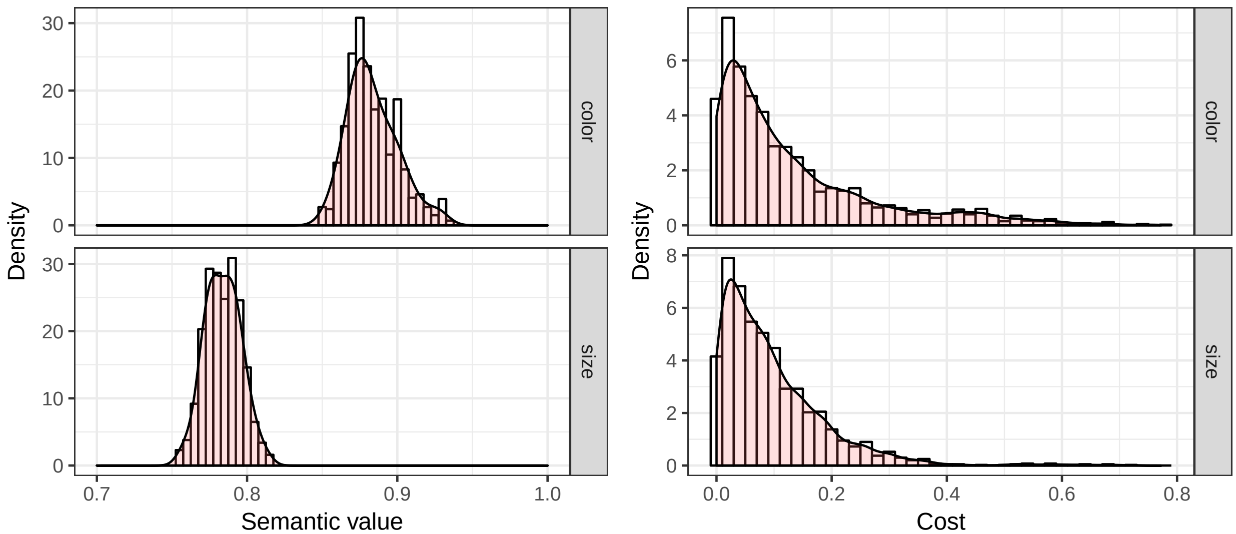


Figure 9: Posterior model parameter distributions for semantic value (left column) and cost (right column), separately for color (top row) and size (bottom row) modifiers. Maximum a posteriori (MAP) $x_{\text{size}} = 0.79$, 95% highest density interval (HDI) = [0.76,0.80]; MAP $x_{\text{color}} = 0.88$, HDI = [0.85,0.92]; MAP $\beta_{c(\text{size})} = .02$, HDI = [0, 0.26]; MAP $\beta_{c(\text{color})} = 0.03$, HDI = [0,0.45].

do color modifiers have inherently higher semantic values than size modifiers? Is the difference constant? What if the color modifier is a less well known one like *mauve*? The way we have formulated the model thus far, there would indeed be no difference in semantic value between *red* and *mauve*. Moreover, the model is not equipped to handle potential object-level idiosyncracies such as the typicality effects discussed in Section 1.2.3. We defer a fuller discussion of the status of the semantic value term to the General Discussion (Section 6.4) and turn first to continuous semantics RSA's potential for capturing these typicality effects.

[jd: will have to discuss here already? not sure how to do that with all the typicality stuff not yet introduced, and i'm actually kinda happy with how the section 6.4 discussion on this issue turned out]

4 Modified referring expressions: color typicality

In Section 3 we showed that continuous semantics RSA successfully captures both the basic asymmetry in overmodification with color vs. size as well as effects of scene variation on overmodification. 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). Continuous semantics 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 semantic value; 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

(a) Typical color. (b) Mid-typical color. (c) Atypical color.

Figure 10: Three hypothetical contexts where color is redundant for referring to the target banana. Banana varies in typicality from left to right. Each context contains one distractor of the same color as the target, and one of a different color.

Table 3: Hypothetical semantic values for utterances (rows) as applied to objects (columns).

	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

to get overmodification behavior from the typicality of features compositionally. This is a problem for all theories of modification (Kamp & Partee, 1995). 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 semantic value 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 the three contexts in Figure 10. The target object in each is the banana, which varies in how typical its color is. The banana is the only object of its type, making type mention (*banana*) sufficient for unique reference and color redundant. Additionally, each context contains a distractor of the same color as the target, further making color redundant, as well as a distractor of a different color. Assume further the hypothetical semantic values shown in Table 3. 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 semantic value of .9. In contrast, brown bananas are less typical instances of *bananas* (.35), and blue bananas are highly atypical *bananas* (.1) but still better than objects of an other non-banana type (.015). Going along the diagonal, we assume for each remaining utterance that its semantic value is very high (.99) when applied to an object in its (deterministic truth-conditional) extension and very low otherwise (.015).

Inputting these contexts and semantic values into the RSA model used thus far, with $\beta_i = 12$ 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, to refer to the yellow banana, the brown banana, and the blue banana, respectively. 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 semantic value of the simple *banana* utterance.

So far we have shown that continuous semantics RSA can capture typicality effects in principle if we assume that semantic values do not operate at the adjective type level but instead captures the typicality of an object for the (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.

This example is somewhat oversimplified. In practice, speakers sometimes mention an object's color without mentioning the noun. In the contexts presented in Figure 10 this does not make much sense because there is always a competitor of the same color present. In contrast, in the contexts in Figure 12a and Figure 12c, color alone disambiguates the target. This suggests that we should consider among the set of utterance alternatives not just the simple type mentions (e.g., *banana*) and color-and-type mentions (e.g., *yellow banana*), but also simple color mentions (e.g., *yellow*). The dynamics of the model proceed as before.

We can now ask whether taking into account this more fine-grained notion of a continuous semantics affects the probability of redundantly mentioning color. Because the stimuli for Exp. 1 were specifically designed to be realistic objects with low color-diagnosticity, they do not include objects with low typicality values or large degrees of variation in typicality. This makes the dataset from Exp. 1 not well-suited for investigating typicality effects.[18] We therefore conducted a separate production experiment in the same paradigm but with two broad changes: first, objects' color varied in typicality; and second, we did not manipulate object size, focusing only on color mention. This allows us to ask three questions: first, do we replicate the typicality effects reported in the literature – that is, are less color-typical objects more likely to lead to redundant color use than more color-typical objects? Second, does RSA with empirically elicited typicality values as proxy for a continuous semantics capture speakers' behavior? Third, does the semantic value depend only on typicality, or is there still a role for modifier type level noise of the kind we investigated in the previous section? In addition, we can investigate the extent to which utterance cost, which we found not to play a role in the previous section, affects the choice of referring expression.

4.1 Experiment 2: color typicality effects

4.1.1 Method

Participants We recruited 61 pairs of participants (122 participants total) over Amazon's Mechanical Turk who were each paid $1.80 for their participation.

Procedure The procedure was identical to that of Exp. 1. See Figure 11 for an example speaker and listener perspective.

Materials Each participant completed 42 trials. In this experiment, there were no filler trials, since pilot studies with and without fillers delivered very similar results. Each array presented to the participants consisted of three objects that could differ in type and color. One of the three objects functioned as a target and the other two as its distractors.

[18] We did elicit typicality norms for the items in Exp. 1 and replicated the previously documented typicality effects on the four items that did exhibit variation in typicality. See Appendix D for details.

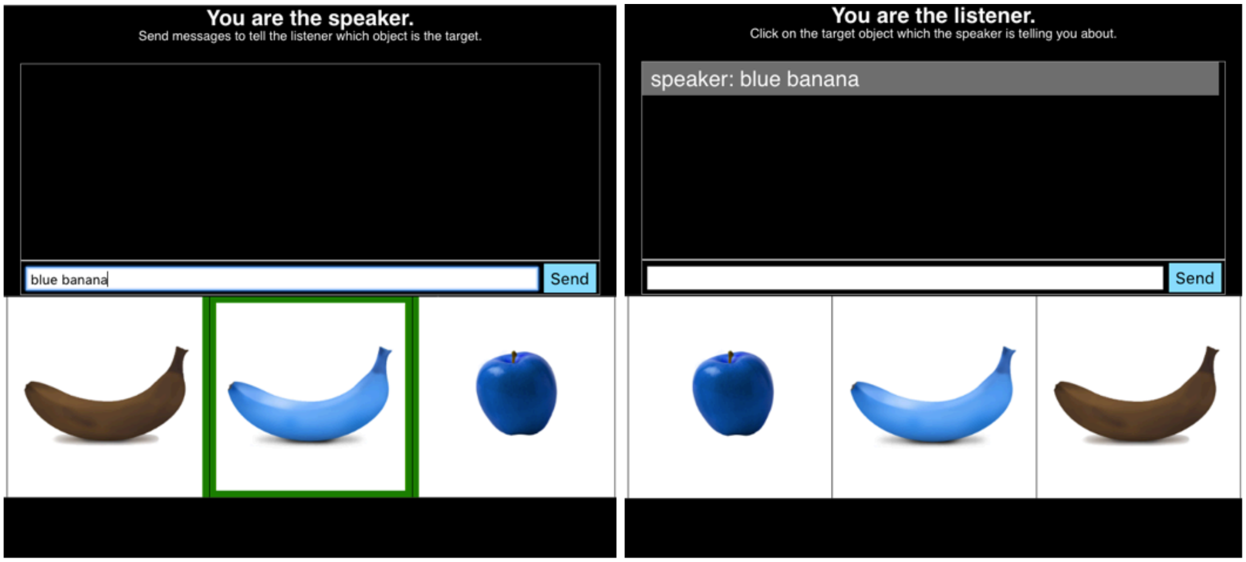


Figure 11: Example displays from the speaker's (left) and listener's (right) perspective in an informative-cc (i.e., presence of another object of the same type and one with the same color) condition.

The stimuli were selected from seven color-diagnostic food items (apple, avocado, banana, carrot, pear, pepper, tomato), which all occurred in a typical, mid-typical and atypical color for that object. For example, the banana appeared in the colors yellow (typical), brown (midtypical), and blue (atypical). All items were presented as targets and as distractors. Pepper additionally occurred in a fourth color, which only functioned as a distractor due to the need for a green color competitor (as explained in the following paragraph).

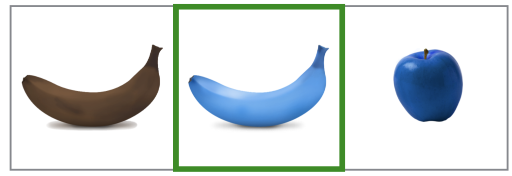
We refer to the different context conditions as "informative", "informative-cc", "overinformative", and "overinformative-cc" (see Figure 12). A context was "overinformative" (Figure 12c) when mentioning the type of the item, e.g., banana, was sufficient for unambiguously identifying the target. In this condition, the target never had a color competitor. This means that mentioning color alone (without a noun) was also unambiguously identifying. In contrast, in the overinformative condition with a color competitor ("overinformative-cc", Figure 12d), color alone was not sufficient. In the informative conditions, color and type mention were necessary for unambiguous reference. Again, one context type did (Figure 12a) and one did not (Figure 12d) include a color competitor among its distractors.

Each participant saw 42 different contexts. Each of the 21 items (color-type combinations) was the target exactly twice, but the context in which they occurred was drawn randomly from the four possible conditions mentioned above. In total, there were 84 different possible configurations (seven target food items, each of them in three colors, where each could occur in four contexts). Trial order was randomized.

