

# When Redundancy Is Useful: A Bayesian Approach to “Overinformative” Referring Expressions

Judith Degen and Robert D. Hawkins  
Stanford University

Caroline Graf and Elisa Kreiss  
Osnabrück University

Noah D. Goodman  
Stanford University

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, we relax the assumption that informativeness is computed with respect to a deterministic Boolean semantics, in favor of a nondeterministic continuous semantics. This innovation 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 increase in specificity in nominal reference as a function of typicality. These findings cast a new light on the production of referring expressions: rather than being wastefully overinformative, reference is usefully redundant.


**Keywords:** language production, reference, overinformativeness, experimental pragmatics, Bayesian modeling

**Supplemental materials:** <http://dx.doi.org/10.1037/rev0000186.supp>

Reference to objects is one of the most basic and prevalent uses of language. In order to refer, speakers must choose from a wealth of referring expressions at their disposal. How does a speaker decide whether to call an object *the animal*, *the dog*, *the dalmatian*, or *the big mostly white dalmatian*? The context within which the object occurs (other nondogs, other dogs, other dalmatians) plays a large part in determining which fea-

tures 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 exhibit what has been claimed to be *overinformativeness*: Referring expressions are often more specific than necessary for establishing unique reference, and they are more specific in systematic ways.

This article was published Online First April 2, 2020.

 Judith Degen and Robert D. Hawkins, Department of Psychology, Stanford University; Caroline Graf and Elisa Kreiss, Institute for Cognitive Science, Osnabrück University; Noah D. Goodman, Department of Psychology, Stanford University.

Robert D. Hawkins is now at the Department of Psychology, Princeton University. Caroline Graf is now at the Department of Sociology, Vrije Universiteit Amsterdam. Elisa Kreiss is now at the Department of Linguistics, Stanford University.

The earliest precursor of this work (the core idea of a continuous semantics RSA model and Experiment 1) was presented as a talk at the RefNet Round Table Event in 2016 and at AMLaP 2016. Experiment 2 and the corresponding model were presented as a talk at the CUNY Conference on Sentence Processing in 2017 and as a poster at the Experimental Pragmatics (XPrag) Conference in 2017. An earlier version of Experiment 3 and an earlier version of the corresponding model appeared in the Proceedings of the 38th Conference of the Cognitive Science Society. All experiments and models have been presented by Judith Degen in various invited talks at workshops and colloquia in linguistics, psychology, philosophy, and cognitive science since 2016.

We are grateful to many audiences and individuals for discussions that have made this article better, in particular Sarah Brown-Schmidt, Herb Clark, Michael Franke, Gerhard Jäger, Florian Jaeger, Emiel Krahmer, Timo Roettger, Paula Rubio-Fernández, Greg Scontras, Mike Tanenhaus, and Roger van Gompel; audiences at the RefNet Round Table Event 2015, CogSci 2016, AMLaP 2016, CUNY 2017, XPrag 2017, the UC Davis Perception, Cognition, and Cognitive Neuroscience Brown Bag, the Bay Area Philosophy of Science Group, the IMBS Quantitative Approaches to Language Workshop 2018; and audiences at colloquium talks given by the first author at Berkeley Psychology, CMU Philosophy, Dartmouth Linguistics, McGill Linguistics, OSU Linguistics, Stanford Linguistics, Tübingen Linguistics, and UC Davis Linguistics. All remaining errors are our own.

This work was supported by Office of Naval Research grants N00014-13-1-0788 and N00014-13-10341. Robert D. Hawkins was supported by the Stanford Graduate Fellowship and the National Science Foundation Graduate Research Fellowship under Grant No. DGE-114747.

Correspondence concerning this article should be addressed to Judith Degen, who is now at the Department of Linguistics, Stanford University, 450 Jane Stanford Way, Stanford, CA 94305. E-mail: [jdegen@stanford.edu](mailto:jdegen@stanford.edu)

This article is concerned with developing a unified quantitative account for these systematic patterns, which has so far proven elusive. We formalize our account as a computational model of referring expression production within the Rational Speech Act (RSA) framework (M. C. Frank & Goodman, 2012; Franke & Jäger, 2016; 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 informativeness of utterances is computed with respect to a deterministic Boolean semantics. Under this relaxed semantics, certain terms may apply better than others to an object without strictly being true or false. This idea has its oldest modern precursor in fuzzy logic (Zadeh, 1965). It is similar in spirit to recently proposed models of meaning in both computational semantics, which assign probabilities rather than truth conditions to sentences (Bernardy, Blanck, Chatzikyriakidis, & Lap-pin, 2018), and in natural language processing, which treat word and sentence meanings as vectors of real numbers (Devlin, Chang, Lee, & Toutanova, 2018; Pennington, Socher, & Manning, 2014; Peters et al., 2018).