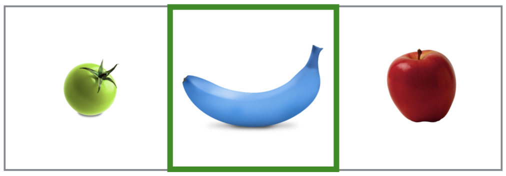
4.1.2 Data pre-processing and exclusion

Two participant-pairs were excluded because they did not finish the experiment and therefore could not receive payment. Trials on which the speaker did not produce any utterances were also excluded,

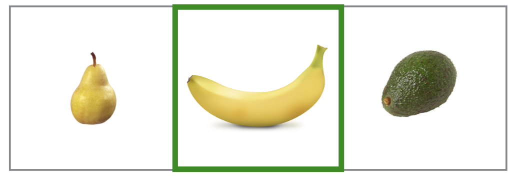
(a) informative (without color competitor)



(b) informative-cc (with color competitor)



(c) overinformative (without color competitor)



(d) overinformative-cc (with color competitor)

Figure 12: Examples of the four different context conditions in Exp. 2. They differed in the presence of an object of the same type (informative vs. overinformative) and in the presence of another object of the same color as the target (with color competitor vs. without color competitor). The green border marks the intended referent.

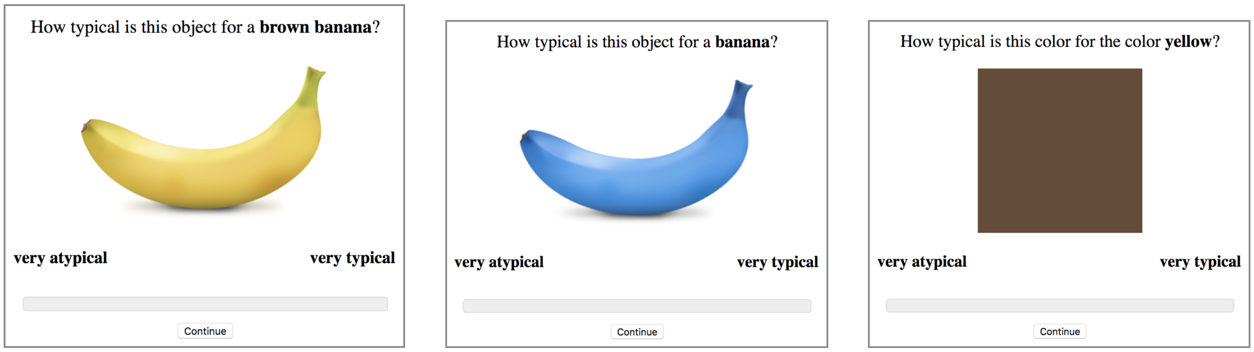
resulting in the exclusion of two additional participant-pairs. Finally, there were 10 speakers who consistently used roundabout descriptions instead of direct referring expressions (e.g., *monkeys love...* to refer to banana). These pairs were also excluded.

We analyzed data from 1974 trials. Just as in Exp. 1, participants communicated freely, which led to a vast amount of different referring expressions. To test the model's predictions, the utterance produced for each trial was classified as belonging to one of the following categories: *type-only* ("banana"), *color-and-type* ("yellow banana"), and *color-only* ("yellow") utterances. Referring expressions that included categories ("yellow fruit"), descriptions ("has green stem"), color-circumscriptions ("funky carrot"), and negations ("yellow but not banana") were regarded as *other* and excluded. To this end we conducted the following semi-automatic data pre-processing.

The referring expressions were analyzed similarly to Exp. 1. First, 32 trials (1.6%) were excluded because the listener selected the wrong referent. 109 trials (5.6%) were excluded because the referring expressions included one of the exceptional cases described above (e.g., using negations). An R script then automatically checked the remaining 1833 utterances for whether they contained a precoded color term (i.e. *green, purple, white, black, brown, yellow, orange, blue, pink, red, grey*) or type (i.e. *apple, banana, carrot, tomato, pear, pepper, avocado*). This way, 96.5% of the remaining cases were classified as mentioning type and/or color.

However, this did not capture that sometimes, participants produced meaning-equivalent modifications of color/type terms for instance by adding suffixes (e.g., *pinkish*), using abbreviations (e.g., *yel* for *yellow*), or using non-precoded color and type labels (e.g., *lavender* or *jalapeno*). In addition, expressions that contained a typo (e.g., *blakc* instead of *black*) could also not be classified automatically. One of the authors (EK) therefore manually hand-coded these cases.

There were 6 cases (0.3%) that could not be categorized. Those were mostly greetings (e.g.,*Hi*), other comments (e.g., *I have instructions to follow sometimes*) and not certainly identifiable utterances (e.g., *re*). These were excluded. After exclusion, 1827 cases classified as one of *color*, *type*, or *color-and-type* entered the analysis.



(a) color-and-type norming. (b) type-only norming. (c) color-only norming.

Figure 13: Example stimuli exemplifying the three different typicality norming studies.

Utterances	Example	Images	Participants	Trials	Items	Excluded Participants
Adj Noun	*yellow banana*	object	174	110	484	14
Noun	*banana*	object	75	90	154 (198)	1
Adj	*yellow*	color patch	110	90	176 (198)	None

Table 4: Overview of the typicality norming studies for Exp. 2. Items column contains the number of unique utterance-object pairs that we elicited responses for. The value in parentheses represents the number of items including utterances *fruit*, *vegetable*, *cup* and *pink*, which were not analyzed further.

4.1.3 Typicality norming

In order to test for typicality effects on the production data and to evaluate RSA's performance, we collected empirical typicality values for each utterance/object pair (as proxy for values of the sort listed in Table 3) across three separate studies. The first study collected typicalities for *color-and-type*/object pairs (e.g., *yellow banana* as applied to a yellow banana, a blue banana, an orange pear, etc., see Figure 13a). The second study collected typicalities for *type-only*/object pairs (e.g., *banana* as applied to a yellow banana, a blue banana, an orange pear, etc., Figure 13b). The third study collected typicalities for *color*/color pairs (e.g., *yellow* as applied to a color patch of the average yellow from the yellow banana stimulus or to a color patch of the average orange from the orange pear stimulus, and so on, for all other colors, Figure 13c).

On each trial, participants saw one of the stimuli used in the production experiment in isolation and were asked: "How typical is this object for a *utterance*", where *utterance* was replaced by an utterance of interest. In the color typicality study, they were asked "How typical is this color for the color *color*?", where *color* was replaced by one of the relevant color terms. They then adjusted a continuous sliding scale with endpoints labeled "very atypical" and "very typical" to indicate their response. An overview of the differences between the three typicality norming studies is shown in Table 4.

Slider values were coded as falling between 0 ('very atypical') and 1 ('very typical'). For each utterance-object combination, we computed mean typicality ratings. The means for the banana items are shown in Table 5. The values are very similar to those hypothesized for the purpose of

Table 5: Mean typicalities for banana items. Combinations where deterministic semantics would return TRUE are marked in boldface.

	Banana items			Other
Utterance	yellow	brown	blue	
banana	**.98**	**.66**	**.42**	.05
yellow banana	**.97**	.30	.15	.05
brown banana	.22	**.91**	.15	.04
blue banana	.16	.15	**.92**	.06

the example in Table 3. The means for all items are visualized in Figure 14.

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, rather than as the degree to which an object's color is representative of that object. 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 distributions of features in objects is beyond the scope of this paper. We expect, however, that the simple Type-object typicalities most closely approximates the Westerbeek question because the employed objects are color-diagnostic – asking whether a blue or a yellow banana is a typical *banana* is similar to asking whether or not the bananas' most salient property – their color – is typical.[19]

Mean typicality values obtained in the norming studies are used in the analyses and visualizations in the following.

4.1.4 Results and discussion

Proportions of type-only (*banana*), color-and-type (*yellow banana*), color-only (*yellow*), and other (*funky carrot*) utterances are shown in Figure 15a as a function of the described item's mean type-only (*banana*) typicality. Visually inspecting just the explicitly marked *yellow banana*, *brown banana*, and *blue banana* cases suggests a large typicality effect in the overinformative conditions as well as a smaller typicality effect in the informative conditions, such that color is less likely to be produced with increasing typicality of the object.

The following questions are of interest. First, do we replicate the previously documented typicality effect on redundant color mention (as suggested by the visual inspection of the banana item)? Second, does typicality affect color mention even when color is informative (i.e., technically necessary for establishing unique reference)? Third, are speakers sensitive to the presence of color competitors in their use of color or are typicality effects immune to the nature of the distractor items?

To address these questions we conducted a mixed effects logistic regression predicting color

[19]See also Appendix D for an independent comparison of our question and the Westerbeek question as applied to typicality norms for the items in Exp. 1. In general, the Type-object values are highly correlated with the Westerbeek question values.

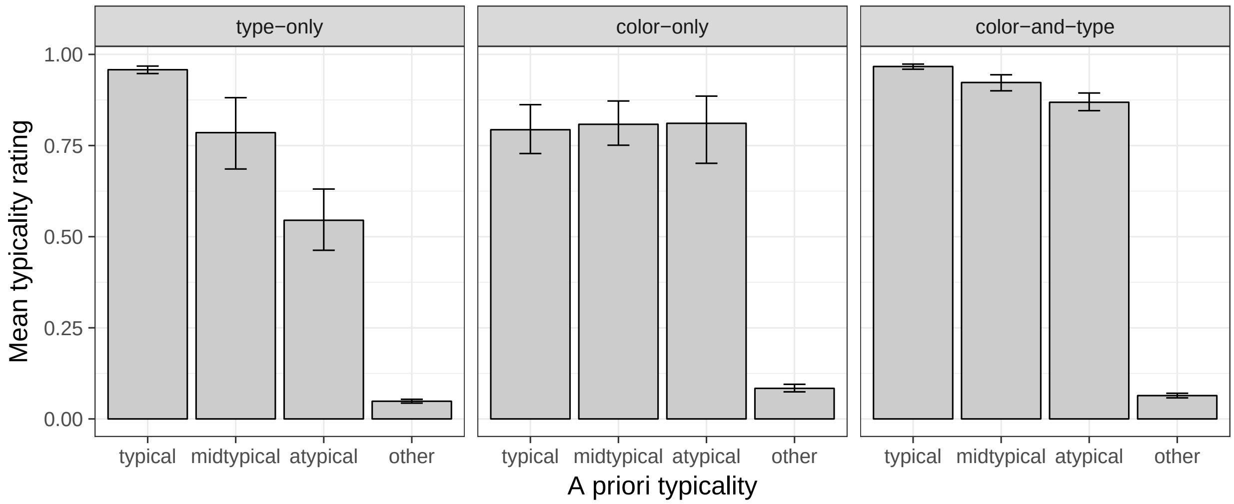
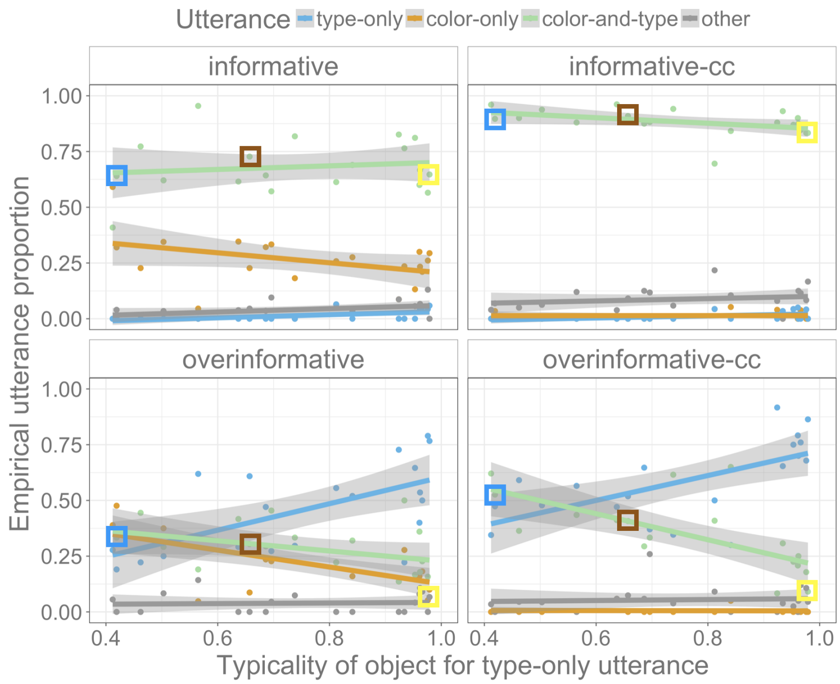


Figure 14: Mean typicality ratings for the three norming studies (type-only, color-only, color-and-type). The results are categorized according to the objects' a priori typicality as determined by the experimenters (yellow banana = typical, brown banana = midtypical, blue banana = atypical). The category *other* comprises all utterance-object combinations where a Boolean semantics would return false. Error bars indicate bootstrapped 95% confidence intervals.

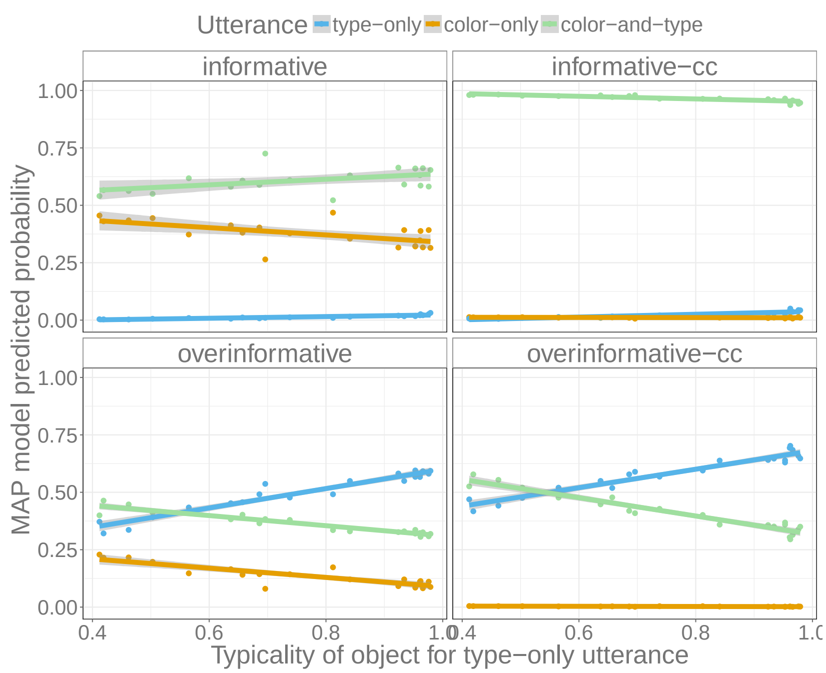
use from fixed effects of typicality, informativeness, and color competitor presence. We used the typicality norms obtained in the *type*/object typicality elicitation study reported above (see Figure 13b) as the continuous typicality predictor. The informativeness condition was coded as a binary variable (color informative vs. color overinformative trial) as was color competitor presence (absent vs. present). All predictors were centered before entering the analysis. The model included by-speaker and by-item random intercepts, which was the maximal random effects structure that allowed the model to converge.

There was a main effect of typicality, such that the more typical an object was for the type-only utterance, the lower the log odds of color mention (β = -4.17, SE = 0.45, $p < .0001$), replicating previously documented typicality effects. Stepwise model comparison revealed that including interaction terms was not justified by the data, suggesting that speakers produce more typical colors less often even when the color is in principle necessary for establishing reference (i.e., in the informative conditions). This is notable: speakers sometimes call a yellow banana simply a *banana* even when other bananas are present, presumably because they can rely on listeners drawing the inference that they must have meant the most typical banana. In contrast, blue bananas' color is always mentioned in the informative conditions.

There was also a main effect of informativeness, such that color mention was less likely when it was overinformative than when it was informative (β = -5.56, SE = 0.33, $p < .0001$). Finally, there was a main effect of color competitor presence, such that color mention was more likely when a color competitor was absent (β = 0.71, SE = 0.16, $p < .0001$). This suggests that speakers are indeed sensitive to the contextual utility of color – color typicality alone does not capture the full set of facts about color mention, as we already saw in Section 3.



(a) Empirical utterance proportions



(b) MAP model predicted utterance probabilities

Figure 15: (a) Empirical utterance proportions in Exp. 2 and (b) MAP model predicted utterance probabilities for each target as a function of mean object typicality for the type-only utterance (e.g., *banana*). Color indicates utterance type: type-only (*banana*), color-only (*yellow*), color-and-type (*yellow banana*), and other (*funky carrot*). Facets indicate conditions. Modified utterance data points for the banana items are circled in the banana's respective color in (a).

4.2 Model evaluation

We evaluated the continuous semantics RSA model on the obtained production data from Exp. 2. In particular, we were interested in using model comparison to address the following issues.

First, we already discussed that at least for the banana item, the empirically elicited typicalities, when interpreted as semantic values, capture qualitative patterns in the data that the simple modifier type noise asymmetry assumed in the previous section cannot capture. Is this more fine-grained treatment of semantic values as capturing an object's typicality as an instance of an utterance enough to account for speakers' choice of expression, or is it necessary to integrate both the empirically elicited values and the utterance type level (color, type) noise values?

Second, we can ask whether or not utterance cost, quantified in two different ways, explains any of the observed production behavior.

While the architecture of the model remained the same as that of the model presented in Section 2.2, we briefly review the minor necessary changes, some of which we already mentioned at the beginning of this section. These changes concerned the lexicon and the cost function.[20] We elaborate on each in turn before turning to the model evaluation.

4.2.1 Lexicon

Whereas for the purpose of evaluating the model in Section 3 we only considered the utterance alternatives *color*, *size*, and *color-size*, collapsing over the precise level of attributes, here we included in the lexicon each possible color adjective, type noun, and combination of the two. This substantially increased the size of the lexicon to 37 unique utterances. For each combination of utterance u and object o that occurred in the experiment, we included a separate semantic value $x_{u,o}$, elicited in the norming experiments described in Section 4.1.3 (rather than inferred as done for Exp. 1, to avoid overfitting).[21] For any given context, we assumed the utterance alternatives that correspond to the individually present features and their combinations. For example, for the context in Figure 12d, the set of utterance alternatives was *yellow, green, pear, banana, avocado, yellow pear, yellow banana*, and *green avocado*.

4.2.2 Semantics

We compared two choices of semantics for the model. In the *empirical semantics* version, the empirically elicited typicality values were directly used as semantic values. In the more complex *fixed plus empirical semantics* version, we introduce an additional parameter interpolating between the empirical typicality values and inferred values for each utterance type as employed in Section 3 (e.g. one value for color terms and another for type terms, which are multiplied when the terms are composed in an utterance). Note that this allows us to perform a nested model comparison, since the first model is a special case of the second.

[20] See Table 7 for an overview of the models reported in the paper.

[21] Ideally, one would like to derive the semantic values for the modified utterances from the semantic values of the individual lexical items. Unfortunately, the compositionality of continuous values is a recognized problem (Kamp & Partee, 1995) and not one we aim to solve here. For our purposes it was therefore sufficient that we had access to the semantic values of the modified utterances elicited experimentally.

4.2.3 Cost function

For the purpose of evaluating the model in Section 3 we inferred two constant costs (one for color and one for size), and found in the Bayesian Data Analysis that the role of cost in explaining the data was minimal at best. Here, we compared two different versions of utterance cost. In the *fixed cost* model we treated cost the same way as in the previous section and included only a color and type level cost, inferred from the data. We then compared this model to an *empirical cost model*, in which we included a more complex cost function. Specifically, we defined utterance cost $c(u)$ as follows:

$$c(u) = \beta_F \cdot p(u) + \beta_L \cdot l(u) \tag{8}$$

Here, $p(u)$ is utterance frequency as estimated from the Google Books corpus (years 1950 to 2008); $l(u)$ is the mean empirical length of the utterance in characters in the production data (e.g., sometimes *yellow* was abbreviated as *yel*, leading to an $l(u)$ smaller than 6); β_F is a weight on frequency; and β_L is a weight on length. Both $p(u)$ and $l(u)$ were normalized to fall into the interval $[0, 1]$. The empirical cost function thus prefers short and frequent utterances (e.g., *blue*) over long and infrequent ones (*turquoise-ish bananaesque thing*). We compared both of these models to a simpler baseline in which utterances were assumed to have no cost.

4.2.4 Model comparison

To evaluate the effect of these choices of semantics and cost, we conducted a full Bayesian model comparison. Specifically, we computed the Bayes Factor for each comparison, a measure quantifying the support for one model over another in terms of the relative likelihood they each assign to the data. As opposed to classical likelihood ratios, which only use the maximum likelihood estimate, the likelihoods in the Bayes Factor integrate over all parameters, thus automatically applying an Occam's Razor correcting for over-fitting. Because it was intractable to analytically compute these integrals for our recursive model, we used Annealed Importance Sampling (AIS), a Monte Carlo algorithm commonly used to approximate these quantities. To ensure high-quality estimates, we took the mean over 100 independent samples for each model, with each chain running for 30,000 steps. The marginal log likelihoods for each model are shown in Table 6. The best performing model used *fixed plus empirical* semantics and did not include a cost term. Despite the greater number of parameters associated with adding the fixed semantics to the empirical semantics, the *fixed plus empirical* semantics models were preferred across the board compared to their empirical-only counterparts ($BF = 3.7 \times 10^{48}$ for fixed costs, $BF = 2.1 \times 10^{60}$ for empirical costs, and $BF = 1.4 \times 10^{71}$ for no cost). In comparison, additional cost-related parameters were not justified, with $BF = 5.7 \times 10^{21}$ for no cost compared to fixed cost and $BF = 2.1 \times 10^{27}$ for compared to empirical cost.

The correlation between empirical utterance proportions and the best model's MAP predictions at the by-item level was $r = .94$. Predictions for the best-performing model are visualized alongside empirical proportions in Figure 15b. The model successfully reproduces the empirically observed typicality effects in all four experimental conditions. However, we note that it diverges somewhat from the empirical data in conditions without color competitors. Here, *color-and-type* utterances are systematically somewhat underpredicted in the informative condition, and systematically somewhat overpredicted in the overinformative condition. The reverse is true for *color-only* utterances. It is worth looking at the posterior over parameters, shown in Figure 16, to understand the pattern. In particular, the utterance type level semantic value of type is inferred to be systematically higher

Table 6: Marginal log likelihood for each model. Best model is in bold. Parentheses indicate number of free parameters.