As we will show, computing utterance informativeness with respect to these more graded meanings can explain a number of seemingly disparate phenomena. 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 restrictive adjectives occurring between the determiner and the noun.<sup>1</sup> This broad class of referring expressions subsumes two domains in language production that have been typically treated as separate. The choice of adjectives in (purportedly) overmodified referring expressions has been a primary focus of the language production literature (Arts, Maes, Noordman, & Jansen, 2011; Engelhardt, Bailey, & Ferreira, 2006; Herrmann & Deutsch, 1976; Koolen, Gatt, Goudbeek, & Kraemer, 2011; Maes, Arts, & Noordman, 2004; Nadig & Sedivy, 2002; Pechmann, 1989; Rubio-Fernandez, 2016; Sedivy, 2003), while the choice of noun in simple nominal expressions 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 (Xu & Tenenbaum, 2007; but see Dale & Reiter, 1995 for a treatment of basic level terms in natural language generation).

This paper is structured as follows: first, we review several key overinformativeness phenomena across these literatures that have presented a puzzle for rational accounts of language use. We then introduce the basic RSA framework with deterministic Boolean semantics and show how it can be extended to a relaxed semantics. In the subsequent three sections we evaluate the relaxed semantics RSA model on data from interactive online reference game experiments that exhibit the phenomena introduced in the following: asymmetries in 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. In each case, our model explains why seemingly overinformative modifiers or overly specific nouns can in fact be useful and informative; not doing so might lead the listener astray, or require them to invest too much processing effort. We wrap up by summarizing our findings and discussing the far-reaching implications of and further challenges for this line of work.

## Production of Referring Expressions: A Case Against Rational Language Use?

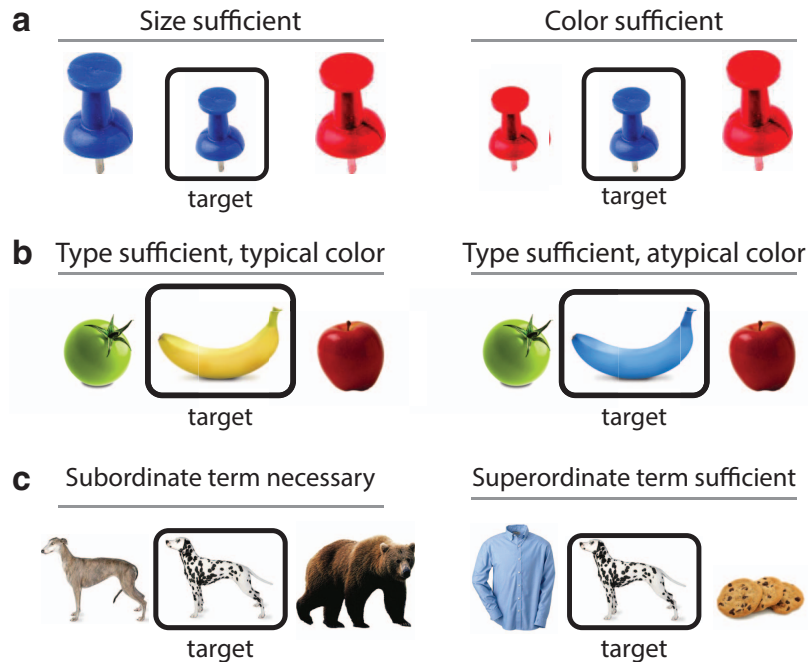
How should a cooperative speaker choose between competing referring expressions? Grice, in his seminal work, provided some guidance by formulating his famous conversational maxims, intended as a guide to listeners' expectations about cooperative speaker behavior (Grice, 1975). His maxim of Quantity, 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; Davies & Katsos, 2013; Engelhardt et al., 2006; Levinson, 1983; Olson, 1970), speakers seem remarkably willing to systematically violate Quantity-2. For example, 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, left; Arts et al., 2011; Gatt, van Gompel, Kraemer, & van Deemter, 2011; Gatt, Kraemer, van Deemter, & van Gompel, 2014; Koolen et al., 2011) and routinely use a basic level term even when a superordinate level term would be sufficient (e.g., *the dog* instead of *the animal* in contexts like Figure 1c, right; R. Brown, 1958; Hoffmann & Ziessler, 1983; Johnson & Mervis, 1997; Rosch et al., 1976; Tanaka & Taylor, 1991).

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 appealed to a built-in preference for referring at the basic level from considerations of conceptual representation or perceptual factors such as shape (Murphy & Smith, 1982; Rosch, 1973; Rosch et al., 1976). 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 itself 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 Boolean 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 *usefully 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.

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



*Figure 1.* Example referential contexts that differ in which property needs to be minimally included in a referring expression to establish unique reference to the intended referent. Intended referent is marked with a border. (a) Contexts where size only (left, e.g., *the small pin*) or color only (right, e.g., *the blue pin*) is sufficient. (b) Contexts where type only (*the banana*) is sufficient and color is typical (left) or atypical (right). (c) Contexts where the subordinate level term is necessary (left, e.g., *the dalmatian*) or the superordinate term is sufficient (right, e.g., *the animal*). See the online article for the color version of this figure.

### Overinformativeness in Modified Referring Expressions

Most of the literature on overinformative referring expressions has been devoted to the use of overinformative modifiers in modified referring expressions. The prevalent observation is that speakers frequently do not include only the minimal modifiers required for establishing reference, but often also include redundant modifiers (Arts et al., 2011; Engelhardt et al., 2006; Koolen et al., 2011; Maes et al., 2004; Nadig & Sedivy, 2002; Pechmann, 1989). 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). Furthermore, these asymmetries interact with features of the context and world knowledge about the typicality of different properties.