		Semantics	
		empirical	*fixed plus empirical*
Cost	*empirical*	-1474.6 (4)	-1354.4 (7)
	fixed	-1434.8 (4)	-1321.9 (7)
	none	-1372.9 (2)	**-1209.8 (5)**

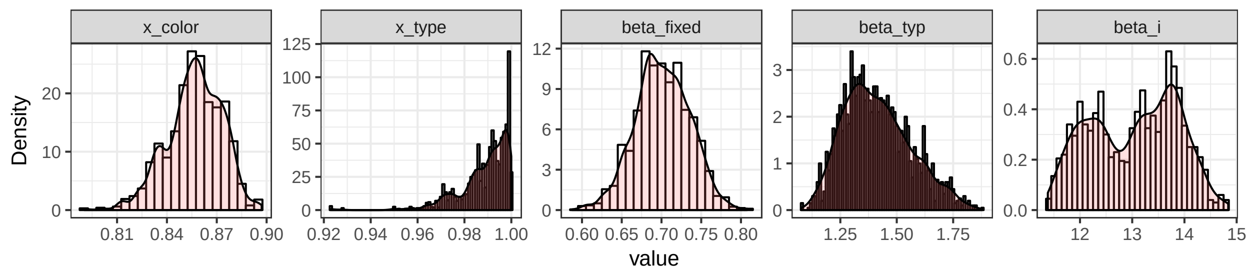


Figure 16: For best model, posterior model parameter distributions for utterance type level semantic values (color, type), interpolation weight on fixed vs. empirical semantics, typicality stretching parameter, and weight on informativity. Maximum a posteriori (MAP) $x_{\text{color}} = 0.86$, 95% highest density interval (HDI) = [0.82,0.89]; MAP $x_{\text{type}} = 0.998$, HDI = [0.97,1.00]; $\beta_{\text{fixed}} = .69$, HDI = [0.64, 0.77]; MAP $\beta_{\text{typ}} = 1.34$, HDI = [1.19,1.75]; MAP $\beta_i = 13.74$, HDI = [11.58,14.37].

than that of color, capturing that type utterances are less noisy than color utterances.[22] An increase in *color-only* mentions in the overinformative condition could be achieved by reducing the semantic value for type. However, that would lead to a further increase in *color-only* mentions in the informative condition as well. That is, the two conditions are in a tug-of-war with each other.

Two further things of note: the interpolation weight between the fixed and empirical semantic values is in the intermediate range: this provides evidence that a noisy truth-conditional semantics as employed in Exp. 1 is justified, but that taking into account graded category membership or typicality in an utterance's final semantic value is also necessary. That the weight on the empirically elicited typicalities is slightly higher than 1 suggests that the empirically elicited typicalities were somewhat less extreme than represented underlyingly.

4.3 Discussion

In this section we demonstrated that the continuous semantics RSA model predicts color typicality effects in the production of referring expressions. The model employed here did not differ in its architecture from that employed in Section 3, but only in that a) semantic values were assumed to operate at the individual utterance/object level in addition to at the utterance type/object level; b) semantic values were empirically elicited via typicality norming studies in addition to being inferred from the data; and c) an utterance's cost was assumed to be a function of its mean empirical length and its corpus frequency as estimated from a large corpus instead of having a

[22]Interestingly, the inferred semantic value for color is very similar in absolute terms to that in Exp. 1.

constant utterance type level value, though utterance cost ultimately was found not to play a role in predicting utterance choice.[23]

This suggests that the dynamics at work in the choice of color vs. size and in the choice of color as a function of the object's color typicality are very similar: speakers choose utterances by considering the fine-grained differences in information about the intended referent communicated by the ultimately chosen utterance compared to its competitor utterances. For noisier utterances (e.g., *banana* as applied to a blue banana), including the 'overinformative' color modifier is useful because it provides information. For less noisy utterances (e.g., *banana* as applied to a yellow banana), including the color modifier is useless because the unmodified utterance is already highly informative with respect to the speaker's intention. These dynamics can sometimes even result in the color modifier being left out altogether, even when there is another object of the same type present that is very atypical, simply because the literal listener asymmetry in probability of choosing the intended referent over the competitor object is big enough.

Crucially, model comparison demonstrated the need for assuming a semantics that interpolates between a noisy truth-conditional semantics as employed in Exp. 1 and empirically elicited semantic values. This suggests that speakers, in choosing utterances, take into account both an utterance's literal semantics and its noisiness, as well as the graded category membership of the objects under consideration. Perhaps surprisingly, we replicated the result from Exp. 1 that utterance cost does not add any predictive power, even when quantified via a more sophisticated cost function that takes into account an utterance's length and frequency.

In the next section, we move beyond the choice of modifier and ask whether continuous semantics RSA provides a good account of content selection in referring expressions more generally. To answer this question we turn to simple nominal referring expressions.

5 Unmodified referring expressions: nominal taxonomic level

In this section we investigate whether continuous semantics RSA accounts for referring expression production beyond the choice of modifier. In particular, we focus on speakers' choice of taxonomic level of reference in nominal referring expressions. A particular object can be referred to at its subordinate (*dalmatian*), basic (*dog*), or superordinate (*animal*) level, among other choices. As discussed in Section 1.3, multiple factors 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, Griffin & Bock, 1998; Jescheniak & Levelt, 1994), and its typicality (an utterance is more likely to be used if the object is a good instance of it, Jolicoeur et al., 1984).

Choice of taxonomic level (e.g., between *dog* and *dalmatian*) is analogous to the choice of modified referring expressions (e.g., between *dog* and *big dog*) in that the two expressions differ in how much information about the intended referent is provided. However, cost-wise, in contrast to the modifier case, the choice between different levels of reference does not result in a choice between a shorter or longer utterance in terms of the number of words required, but it is possible that differences in length and frequency of the different nouns affect the choice. Thus, effects of the same factors we previously tested for in the choice of modifier in the previous two sections can also be tested for their involvement in nominal choice.

[23]See Table 7 for a more extensive overview of the ways in which the models reported across sections differed.

In order to evaluate continuous semantics RSA for nominal choice, we proceeded as in Section 4: we collected production data within the same reference game setting, but varied the contextual informativeness of utterances by varying whether distractors shared the same basic or superordinate category with the target (see Figure 17). We also elicited typicality ratings for object-utterance combinations, which entered the model as the semantic values via the lexicon. We then conducted the same type of Bayesian data analysis as reported in the previous sections for model comparison.

5.1 Experiment 3: taxonomic level of reference in nominal referring expressions

5.1.1 Method

Participants We recruited 58 pairs of participants (116 participants total, the same participants as in Exp. 1) 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.[24] 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 9 in Appendix E 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. These target items were randomly assigned distractor items which were selected from three different context conditions, corresponding to different communicative pressures (see Figure 17). The *subordinate necessary* contexts contained one distractor of the same basic category and one distractor of the same superordinate category (e.g., target: *dalmatian*, distractors: *greyhound* (also a dog) and *squirrel* (also an animal)). The *basic sufficient* contexts contained either two distractors of the same superordinate category but different basic category as the target (e.g., target: *husky*, distractors: *hamster* and *elephant*) or one distractor of the same superordinate category and one unrelated item (e.g., target: *pug*, distractors: *cow* and *table*). The *superordinate sufficient* contexts contained 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 subordinate (*dalmatian*), basic (*dog*) or superordinate (*animal*) level. However, the level of reference necessary for uniquely referring differed across contexts.

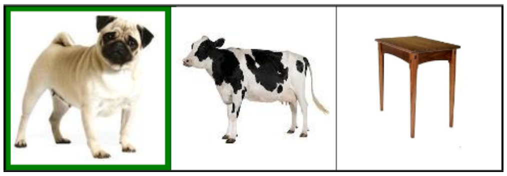
5.1.2 Data pre-processing and exclusion

We collected 2193 referring expressions. To determine the level of reference for each trial, we followed the following procedure. First, speakers' and listeners' messages were parsed automatically; the referring expression used by the speaker was extracted for each trial and checked for whether

[24] Exp. 3 constitutes a replication of the production experiment reported in Graf et al. (2016).

(a) subordinate necessary

(b) basic sufficient (two superordinate distractors)



(c) basic sufficient (one superordinate distractor)

(d) superordinate sufficient

Figure 17: Example contexts in which different levels of reference are necessary for establishing unique reference to the target marked with a green border: (a) subordinate necessary (*dalmatian*); (b, c) basic sufficient (*dog*) and subordinate possible (*husky*, *pug*); (d) superordinate sufficient (*animal*) and basic or subordinate possible (*dog*, *German Shepherd*).

it contained the current target's correct *sub*(ordinate), *basic*, or *super*(ordinate) level term using a simple grep search. In this way, 71.4% of trials were labelled as mentioning a pre-coded level of reference. In the next step, remaining utterances were checked manually by one of the authors (CG) 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 15.0% of trials. 41 trials on which the listener selected the wrong referent were excluded, leading to the elimination of 2.1% of trials. Six trials were excluded because the speaker did not produce any utterances. Additionally, a total of 12.5% 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, 18.9% of basic level terms, and 60.9% 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, 1872 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 Figure 18. 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 (e.g., target: dalmatian, distractor: greyhound; see Figure 17a). Nevertheless, even in these contexts

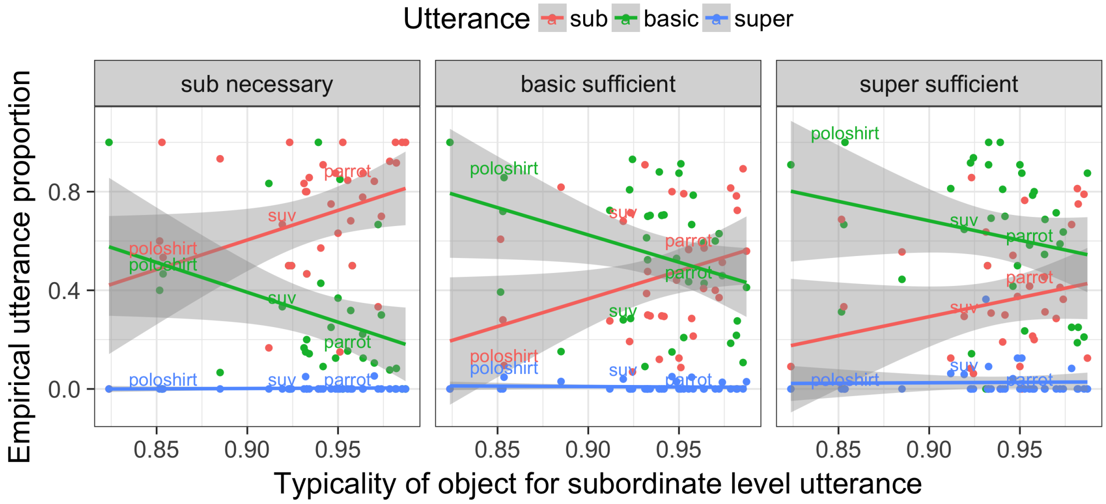


Figure 18: Utterance proportions for each target item across different informativeness conditions as a function of the object's subordinate level typicality. Example target items *polo shirt* (basic: *shirt*, super: *clothes*), *SUV* (basic: *car*, super: *vehicle*), and *parrot* (basic: *bird*, super: *animal*) that were characteristic of relatively low to relatively high sub typicality items are labeled explicitly.

there is a non-negligible proportion of basic level mentions (28%).[25] In the remaining contexts, where the sub and basic level are equally informative, there is a clear preference for the basic level. In addition, mitigating this context effect, sub level mentions increased with increasing typicality of the object as an instance of the sub level utterance.

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).[26]

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

[25] This context is analogous to the informative contexts from Exp. 2, where we also observed some 'underinformative' referring expressions, but note that here this proportion is even higher.

[26] In order to address convergence issues with `lmer` when specifying the full random effects structure – i.e., by-speaker and by-item random intercepts and slopes for all fixed effects – we also ran a Bayesian binomial mixed effects model with weakly informative priors using the `brms` package (Bürkner, 2017) that included the same fixed effects structure as the lmer model and the full random effects structure. The results were qualitatively identical, yielding evidence for main effects of context (sub vs basic sufficient: posterior mean $\beta = 2.44$, 95% CI = [1.87,3.06], $p(\beta > 0) = 1$; basic vs super sufficient: posterior mean $\beta = 0.70$, 95% CI = [0.32,1.09], $p(\beta > 0) = 1$), typicality (posterior mean $\beta = 9.96$, 95% CI = [3.55,17.51], $p(\beta > 0) = 1$), and length (posterior mean $\beta = -1.12$, 95% CI = [-2.00,-0.31], $p(\beta < 0) = 1$).

compared to the basic level, while a higher length ratio reflects a *higher* cost for the sub level term compared to the basic level.[27]

Typicality was coded as the ratio of the target's sub to basic level label typicality.[28] 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 "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 *basic sufficient (two superordinate distractors)* and *basic sufficient (one superordinate distractor)* 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 18.

The log odds of mentioning the sub level term were greater in the *sub necessary* condition than in either of the other two conditions ($\beta = 2.11$, $SE = .17$, $p < .0001$), and greater in the *basic sufficient* condition than in the *super sufficient* condition ($\beta = .60$, $SE = .15$, $p < .0001$), suggesting that the contextual informativeness of the sub level mention has a gradient effect on utterance choice.[29] 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 ($\beta = 4.82$, $SE = 1.35$, $p < .001$). 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", $\beta = -.95$, $SE = .27$, $p < .001$). The main effect of frequency did not reach significance ($\beta = .08$, $SE = .11$, $p < .45$).

Unsurprisingly, there was also significant by-participant and by-domain variation in sub level 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., *plant* in the "flower" domain) to being used more frequently (e.g., *furniture* in the "table" domain and *vehicle* in the "car" domain).

We thus replicated 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, mirroring the results from Exp. 2.

Perhaps surprisingly, we did not observe an effect of frequency on sub level term mention. This is likely due to the modality of the experiment: 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. 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

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

[28] 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.

[29] 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.

effect of frequency. Support for this hypothesis comes from studies comparing written and spoken language, which has found that spoken descriptions are likely to be longer than written descriptions and, in English, seem to have a lower propositional information density than written descriptions (van Miltenburg, Koolen, & Krahmer, 2018).

5.2 Model evaluation

We evaluated the continuous semantics RSA model on the obtained production data from Exp. 3. The architecture of the model is identical to that of the model presented in Section 4.2. The only difference is that the set of alternatives contained only the three potential target utterances (i.e., the target's sub, basic, and super label).[30] Whereas the modifier models from the previous sections treat 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.

For the previous dataset, we tested which of three different semantics was most justified – a fixed semantics with type-level semantic values, the empirically elicited semantics, or a combination of the two. For the current dataset, this question did not arise, because we investigated only one type of utterance – nouns. We hence only considered the *empirical* semantics. However, like in the previous dataset, we evaluated which cost function was best supported by the data: the one defined in (8) (a linear weighted combination of an utterance's length and its frequency) or a simpler baseline in which utterances were assumed to have no cost.

We employed the same procedure as in the previous section to compute the Bayes Factor for the comparison between the two cost models, and to compute the posteriors over parameters $\beta_i \sim \mathcal{U}(0, 20)$, $\beta_F \sim \mathcal{U}(0, 5)$, $\beta_L \sim \mathcal{U}(0, 5)$,, $\beta_t \sim \mathcal{U}(0, 5)$.

Despite the greater number of parameters associated with adding the cost function, the model that includes non-zero costs was preferred compared to its no-cost counterpart ($BF = 2.8 \times 10^{77}$). Posteriors over parameters are shown in Figure 19. It is worth noting that the weight on frequency is close to zero. That is, in line with the results from the mixed effects regression, it is an utterance's length, but not its frequency, that affects the probabilitiy with which it is produced in this paradigm.[31] The posterior on the typicality weight (values < 1) also suggests that, in contrast to Exp. 2, the typicality values elicited empirically are somewhat more extreme than the semantic values that are appropriate for computing informativity in the context of nominal choice.

Empirical utterance proportions are shown against MAP model predictions in Figure 20 ($r = .86$). While the model overpredicts subordinate level and underpredicts basic level choices in the *sub necessary* condition, it generally captures the patterns in the data. The correlation between empirical utterance proportions and the best model's MAP predictions at the by-item level was $r = .86$.

[30] See Table 7 for an overview of the models reported in the paper.

[31] As discussed in previous sections, the lack of importance of a word's frequency may well be attributable to the written modality within which participants generated referring expressions. In the written modality, time pressure on lexical retrieval is lower than in spoken language, so a word's retrievability (which its frequency is an indicator of) is less likely to affect production choices than in the spoken modality; vice versa for an utterance's length – each additional character incurs a typing cost in the written modality.

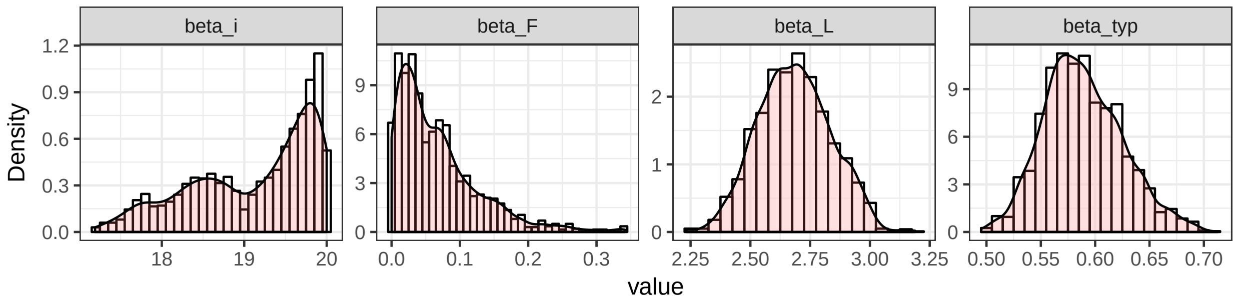


Figure 19: Posterior model parameter distributions for informativity weight (β_i), frequency cost weight (β_F), length cost weight (β_L), and typicality weight (β_t). Maximum a posteriori (MAP) β_i = 19.8, 95% highest density interval (HDI) = [17.71,20.0]; MAP β_F = 0.0.02, HDI = [0.00,0.19]; MAP β_L = 2.69, HDI = [2.42,2.99]; MAP β_t = 0.57, HDI = [0.53,0.67].

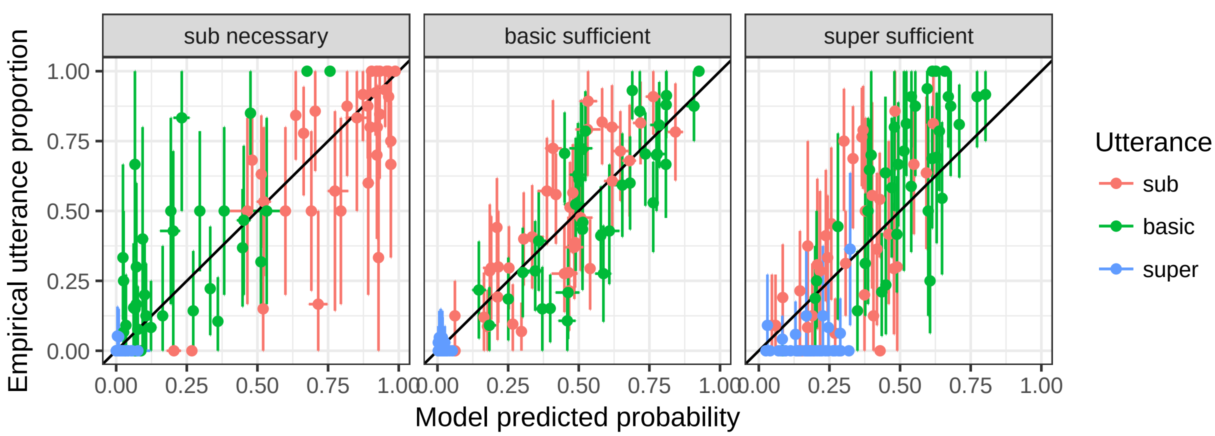


Figure 20: Scatterplot of empirical utterance proportions against MAP model predictions across conditions for nominal choice dataset (Exp. 3).

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, including the contextual informativeness of the referring expression and its cognitive cost, compared to its plausible alternatives. Importantly, computing contextual informativeness with respect to a *continuous* rather than Boolean underlying semantics was crucial for capturing various aspects of speakers' referring behavior. First, the continuous semantics allowed us to capture the basic well-documented asymmetry for speakers to be more likely to redundantly use color adjectives 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. And finally, treating empirically elicited typicality values as continuous semantic values yielded well-documented effects of typicality in both color modifier choice and noun choice. A modifier was more likely to be mentioned redundantly when the object was (out of context) 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 produced when the object was (out of context) a substantially less good instance of the higher than of the lower level.

We have thus shown that with one key innovation – a continuous 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 – because in this case, the prima facie sufficiently informative expression is more likely to lead a literal listener to draw a wrong inference about the intended referent. 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. And in contrast to previously proposed computational models, it is likely to be straightforwardly extendable to any instance of definite referring expressions of the sort we have examined here.

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 as it is commonly employed in the linguistic and psychological literature. The reason that redundant referring expressions are interesting for psycholinguists to study is that 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 one feature uniquely distinguishes the target object from the rest and a second one does not, mentioning the second does not contribute any information that is not already communicated by the first. Hence, the second is considered 'overinformative', a referring expression that contains it 'overspecified.'

Table 7: Overview of the models used for the three different production datasets Color/size (Exp. 1), Color typicality (Exp. 2), and Nominal choice (Exp. 3). Parameter names: x_{color}: semantic value of color; x_{size}: semantic value of size; $\beta_{c(\text{color})}$: cost of color; $\beta_{c(\text{size})}$: cost of size; β_i: weight on informativity; β_F: weight on cost (as estimated by utterance frequency); β_L: weight on cost (as estimated by utterance length); β_t: weight on elicited typicality values

	Color/size	Color typicality	Nominal choice
Semantic values applied to ...	modifier type (size, color)	utterance type (color, type) and utterance-object combination (e.g., *yellow banana* as referring to various bananas)	utterance-object combination (e.g., *dalmatian* as referring to various dogs)
Semantic values were...	inferred	inferred and elicited experimentally	elicited experimentally
Size of lexicon $\mathcal{L}(u, o)$	8 (all combinations of *size*, *color*, *othersize*, and *othercolor*)	814 (1 for each u, o combination)	51 (1 for each target u, o combination)
Contextual set of alternatives	all contextually available feature combinations (size, color)	all contextually available feature combinations (type, color)	3 target alternatives (level of reference: *sub, basic, super*)
Size of contextual set of alternatives	8	8 or 9	3
Compositionality	$\mathcal{L}(u_{\text{size}}, o) \times \mathcal{L}(u_{\text{color}}, o)$	$\mathcal{L}(u_{\text{color}}, o) \times \mathcal{L}(u_{\text{type}}, o)$ for utterance type values; elicited experimentally for utterance-object combination values	NA
Cost function	$\beta_{c(\text{color})} \mathbb{1}_{color}(u) + \beta_{c(\text{size})} \mathbb{1}_{size}(u)$	comparison between $\beta_F p(u) + \beta_L l(u)$ vs. $\beta_{c(\text{color})} \mathbb{1}_{color}(u) + \beta_{c(\text{type})} \mathbb{1}_{size}(u)$ vs. none	comparison between $\beta_F p(u) + \beta_L l(u)$ vs. none
Free parameters inferred in BDA	x_{color}, x_{size}, $\beta_{c(\text{color})}$, $\beta_{c(\text{size})}$, β_i	β_F, β_L, β_i, β_t	β_F, β_L, β_i, β_t

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 continuous semantics, a modifier that would be regarded as overinformative under the traditional conception may in fact communicate information about the referent. 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 the redundant modifier 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 captures the quantitative variation observed in speakers' production in a principled way while still assuming that speakers are aiming to be informative, and is compatible with other efficiency-based accounts of 'overinformative' referring expressions (e.g., Sedivy, 2003; Rubio-Fernandez, 2016).