**Asymmetry in redundant use of color and size adjectives.** In Figure 1a (left), distinguishing the object highlighted by the thick border requires only mentioning its size (*the small pin*). 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 (Belke & Meyer, 2002; Gatt et al., 2011; Pechmann, 1989). However, the same is not true for size: In contexts like Figure 1a (right), 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 (Gatt et al., 2011; Koolen, Goudbeek, & Kraemer, 2013; Pechmann, 1989; Rubio-Fernandez, 2016; Sedivy, 2003; Westerbeek et al., 2015).

**Scene variation.** Speakers' propensity to overmodify with color is highly dependent on features of the distractor objects in the context. In particular, as the variation present in the scene increases, so does the probability of overmodifying. For example Koolen, Goudbeek, and Kraemer (2013) consistently found higher rates of overmodification with color adjectives in high-variation scenes (28–27%) compared with the low-variation ones (4–10%). Scene variation has been quantified in several different ways: the number of dimensions along which objects differ (Koolen et al., 2013), the number of distractors present in a scene (Gatt, Kraemer, van Deemter, & van Gompel, 2017), and whether objects are “simple” or “compositional” (Davies & Katsos, 2013). A model of referring expression generation should ideally capture all of these types of variation in a unified way.

**Feature typicality.** Overmodification with color has also been shown to be systematically related to the typicality of the color for the object. Westerbeek, Koolen, and Maes (2015) 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 (see also Rubio-Fernandez, 2016; Sedivy, 2003). For example, speakers never refer to a yellow banana in the absence of other bananas as *the yellow banana* (see Figure 1b, left), but they sometimes refer to a brown banana as *the brown banana*, and they almost always

Table 1  
*List of Effects a Theory of Referring Expression Production Should Account for and Experiment in Which They Are Addressed*

| Experiment | Effect                 | Description  |
|------------|------------------------|--|
| 1          | Color/size asymmetry   | More redundant use of color than size <sup>2</sup>                         |
| 1          | Scene variation        | More redundant use of color with increasing scene variation <sup>3</sup>   |
| 2          | Color typicality       | More redundant use of color with decreasing color typicality <sup>4</sup>  |
| 3          | Basic level preference | Preference for basic level term when superordinate sufficient <sup>5</sup> |
| 3          | Subordinate level use  | Unnecessary use of subordinate level term <sup>6</sup>                     |

refer to a blue banana as *the blue banana* (see Figure 1b, right). Similar typicality effects have been shown for other (noncolor) 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.

### Overinformativeness in Nominal Referring Expressions

Even in the absence of modifying 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. A well-documented effect from the concepts and categorization literature is that speakers prefer to refer at the *basic level* (Rosch et al., 1976; Tanaka & Taylor, 1991). That is, in the absence of other constraints, even when a superordinate level term would be sufficient for establishing reference (as in Figure 1c, right), speakers prefer to say *the dog* rather than *the animal*. However, there are systematic exceptions: in some cases when the basic level would be sufficient, speakers prefer the *subordinate* term. For example, atypical birds like penguins are often referred to at the subordinate level rather than at the basic level *bird* (Jolicoeur, Gluck, & Kosslyn, 1984). Indeed, children may even use expectations about such referential preferences to infer narrower categories from atypical exemplars during word learning (Emberson, Loncar, Mazzei, Treves, & Goldberg, 2019).

### Modeling Speakers' Choice of Referring Expression

To date, there is no theory to account for all of these different phenomena (see Table 1), and no model has attempted to unify the domains of modified and nominal referring expressions. Here we propose an explicit computational account of how multiple factors—including an utterance's literal semantics, its contextual informativity, its cost relative to alternative utterances, and the typicality of an object or its features—interact in referring expression production. We argue that this model provides a principled explanation for the phenomena reviewed in the previous section and holds promise for being generalizable to many further production phenomena related to overinformativeness, which we discuss in relation to previous accounts in the General Discussion.

Our model is formulated within the RSA framework (M. C. Frank & Goodman, 2012; Goodman & Frank, 2016).<sup>7</sup> We proceed by first presenting the general production framework and show why the most basic model, as formulated by M. C. Frank and Goodman (2012), does not produce the phenomena outlined above due to its strong focus on speakers maximizing the informativeness of expressions under a deterministic Boolean semantics. We then introduce our crucial innovation: relaxing the semantics.

### The Basic Rational Speech Act 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:

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

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:

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

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 (left),  $U = \{big, small, blue, red\}$  and  $O = \{o_{big\_blue}, o_{big\_red}, o_{small\_blue}\}$ . Upon observing *blue*, the literal listener therefore assigns equal probability to  $o_{big\_blue}$  and  $o_{small\_blue}$ . Values of  $P_{L_0}(o|u)$  for each  $u$  are shown on the left in Table 2.

<sup>2</sup> Reported by many (e.g., Engelhardt et al., 2006; Gatt et al., 2011; Pechmann, 1989; Rubio-Fernandez, 2016).

<sup>3</sup> Multiple replications reported (e.g., Davies & Katsos, 2013; Koolen et al., 2013).