But this raises a question: what counts as a truly overinformative utterance under continuous semantics RSA? RSA shifts the standard for overinformativeness and turns it into a graded notion: the less expected the use of a redundant modifier is contextually, the more the use of that modifier should be considered overinformative. For example, consider again Figure 8: the less scene variation there is, the more truly overinformative the use of the redundant modifier is. Referring to *the big purple stapler* when there are only purple staplers in the scene should be considered overinformative. If there is one red stapler, the utterance should be judged less overinformative, and the more non-purple staplers there are, the less overinformative the utterance should be judged. We leave a systematic test of this prediction for our stimuli for future research, though we point to some qualitative examples where it has been borne out previously in the next subsection.

6.3 Comprehension

While the account proposed in this paper is an account of the *production* of referring expressions, it can be extended straightforwardly to *comprehension*. RSA models typically assume that listeners interpret utterances 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? There are two interpretations of this question: first, what is the ultimate interpretation that listeners who reason about speakers characterized by the model provided in this paper arrive at, i.e. what are the predictions for referent choice? And second, how do the production probabilities enter into online processing of prima facie overinformative utterances? The first question has a clear answer. For the second question we offer a more speculative answer.

6.3.1 Choice of referent

Most RSA reference models, unlike the one reported in this paper, have focused on comprehension (M. C. Frank & Goodman, 2012; Degen, Franke, & Jäger, 2013; Qing & Franke, 2015; Franke & Degen, 2016). The formula that characterizes pragmatic listeners' referent choices is:

$$P_{L_1}(o|u) \propto P_{S_1}(u|o) \cdot P(o) \tag{9}$$

That is, the pragmatic listener interprets utterance u (e.g., *the big purple stapler*) via Bayesian inference, taking into account both the speaker probability of producing *the big purple stapler* and its alternatives, given a particular object o the speaker had in mind, as well as the listener's prior beliefs about which object the speaker is likely to intend to refer to in the context. For the situations considered in this paper, in which the utterance is semantically compatible with only one of the referents in the context, this always predicts that the listener should choose the target. And indeed, in Exps. 1-3 the error rate on the listeners' end was always below 1%. From a referent choice point of view, then, these contexts are not very interesting. They are much more interesting from an online processing point of view, which we discuss next.

6.3.2 Online processing

The question that has typically been asked about the online processing of redundant utterances is this: do redundant utterances, compared to their minimally specified alternatives, help or hinder comprehenders in choosing the intended referent? 'Help' and 'hinder' are typically translated into 'speed up' and 'slow down', respectively. What does the RSA model presented here have to say about this?

In sentence processing, the current wisdom is that the processing effort spent on linguistic material is related to how surprising it is (Hale, 2001; Levy, 2008). In particular, an utterance's log reading time is linear in its surprisal (Smith & Levy, 2013), where surprisal is defined as $-\log p(u)$. In these studies, surprisal is usually estimated from linguistic corpora. Consequently, an utterance of *the big purple stapler* receives a particular probability estimate independent of the non-linguistic context it occurred in. Here we provide a speaker model from which we can derive estimates of *pragmatic surprisal* directly for a particular context. We can thus speculate on a linking hypothesis: the more expected a redundant utterance is under the pragmatic continuous semantics speaker model, the faster it should be to process compared to its minimally specified alternative, all else being equal. We have shown that redundant expressions are more likely than minimal expressions when the sufficient dimension is relatively noisy and scene variation is relatively high. Under our speculative linking hypothesis, the redundant expression should be easier to process in these sorts of contexts than in contexts where the redundant expression is relatively less likely.

Is there evidence that listeners do behave in accordance with this prediction? Indeed, the literature reports 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 redundant 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 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 continuous semantics RSA account, this cursory overview of the patterns reported in the existing literature suggests that pragmatic surprisal may be a plausible linking function from model predictions to processing times. Excitingly, it has the potential for unifying the equivocal processing time evidence by providing a model of utterance probabilities that can be computed from the features of the objects in the context.

6.4 Continuous semantics

The crucial compoenent of the model that allows for capturing 'overinformativeness' effects is the continuous semantics. For the purpose of Exp. 1 (modifier choice), a semantic value was assigned to modifier *type*. The semantics of modifiers was underlyingly truth-conditional and the semantic value captured the probability that a modifier's truth conditions would accidentally be inverted. This model included only two semantic values, one for size and one for color, which we inferred from the data. For the datasets from Exps. 2 and 3, we then extended the continuous semantics to apply at the level of utterance-object combinations (e.g., *banana* vs. *blue banana* as applied to the blue banana item, *dalmatian* vs. *dog* as applied to the dalmatian item) to account for typicality effects in modifier and nominal choice. In this instantiation of the model, the semantic value differed for every utterance-object combination and captured how good of an instance of an utterance an object was. These values were elicited experimentally to avoid over-fitting, and for the dataset from Exp. 2 we found further that a combination of a noisy truth-conditional semantics and the empirically elicited semantics best accounted for the obtained production data.

What we have said nothing about thus far is what determines these semantic values; in particular, which aspects of language users' experience – perceptual, conceptual, communicative, linguistic – they represent. We will offer some speculative remarks and directions for future research here.

First, semantic values may 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 (at least in our stimuli). 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, consequently yielding a lower semantic value for a blue banana than for a yellow banana as an instance of *banana*. Further, the value may be context-invariant or context-dependent. If it is context-invariant, the semantic value inferred for color vs. size, for instance, should not vary by making size differences more salient and color differences less salient. If, instead, it is context-dependent, increasing the salience of size differences and decreasing the salience of color differences should result, e.g., in color modifiers being more noisy, with concomitant effects on production, i.e., redundant color modifiers should become less likely. This is indeed what Viethen, van Vessem, Goudbeek, & Krahmer, 2017 found.

Another possibility is that semantic values represent aspects of agents' prior beliefs (world knowledge) about the correlations between features of objects. For example, conditioning on an object being a banana, experience dictates that the probability of it being yellow is much greater than of it being blue. This predicts the relative ordering of the typicality values we elicited empirically, i.e., the blue banana received a lower semantic value than the yellow banana as an instance of *banana*.

Another possibility is that the semantic values capture the past probability of communicative success in using a particular expression. For example, the semantic value of *banana* as applied to a yellow banana may be high because in the past, referring to yellow bananas simply as *banana* was on average successful. Conversely, the semantic value of *banana* as applied to a blue banana may be low because in the past, referring to blue bananas simply as *banana* was on average unsuccessful (or the speaker may have uncertainty about its communicative success because they have never encountered blue bananas before). Similarly, the noise difference between color and size modifiers may be due to the inherent relativity of size modifiers compared to color modifiers – while color modifiers vary somewhat in meaning across domains (consider, e.g., the difference in redness between *red hair* and *red wine*), the interpretation of size modifiers is highly dependent on a comparison class (consider, e.g., the difference between a *big phone* and a *big building*). In negotiating what counts as *red*, then, speakers are likely to agree more often than in negotiating what counts as *big*. That is, size adjectives are more subjective than color adjectives. If semantic values encode adjective subjectivity, speakers should be even more likely to redundantly use adjectives that are more objective than color. In a study showing that adjective subjectivity is almost perfectly correlated with an adjective's average distance from the noun, Scontras et al. (2017) collected subjectivity ratings for many different adjectives and found that material adjectives like *wooden* and *plastic* are rated to be even more objective than color adjectives. Thus, under the hypothesis that semantic values represent adjective subjectivity, material adjectives should be even more likely to be used redundantly than color adjectives. This is not the case. For instance, Sedivy (2003) reports that material adjectives are used redundantly about as often as size adjectives. Hence, while the hypothesis that semantic values capture the past probability of communicative success in using a particular expression has yet to be systematically investigated, subjectivity alone seems not to be the determining factor.

Finally, it is also possible that semantic values are simply an irreducible part of the lexical entry of each utterance-object pair. This seems unlikely because it would require a separate semantic value for each utterance and object token, and most potentially encounterable object tokens in the world have not been encountered, making it impossible to store utterance-token-level values. However, it is possible that, reminiscent of prototype theory, semantic values are stored at the level of utterances and object *types*. This view of semantic values suggests that they should not be updated in response to further exposure of objects. For example, if semantic values 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 merits further investigation.

The various possibilities for the interpretation of the continuous semantic values included in the model are neither independent nor incompatible with each other. Disentangling these possibilities presents an exciting avenue for future research.

6.5 Audience design

One question which has plagued the literature on language production is that of whether, and to what extent, speakers actually tailor their utterances to their audience (Clark & Murphy, 1982; Horton & Keysar, 1996; Brown-Schmidt & Heller, 2014). This is also known as the issue of *audience design*. With regards to redundant referring expressions, the question is whether speakers produce redundant expressions because it is helpful to them (i.e., due to internal production pressures) or because it is helpful to their interlocutor (i.e., due to considerations of audience design). For instance, Walker (1993) shows that redundancy is more likely when processing resources are limited. On the other hand, there is evidence that redundant utterances are frequently used in response to signs of listener non-comprehension, when responding to listener questions, or when speaking to strangers (Baker, Gill, & Cassell, 2008), suggesting at least some consideration of listeners' needs.

RSA seems to make a claim about this issue: speakers are trying to be informative with respect to a literal listener. 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 listeners are, in line with previous work and briefly discussed above, more likely fall into the class of pragmatic L_1 listeners; listeners who reason about the speaker's intended meaning via Bayesian inference (M. C. Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013).[32]

Because RSA is a computational-level theory (Marr, 1982) of language use, it does not claim that the mechanism of language production requires that speakers actively consult an internal model of a listener every time they choose an utterance, just that the distribution of utterances they produce reflect informativity with respect to such a model. It is possible that this distribution is cached or computed using some other algorithm that doesn't explicitly involve a listener component.

Thus, the RSA model as formulated here remains agnostic about whether speakers' (over-) informativeness should be considered geared towards listeners' needs or simply a production-internal process. Instead, the claim is that redundancy emerges as a property of the communicative situation as a whole.

6.6 Other factors that affect redundancy

Continuous semantics 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) showed that colors are mentioned more often redundantly for clothes than for geometrical shapes. Her explanation is that knowing an object's color is generally more useful for clothing than it is for shapes. 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 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.