<sup>4</sup> Multiple replications reported (e.g. Rubio-Fernandez, 2016; Sedivy, 2003; Westerbeek et al., 2015).

<sup>5</sup> Originally reported by Rosch et al. (1976), dozens of replications.

<sup>6</sup> Reported by Jolicoeur et al. (1984).

<sup>7</sup> All RSA models and Bayesian data analyses reported in this article were implemented in the probabilistic programming language WebPPL (Goodman & Stuhlmüller, n.d.) and can be viewed at [https://github.com/thericean/RE\\_production](https://github.com/thericean/RE_production). All experimental materials and analysis scripts are available in the same repository. An interactive browser-based toy model is provided at <http://forestdb.org/models/overinf.html>.

Table 2  
 Row-Wise Literal Listener Distributions  $P_{L_0}(o|u)$  for Each Utterance  $u$  in the Size-Sufficient Context Depicted in the Left Panel of Figure 1a, Under a Deterministic Boolean Semantics (Left) or Under a Continuous Semantics (Right) With  $x_{size} = .8$ ,  $x_{color} = .99$

| Utterance         | Boolean         |                |                   | Continuous      |                |                   |
|-------------------|-----------------|----------------|-------------------|-----------------|----------------|-------------------|
|                   | $o_{big\_blue}$ | $o_{big\_red}$ | $o_{small\_blue}$ | $o_{big\_blue}$ | $o_{big\_red}$ | $o_{small\_blue}$ |
| <i>big</i>        | .5              | .5             | 0                 | .39             | .39            | .22               |
| <i>small</i>      | 0               | 0              | <b>1</b>          | .26             | .26            | <b>.48</b>        |
| <i>blue</i>       | .5              | 0              | .5                | .42             | .16            | .42               |
| <i>red</i>        | 0               | 1              | 0                 | .21             | .57            | .21               |
| <i>big blue</i>   | 1               | 0              | 0                 | .50             | .23            | .27               |
| <i>big red</i>    | 0               | 1              | 0                 | .24             | .52            | .24               |
| <i>small blue</i> | 0               | 0              | <b>1</b>          | .27             | .23            | <b>.50</b>        |

Note. Bolded numbers indicate crucial comparisons between literal listener probabilities in correctly selecting the intended referent  $o_{small\_blue}$  in response to observing the sufficient small and the redundant small blue utterances.

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

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

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,  $\beta_i$  and  $\beta_c$ , enter the computation, weighting the respective contributions of informativeness and utterance cost, respectively.<sup>8</sup> In order to understand the effect of  $\beta_i$ , it is useful to explore its effect when utterances are cost-free. In this case, as  $\beta_i$  approaches infinity, the speaker increasingly only chooses utterances that maximize informativeness; if  $\beta_i$  is 0, informativeness is disregarded and the speaker chooses randomly from the set of all available utterances; if  $\beta_i$  is 1, the speaker probability-matches, that is, 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  $o_{small\_blue}$  they have two semantically possible utterances, *small* and *blue*, where *small* is twice as informative as *blue*. They produce *small* with probability 1 when  $\beta_i \rightarrow \infty$ , probability 2/3 when  $\beta_i = 1$  and probability 1/4 when  $\beta_i = 0$ .<sup>9</sup> Conversely, disregarding informativeness and focusing only on cost, any asymmetry in costs will be exaggerated with increasing  $\beta_c$ , such that the speaker will choose the least costly utterance with higher and higher probability as  $\beta_c$  increases.

As has been pointed out by van Gompel, van Deemter, Gatt, Snoeren, and Krahmer (2019), the basic RSA model described so far (M. C. Frank & Goodman, 2012) does not generate overinformative referring expressions for two 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  $u_{size} \in \{big, small\}$  and  $u_{color} \in \{blue, red\}$ , such that the meaning of a complex two-word utterance is defined as

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

The resulting renormalized literal listener distributions for our example size-sufficient context in the left panel of Figure 1a are shown in the left columns in Table 2,<sup>10</sup> and the concomitant pragmatic speaker distributions are shown in the left columns in Table 3.<sup>11</sup>

Unfortunately, simply including complex utterances in the set of alternatives does not solve the problem. We turn again to the case where the speaker wants to communicate the small blue object. There are now two useful utterances, *small* and *small blue*, for referring to this object. Because they are equally informative (see bolded numbers in Table 2, column 3), the pragmatic speaker is equally likely to produce them (see bolded numbers in Table 3, column 3). The only way for the more 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

<sup>8</sup> 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 et al., 2013; Rohde et al., 2012) 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; van Gompel, van Deemter, Gatt, Snoeren, & Krahmer, 2019). 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 and 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.

<sup>9</sup> Note that instead of a  $\beta_i$  parameter weighting informativeness inside the utility function, other recent formulations have used an  $\alpha$  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.

<sup>10</sup> "Normalization" refers to the process of turning a set of numbers into a probability distribution by dividing each number by the sum of all the numbers in the set, such that they add up to 1.

<sup>11</sup> An interactive toy version of this model is provided at <http://forestdb.org/models/overinf.html>.

Table 3

Column-Wise Pragmatic Speaker Distributions  $P_{S_i}(u|o)$  for Each Object  $o$  in the Size-Sufficient Context Depicted in the Left Panel of Figure 1a, Under a Deterministic Boolean Semantics (Left) or Under a Continuous Semantics (Middle, Right) With  $x_{size} = .8$ ,  $x_{color} = .99$ , With  $\beta_i$  Set to 1 (Middle) or 30 (Right)

| Utterance         | Boolean         |                |                   | Continuous ( $\beta_i = 1$ ) |                |                   | Continuous ( $\beta_i = 30$ ) |                |                   |
|-------------------|-----------------|----------------|-------------------|------------------------------|----------------|-------------------|-------------------------------|----------------|-------------------|
|                   | $O_{big\_blue}$ | $O_{big\_red}$ | $O_{small\_blue}$ | $O_{big\_blue}$              | $O_{big\_red}$ | $O_{small\_blue}$ | $O_{big\_blue}$               | $O_{big\_red}$ | $O_{small\_blue}$ |
| <i>big</i>        | .25             | .2             | 0                 | .17                          | .17            | .09               | 0                             | 0              | 0                 |
| <i>small</i>      | 0               | 0              | <b>.4</b>         | .11                          | .11            | <b>.20</b>        | 0                             | 0              | <b>.21</b>        |
| <i>blue</i>       | .25             | 0              | .2                | .18                          | .07            | .18               | .01                           | 0              | 0                 |
| <i>red</i>        | 0               | .4             | 0                 | .09                          | .24            | .09               | 0                             | .93            | 0                 |
| <i>big blue</i>   | .5              | 0              | 0                 | .22                          | .10            | .12               | .99                           | 0              | 0                 |
| <i>big red</i>    | 0               | .4             | 0                 | .10                          | .22            | .10               | 0                             | .07            | 0                 |
| <i>small blue</i> | 0               | 0              | <b>.4</b>         | .12                          | .10            | <b>.21</b>        | 0                             | 0              | <b>.79</b>        |

Note. Bolded numbers indicate the relevant speaker probabilities for the minimal (small) and redundant (small blue) utterances when intending to communicate referent  $o_{small\_blue}$ .

being cheaper than simple utterances is highly dubious.<sup>12</sup> Thus we must look elsewhere to break the symmetry and account for overinformativeness. We propose that the place to look is the computation of informativeness itself.

### RSA With Continuous Semantics

Here we introduce the crucial innovation: rather than assuming a deterministic Boolean 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:

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

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, prototype theory, and recent developments in natural language processing; Bernardy et al., 2018; Rosch, 1973; Zadeh, 1965).

Another approach to relaxing the deterministic Boolean semantics would be to relax the determinism. This can be done either by assuming a semantics which is fundamentally Boolean, but whose truth-values contain an element of randomness; or by assuming a fully deterministic Boolean semantics with intensional parameters that are themselves random variables. This is appealing because it would preserve the existing machinery of standard truth-functional compositional semantics. It can be shown that using continuous semantic values in the RSA model is equivalent to using Boolean values that are chosen nondeterministically. Conversely, marginalizing over the randomness in a Boolean semantics yields a probability of truth, which is a value between 0 and 1. For this reason we will sometimes refer to the relaxed semantics as a “noisy” semantics, and the deviation of the semantic value from 0 or 1 as the degree of noise. We will generally treat the relaxed semantics in its continuous value guise, as it simplifies exposition and development.

We now show via simulations that this model can qualitatively account both for speakers’ asymmetric propensity to overmodify with color rather than with size and for speakers’ propensity to overmodify more with increasing scene variation.

The intuition, using the example from Figure 1a (left), is that *blue* and *small* do not apply equally well to all roughly blue, roughly small objects, and that a speaker might opt to include more modifiers when any one alone might not be a perfectly apt descriptor. Assuming that *blue* is more precise than *small* leads the speaker to overmodify more with color than with size—and further, the more variability is present in the scene, the more the precision of color helps weed out nonintended referents, that is, the more color overmodification occurs.

**Simulation 1: Color-size asymmetry.** 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 semantics in which all colors are treated the same, all sizes are as well, and the two compose via a product rule. That is, when an object  $o$  is in the extension of a size adjective under a Boolean semantics—that is, when the size can be truthfully predicated of  $o$ —we take  $\mathcal{L}(u, o) = x_{size}$ , a constant; when it is not in the extension of the adjective—that is, when the size cannot be truthfully predicated of  $o$ — $\mathcal{L}(u, o) = 1 - x_{size}$ . 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, Equation 5, this fully specifies a relaxed semantic function for our reference domain.<sup>13</sup>

Now consider the RSA literal listener, Equation 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  $x_{size} = .8$  and  $x_{color} = .99$  on the right in Table 2.<sup>14</sup> 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

<sup>12</sup> See also the discussion of cost functions in Kraher, van Erk, and Verleg (2003), who explicitly introduce this monotonicity constraint as a constraint on the search space of possible referring expressions within a graph-based framework.

<sup>13</sup> An interactive toy version of this model is provided at <http://forestdb.org/models/overinf.html>.