One factor that has been repeatedly discussed in the literature and that we have not taken up here is the *incrementality* of language production. For instance, according to Pechmann (1989), incrementality is to blame for redundancy: speakers retrieve and subsequently produce words as soon as they can. Because color modifiers are easier to retrieve than size modifiers, speakers produce them regardless of whether or not they are redundant. The problem with this account is that it predicts that the preferred adjective order should be reversed, i.e., color adjectives should occur before size adjectives. Pechmann does observe some instances of this occurring, but not many. In addition, it is unclear how incrementality could account for the systematic increase in color redundancy with increasing scene variation and decreasing color typicality, unless one makes the auxiliary assumption that the more contextually discriminative or salient color is, the more easily retrievable the modifier is. Indeed, Clark and Bangerter (2004) emphasize the importance of *salience against the common ground* in speakers' decisions about which of an object's properties to include in a referring expression.[33] However, there are other ways incrementality could play a role.

[32]But see Franke and Degen (2016) for an evaluation of the distribution of listener and speaker types in Quantity inferences.

[33]This is also what Sarah Brown-Schmidt seems to have had in mind when, in a comment during the Q&A at the CUNY Sentence Processing conference in 2016, where part of this work was presented, she exclaimed "I'm calling

For example, mentioning the color adjective may buy the speaker time when the noun is hard to retrieve. This predicts that in languages with post-nominal adjectives, where this delay strategy cannot be used for noun planning, there should be less redundant color mention; indeed, this is what Rubio-Fernandez (2016) shows for Spanish. The ways in which considerations of incremental language production can and should be incorporated in RSA are yet to be explored.

6.7 Extensions to other language production phenomena

In this paper we focused on providing a computationally explicit account of definite modified and nominal referring expressions in reference games, focusing on the use of prenominal size and color adjectives as well as on the taxonomic level of noun reference. The continuous semantics RSA model can be straightforwardly extended to different nominal domains and different properties. For instance, the literature has also explored 'overinformative' referring expressions that include material (*wooden, plastic*), other dimensional (*long, short*), and other physical (*spotted, striped*) adjectives.

However, beyond the relatively limited linguistic forms we have explored here, future research should also investigate the very intriguing potential for this approach to be extended to any language production phenomenon that involves content selection, including in the domain of reference (pronouns, names, definite descriptions with post-nominal modification) and event descriptions. For example, in investigations of optional instrument mentions, P. 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.

More generally, the approach should extend to any content selection phenomenon that affords a choice between a more or less specific utterance. Whenever the more specific utterance adds sufficient information, it should be included. This is related to surprisal based theories of production like Uniform Information Density (UID, Jaeger, 2006; Levy & Jaeger, 2007; A. Frank & Jaeger, 2008; Jaeger, 2010), where researchers have found that speakers are more likely to omit linguistic signal if the underlying meaning or syntactic structure is highly predictable. Importantly, UID diverges from our account in that it is an account of the choice between meaning-equivalent alternative utterances and includes no pragmatic reasoning component.

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.

it *the blue banana* because the banana is friggin' blue!". We agree that property salience is likely to factor into the redundant use of these adjectives. Investigating the extent to which salience is captured by the semantic value of the adjective that denotes the property in question, and to what extent it factors in independently, is a question we leave for further investigation.

A Effects of semantic value on utterance probabilities

Here we visualize the effect of different adjective types' semantic value on the probability of producing the insufficient color-only utterance (*blue pin*), the sufficient size-only utterance (*small pin*), or the redundant color-and-size utterance (*small blue pin*) to refer to the target in context Figure 1a under varying β_i values, in Figure 21. This constitutes a generalization of Figure 4, which is duplicated in row 6 ($\beta_i = 30$).

B Pre-experiment quiz

Before continuing to the main experiment, each participant was required 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)
- 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)

C Exp. 1 items

The following table lists all 36 object types from Exp. 1 and the colors they appeared in:

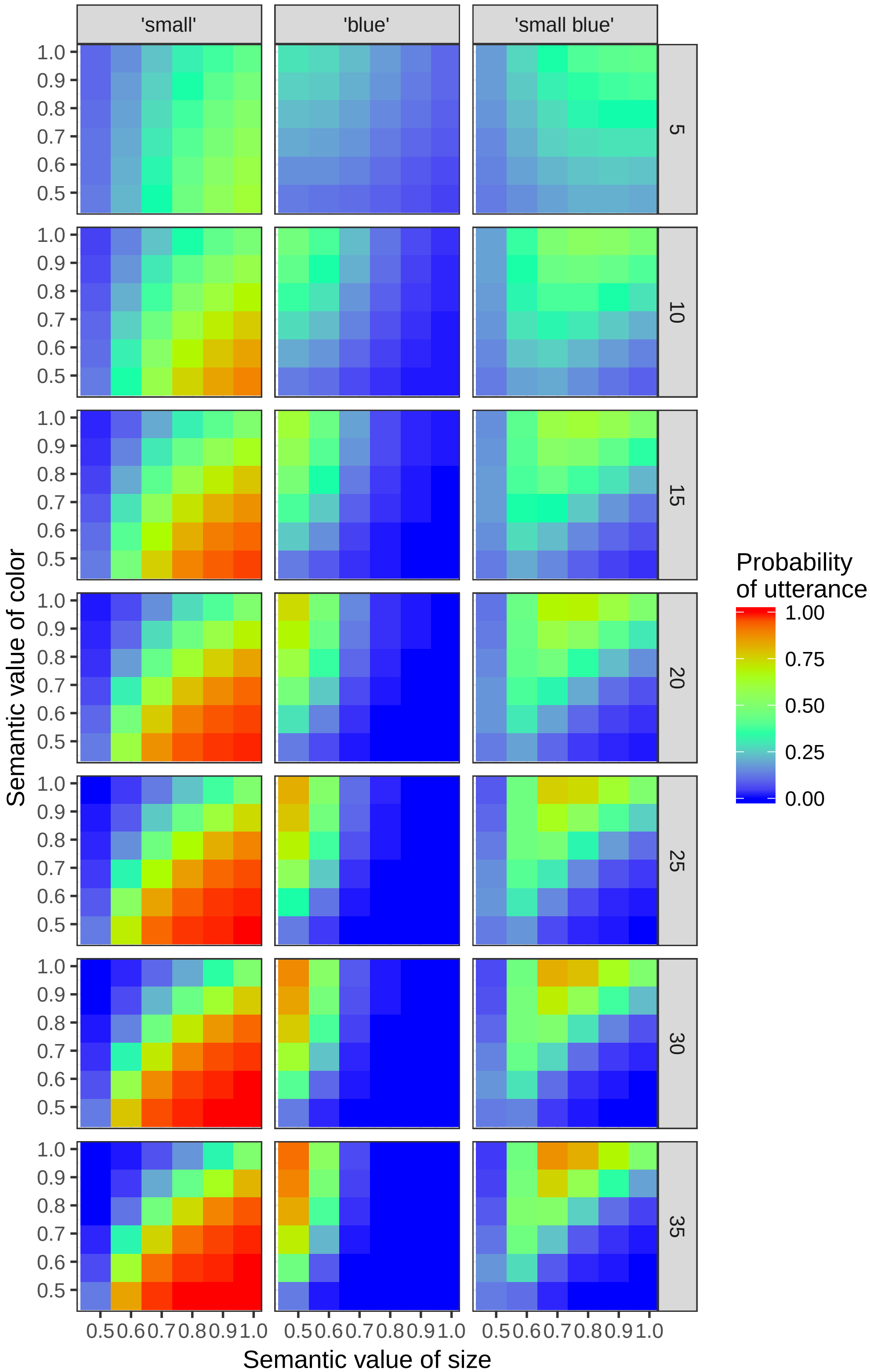


Figure 21: Probability of producing sufficient *small pin*, insufficient *blue pin*, and redundant *small blue pin* in contexts as depicted in Figure 1a, as a function of semantic value of color and size utterances and varying β_i row-wise (for $\beta_c = 0$).

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

D 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 the items in Exp. 1 and then included typicality as an additional predictor of redundant adjective use in the regression analysis reported in Section 3.1.3.

D.1 Methods

D.1.1 Participants

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

D.1.2 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 22 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.

Figure 22: A modified example trial from the typicality elicitation experiment.

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 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, i.e. when color is redundant.

D.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 23. 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 23, 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 (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 continuous semantics 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. continuous semantics RSA,

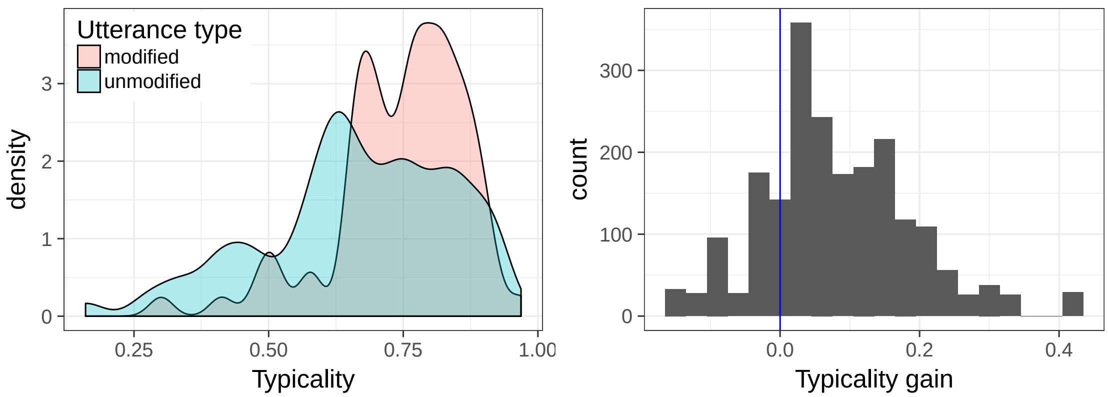


Figure 23: Typicality densities for modified and unmodified utterances (left) and histogram of typicality gains (differences between modified and unmodified typicalities, right).

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 24 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 8. 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.[34] Simple effects analysis reveals that this interaction 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.

[34]Conducting the same analysis on the entire dataset (i.e., using all of the noisy typicality estimates, replicated the 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 ($\beta = 1.52$, $SE = 1.45$, $p < .29$).

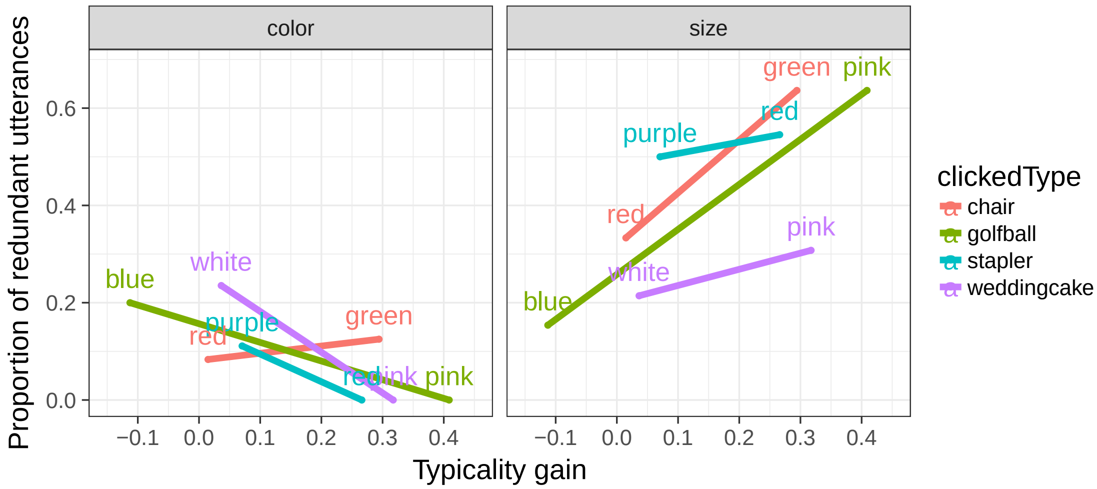


Figure 24: Utterance probability for four items as a function of difference in typicality between modified and unmodified utterance (x-axis) and sufficient dimension (columns).