<sup>14</sup> 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 van Gompel et al. (2019): 79% in size-sufficient contexts; 7% in color-sufficient contexts.

this purpose, we compare the literal listener probabilities in the  $o_{\text{small\_blue}}$  column. The two best utterances under both the Boolean and the continuous semantics are bolded in the table: under the Boolean 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 (.48) of retrieving the intended referent than the redundant *small blue* (.50). While this difference may appear small, it is enough to break the symmetry in utterance informativeness. 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  $\beta_c$  and  $\beta_r$ . Table 3 shows the resulting pragmatic speaker probabilities under a low and a high  $\beta_i$  with no utterance cost.

Crucially, the reverse is not the case when color is the distinguishing dimension. Consider the speaker in the same context wanting to communicate the big red pin. The two best utterances for this purpose are *red* (.57) and *big red* (.52). In contrast to the results for the small blue pin, the redundant utterance does not increase the literal listener probability of inferring 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 Boolean 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 Boolean and continuous semantics in Table 2.

To better understand the consequences of continuous meanings in contexts like that depicted in Figure 1a, we visualize the results of varying  $x_{\text{size}}$  and  $x_{\text{color}}$  in Figure 2. The deterministic Boolean 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 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

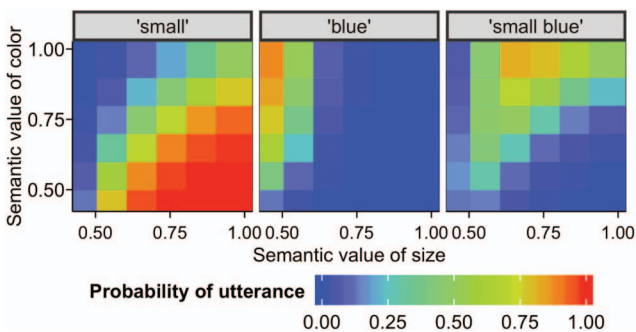


Figure 2. Probability of producing sufficient *small pin*, insufficient *blue pin*, and redundant *small blue pin* in contexts as depicted in Figure 1a (left), as a function of semantic value of color and size utterances (for  $\beta_i = 30$  and  $\beta_c = 0$ ). For a visualization of model behavior under varying  $\beta_r$ , see online supplementary material. See the online article for the color version of this figure.

emerges. This asymmetry is found in the warmer region of the “small blue” facet, characterized by values of  $x_{\text{size}}$  that are lower than  $x_{\text{color}}$ , with high values for  $x_{\text{color}}$ . 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 (Gatt et al., 2011; Pechmann, 1989). Note also that no difference in adjective *cost* is necessary for obtaining the overinformativeness asymmetry, though assuming a greater cost for size than for color does further increase the observed asymmetry.

**Simulation 2: Scene variation.** In the previous section, we showed that extending RSA with continuous adjective semantics gives rise to color-size asymmetries when the semantics of color adjectives is closer to deterministic Boolean truth-functions than size adjectives. When modifiers are noisy, adding “stricter” modifiers adds information. From this perspective, these additional modifiers are not *overinformative*; they are usefully redundant given the needs of the listener. Next, we show how the same mechanism accounts for why increased scene variation increases the probability that referring expressions are overmodified with color.

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).<sup>15</sup> Scene variation was manipulated across two experiments, which differed in the dimension necessary for unique reference (color was always redundant). In Experiment 1, only type was necessary (*fan* and *couch* in the low and high variation conditions in Figure 3, respectively). In Experiment 2, size and type were necessary (*big chair* and *small chair* in Figure 3, respectively). Across both experiments, lower rates of redundant color use were found in the low variation conditions (4% and 9%) than in the high variation conditions (24% and 18%). Here, we use simulations to explore the predictions that continuous semantics RSA—henceforth cs-RSA—makes for these situations.

Following Koolen et al. (2013), we considered any mention of color as a redundant mention. In Experiment 1, this includes simple redundant utterances like *blue couch* as well as complex redundant utterances like *small blue couch*. In Experiment 2, where size was necessary for unique reference, only the complex redundant utterance *small brown chair* was truly redundant (*brown chair* was insufficient, but still included in counts of color mention). Because object type was a distinguishing dimension, we introduce an additional semantic value  $x_{\text{type}}$ , which encodes how noisy nouns are. The results of simulating these conditions with parameters  $\beta_i = 30$ ,  $\beta_c = c(u_{\text{size}}) = c(u_{\text{color}}) = 1$ ,  $x_{\text{size}} = .8$ ,  $x_{\text{color}} = .999$ , and  $x_{\text{type}} = .9$  are shown in Figure 3, under the assumption that the cost of a two-word utterance  $c(u)$  is the sum of

<sup>15</sup> 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.

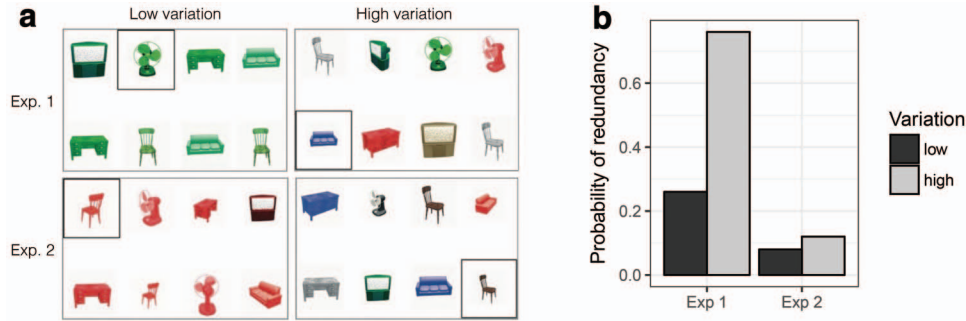


Figure 3. Visual contexts employed in experiments by Koolen et al. (2013) alongside RSA model predictions for the use of redundant modifiers in those contexts. (a) Contexts from Koolen et al. (2013)’s low variation (left column) and high variation (right column) conditions in Exp. 1 (top row) and Exp. 2 (bottom row). (b) Predicted probability of redundant color utterance in Koolen et al. (2013) conditions for  $\beta_i = 30$ ,  $\beta_c = c(u_{size}) = c(u_{color}) = 1$ ,  $x_{size} = .8$ ,  $x_{color} = .999$ ,  $x_{type} = .9$ . See the online article for the color version of this figure.

the costs of the one-word subutterances.<sup>16</sup> For both experiments, the model exhibits 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, this effect of scene variation is predicted by the model anytime the semantic values for size, type, and color are ordered as:  $x_{size} \leq x_{type} < x_{color}$ . If, on the other hand,  $x_{type}$  is greater than  $x_{color}$ , the probability of redundantly mentioning color is close to zero and does not differ between variation conditions (in those cases, color mention reduces, rather than adds, information about the target).

To further explore the scene variation effect predicted by RSA, we turn again to the left panel of 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. We can measure scene variation as 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 from the target, in the value of the insufficient feature divided by the total number of distractors  $n_{total}$ :

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

In this example context, there is one distractor that differs from the target in color (the big red pin) and there are two distractors in total. Thus, scene variation =  $\frac{1}{2} = .5$ . In general, this measure of 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  $\frac{n_{total}-1}{n_{total}}$ .<sup>17</sup>

Using the same parameter values as above, we generate model predictions for size-sufficient and color-sufficient contexts, manipulating scene variation by varying number of distractors (2, 3, or 4) and number of distractors that do not share the insufficient feature value. The resulting model predictions are shown in Figure 4. The predicted probability of redundant adjective use is largely (though not completely) correlated with scene variation. Redundant adject-

ive use increases with increasing scene variation when size is sufficient (and color redundant), but not when color is sufficient (and size redundant). The latter prediction depends, however, on the actual semantic value of color—with slightly lower semantic values for color, the model predicts small increases in redundant size use. In general: Increased scene variation is predicted to lead to a greater increase in redundant adjective use for less noisy adjectives.

RSA with a continuous semantics (cs-RSA) thus captures the qualitative effects of color-size asymmetry and scene variation in production of redundant expressions, and it makes quantitative predictions for both. Testing these quantitative predictions, however, will require more data. In the remainder of the article, we quantitatively evaluate cs-RSA on new data sets capturing the phenomena described in the Introduction (see Table 1): modifier type and scene variation effects on modified referring expressions, typicality effects on color mention, and the choice of taxonomic level in nominal choice.

### Experiment 1: Size and Color Modifiers Under Different Scene Variation Conditions

Adequately assessing the explanatory value of RSA with continuous semantics requires evaluating how well it does at predicting the probability of various types of utterances occurring in large data sets of naturally produced referring expressions. While we showed above that cs-RSA *qualitatively* predicts the pattern of overmodification under scene variation, we now test the model’s *quantitative* predictions more rigorously in an interactive web-based reference game paradigm. We then per-

<sup>16</sup> These parameter values were chosen merely for convenience in illustrating the qualitative model predictions. We reused values from the previous example, where possible, but also included a cost per word.

<sup>17</sup> 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. We choose to explore this aspect of variation here as a reasonable first step; RSA makes predictions for other kinds of variation that would be equally straightforward to test.



form a Bayesian data analysis to both assess how likely the model is to generate the observed data—that is, to obtain a measure of model quality—and to explore the posterior distribution of parameter values—that is, to understand whether the asymmetries in adjectives' semantic values and/or costs explored in the previous section are validated by the data.

## 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.<sup>18</sup> Data from another seven pairs who prematurely dropped out of the experiment and who could therefore not be compensated for their work, were also included. Here and in all other experiments reported in this article, participants' IP address was limited to U.S. 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 multiplayer interface (Hawkins, 2015). One participant was assigned the speaker role and one the listener role. Before continuing to the experiment, participants were required to correctly answer a series of questions about the experimental procedure (see [supplementary material](#) for details). On each trial, both participants saw the same array of objects in independently randomized locations. One of these objects was privately designated as the *target* object to the speaker, and marked by a thick border (see [Figure 5](#)). The speaker's task was to use an unrestricted chat box to send a message communicating the target to the listener, who subsequently clicked an object to make a response. Both participants then received feedback about whether the intended referent was selected and advanced to the next trial. 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. For natural interaction, we allowed both speakers and listeners to write freely in the chat window at any point, but listeners could only click on an object to advance to the next trial after the speaker sent an initial message. At the end of the experiments, participants completed a questionnaire in which they indicated whether their native language was

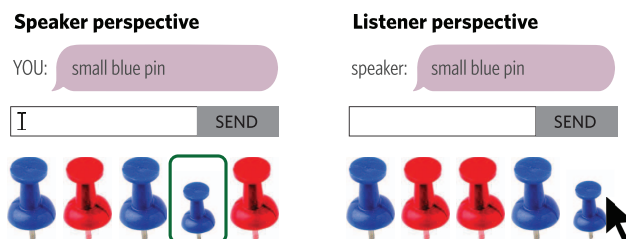


Figure 5. Example displays from the speaker's (left) and the listener's (right) perspective on a *size-sufficient* 4–2 trial. See the online article for the color version of this figure.