Table 8: 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**

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 ($\beta = 3.09$, $SE = 1.45$, $p < .04$).[35]

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.

E Experiment 3 items

The following table lists all items used in Exp. 3 and the mean empirical utterance lengths that participants produced to refer to them:

F Typicality norms for Experiment 3

Analogous to the color typicality norms elicited for utterances in Exps. 1-2, we elicited typicality norms for utterances in Exp. 3. The elicited typicalities were used in the Bayesian Data Analysis reported in Section 5.2.

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. 3 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. 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 superordinate category (*dist-samesuper*,

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

Table 9: 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)

e.g., elephants share the superordinate category animal with dogs) on both the basic level and the superordinate level term (e.g., *dog* for elephant and *animal* for elephant), for a total of 469 combinations. Finally, we also tested each distractor of a different superordinate category than the target on the target's superordinate level term (*dist-diffsuper*, e.g., *animal* for rose). 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 25. As expected, typicality was close to 0 for distractor items with a different superordinate category as the target, and for subordinate/basic level terms used with distractors of the same superordinate category. 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.

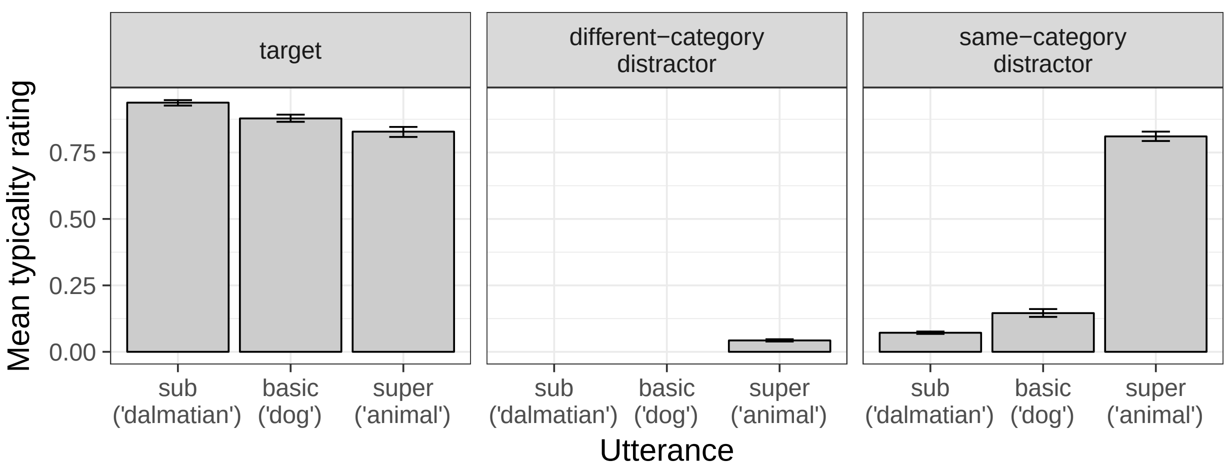


Figure 25: Mean typicality ratings by utterance (target subordinate, basic, and superordinate level term) for targets (e.g., *dalmatian*, left panel), distractors with a different superordinate category from the target (e.g., *rose*, middle panel), and distractors with the same superordinate category as the target (e.g., *elephant*, right panel). Error bars indicate bootstrapped 95% confidence intervals.

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

Baker, R., Gill, A. J., & Cassell, J. (2008). Reactive redundancy and listener comprehension in direction-giving. In *9th sigdial workshop on discourse and dialogue* (pp. 37–45). Retrieved from http://acl.eldoc.ub.rug.nl/mirror/W/W08/W08-0105.pdf doi: 10.3115/1622064.1622071

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

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

Bergen, L., Levy, R., & Goodman, N. (2016). Pragmatic reasoning through semantic inference. *Semantics and Pragmatics*, *9*(1984), 1–46. Retrieved from http://semprag.org/article/view/sp.9.20 doi: 10.3765/sp.9.20

Bloomfield, L. (1933). *Language.* New York: Holt.

Brennan, S. E., & Clark, H. H. (1996, nov). Conceptual pacts and lexical choice in conversation. *Journal of experimental psychology. Learning, memory, and cognition*, *22*(6), 1482 – 1493. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8921603

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

Brown, R. (1958). Words and things.

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

Bürkner, P.-C. (2017). brms: An R package for bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. doi: 10.18637/jss.v080.i01

Clark, H. H., & Bangerter, A. (2004). Changing Ideas about Reference. In I. A. Noveck & D. Sperber (Eds.), *Experimental pragmatics* (pp. 25–49). Basingstoke, UK: Palgrave MacMillan. doi: 10.1057/9780230524125_2

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

Cohen, B., & Murphy, G. L. (1984). Models of concepts. *Cognitive science*, *8*(1), 27–58.

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 *Cogsci.*

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. (2006, 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.
Franke, M., & Degen, J. (2016). Reasoning in Reference Games : Individual- vs . Population-Level Probabilistic Modeling. *PLoS ONE*, *11*(5), 1–25. doi: 10.1371/journal.pone.0154854
Gatt, A., Krahmer, E., van Deemter, K., & van Gompel, R. P. (2014). Models and empirical data for the production of referring expressions. *Language, Cognition and Neuroscience*, *29*(8), 899–911.
Gatt, A., Krahmer, E., Van Deemter, K., & van Gompel, R. P. (2017). Reference production as search: The impact of domain size on the production of distinguishing descriptions. *Cognitive science*, *41*, 1457–1492.
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., & Frank, M. C. (2016). Pragmatic Language Interpretation as Probabilistic Inference. *Trends in Cognitive Sciences*, *20*(11), 818–829. Retrieved from http://dx.doi.org/10.1016/j.tics.2016.08.005 doi: 10.1016/j.tics.2016.08.005
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.
Griffin, Z. M., & Bock, K. (1998). Constraint, word frequency, and levels of processing in spoken word production. *Journal of Memory and Language*, *38*(38), 313–338.
Hale, J. (2001). A Probabilistic Earley Parser as a Psycholinguistic Model. In *Proceedings of the second meeting of the north american chapter of the asssociation for computational linguistics* (pp. 1–8). Association for Computational Linguistics.
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.

Hoffmann, J., & Ziessler, C. (1983). Objektidentifikation in künstlichen begriffshierarchien. *Zeitschrift für Psychologie mit Zeitschrift für angewandte Psychologie*.

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.

Jescheniak, J. D., & Levelt, W. J. M. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4), 824–843. Retrieved from `http://doi.apa.org/getdoi.cfm?doi=10.1037/0278-7393.20.4.824` doi: 10.1037//0278-7393.20.4.824

Johnson, K. E., & Mervis, C. B. (1997). Effects of varying levels of expertise on the basic level of categorization. *Journal of Experimental Psychology: General*, *126*(3), 248–277. Retrieved from `papers3://publication/uuid/F6C763F3-CAD3-479E-BFED-1EF941293840` doi: 10.1037/0096-3445.126.3.248

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

Kamp, H., & Partee, B. (1995, nov). Prototype theory and compositionality. *Cognition*, *57*(2), 129–91. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/8556840`

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.

Kennedy, C., & McNally, L. (2005). Scale Structure, Degree Modification, and the Semantics of Gradable Predicates. *Language*, *81*(2), 345–381. Retrieved from `http://muse.jhu.edu/content/crossref/journals/language/v081/81.2kennedy.pdf` doi: 10.1353/lan.2005.0071

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

Levinson, S. C. (1983). Pragmatics (cambridge textbooks in linguistics).

Levy, R. (2008, mar). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–77. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/17662975` doi: 10.1016/j.cognition.2007.05.006

Levy, R., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In B. Schlökopf, J. Platt, & T. Hoffman (Eds.), *Advances in neural in-*

formation 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`

Luce, R. D. (1959). *Individual Choice Behavior: A Theoretical Analysis.* New York: Wiley.

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

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* San Francisco: W.H. Freeman.

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`

Murphy, G. L., & Smith, E. E. (1982). Basic-level superiority in picture categorization. *Journal of verbal learning and verbal behavior*, *21*(1), 1–20.

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

Oldfield, R. C., & Wingfield, A. (1965). Response latencies in naming objects. *Quarterly Journal of Experimental Psychology*, *17*(4), 273–281.

Olson, D. R. (1970). Language and thought: Aspects of a cognitive theory of semantics. *Psychological review*, *77*(4), 257.

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

Qing, C., & Franke, M. (2015). Bayesian Natural Language Semantics and Pragmatics. In H. Zeevat & H. Schmitz (Eds.), *Bayesian natural language semantics and pragmatics* (pp. 201–220). Cham: Springer. Retrieved from `https://books.google.com/books?hl=en{\&}lr={\&}id=Yt3yCQAAQBAJ{\&}oi=fnd{\&}pg=PR5{\&}dq=pragmatics+language{\&}ots=BbyeJKxAPf{\&}sig=Bh4ymPWUhaDuu9VLnZKEWacg3jE` doi: 10.1007/978-3-319-17064-0

R Core Team. (2017). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Rohde, H., Seyfarth, S., Clark, B., Jäger, G., & Kaufmann, S. (2012). Communicating with Cost-based Implicature: a Game-Theoretic Approach to Ambiguity. In *Proceedings of the 16th workshop on the semantics and pragmatics of dialogue* (pp. 107 – 116).

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

Scontras, G., Degen, J., & Goodman, N. D. (2017). Subjectivity Predicts AdjectiveOrdering Preferences. *Open Mind: Discoveries in Cognitive Science*, *1*(1), 53–65. doi: 10.1162/opmi

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`

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319. Retrieved from `http://dx.doi.org/10.1016/j.cognition.2013.02.013` doi: 10.1016/j.cognition.2013.02.013

Sproat, R., & Shih, C. (1991). The cross-linguistic distribution of adjective ordering restrictions. In *Interdisciplinary approaches to language* (pp. 565–593). Springer Netherlands.

Tanaka, J. W., & Taylor, M. (1991a). Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive psychology*, *23*(3), 457–482.

Tanaka, J. W., & Taylor, M. (1991b). 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

van Miltenburg, E., Koolen, R., & Krahmer, E. (2018). Varying image description tasks : spoken versus written descriptions. In *Proceedings of the fifth workshop on nlp for similar languages, varieties and dialects* (pp. 88–100).

Viethen, J., van Vessem, T., Goudbeek, M., & Krahmer, E. (2017). Color in Reference Production: The Role of Color Similarity and Color Codability. *Cognitive Science*, *41*, 1493–1514. doi: 10.1111/cogs.12387

Walker, M. A. (1993). *Informational Redundancy and Resource Bounds in Dialogue* (Unpublished doctoral dissertation). University of Pennsylvania.

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

Zadeh, L. A. (1965). Fuzzy sets. *Information and control*, *8*(3), 338–353.