English, whether they thought their partner was human, and how much they liked their partner.

**Materials.** Participants proceeded through 72 trials. Of these, half were critical trials of interest and half were filler trials. On critical trials, we varied which feature was sufficient 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 5](#) 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 (two, three, or four) 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, because otherwise the redundant property would have been sufficient for uniquely establishing reference; that is, 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 5](#)) but always differed in type across trials. Each object type could occur in two different sizes and two different colors. We used photo-realistic objects of intuitively fairly typical colors. The 36 different object types and the colors they could occur with are listed in [Table 4](#).

Fillers were target trials from Experiment 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 the Method section of Experiment 3 for a description of these materials.

**Data preprocessing and exclusion.** We collected data from 2,177 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

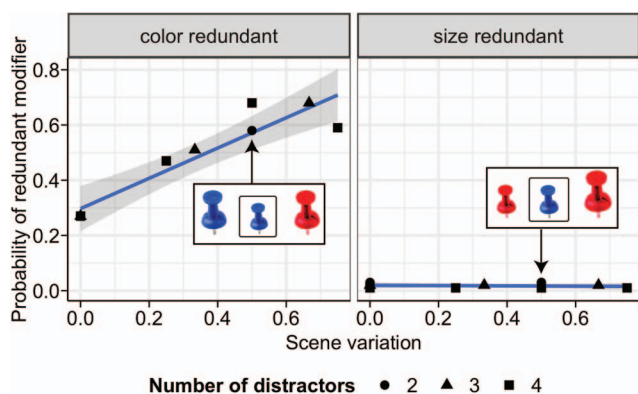


Figure 4. Predicted probability of redundant utterance (*small blue pin*) as a function of scene variation when size is sufficient (and color redundant, left) and when color is sufficient (and size redundant, right), for  $\beta_1 = 30$ ,  $\beta_c = c(u_{\text{size}}) = c(u_{\text{color}}) = 1$ ,  $x_{\text{size}} = .8$ ,  $x_{\text{color}} = .999$ . Linear smoothers overlaid. See the online article for the color version of this figure.

<sup>18</sup> We aim to pay Mechanical Turk workers at a rate of \$12–\$14 per hour.

Table 4  
*Experiment 1 Items and the Colors They Appeared In*

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

each trial, classifying the produced utterance as an instance of a *color-only* mention (e.g., *blue pin*), a *size-only* mention (e.g., *big pin*), or a redundant *color-and-size* mention (e.g., *big blue pin*). To this end we applied a semiautomatic data preprocessing procedure in which a script first checked whether the speaker's utterance contained a color or size term. In a second step, one of the authors (CG) manually checked and, if necessary, corrected the automatic classification. If no classification was possible, the trial was excluded. After exclusions, 2,076 cases entered the analysis. See the [online supplementary materials](#) for details on the preprocessing procedure.

## Behavioral Results

Proportions of redundant *color-and-size* utterances are shown in [Figure 6](#) alongside model predictions (to be explained further below). 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 by conducting a single 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 do not share the insufficient property value with the target), and the interaction between the two.<sup>19</sup> All predictors were centered before entering the analysis. The model included the most sophisticated random effects structure that allowed the model to converge: by-speaker and by-item random intercepts.

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. That is, while the *color-sufficient* condition indeed showed a scene variation effect—and as far as we know, this is the first demonstration of an effect of scene variation on redundant size use—this effect was tiny compared with that of the *size-sufficient* condition.<sup>20</sup>

## Model Evaluation

In order to evaluate RSA with continuous semantics we conducted a Bayesian data analysis. This allowed us to 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 ( $x_{\text{color}}$ ,  $x_{\text{size}}$ ) and an informativeness parameter  $\beta_i$ . 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 10,000 and lag of 10 to draw 2,000 samples from the joint posteriors on the five free parameters.

Point-wise maximum a posteriori (MAP) estimates of the model's posterior predictives for just redundant utterance probabilities are shown alongside the empirical data in [Figure 6](#). In addition, 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 plotted against all empirical utterance proportions in [Figure 7](#). At this level, the model achieves a correlation of  $r = .99$ . Looking at results additionally on the by-item level yields a correlation of  $r = .85$  (this correlation is expected to be lower both because each item contains less data, and because we did not provide the model any means to refer differently to, e.g., *combs* and *pins*). The model thus does a very good job of capturing the quantitative patterns in the data.

Posteriors over parameters are shown in [Figure 8](#). 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

<sup>19</sup> All mixed effects analyses reported in this paper were conducted with the `lme4` package (Bates, Mächler, Bolker, & Walker, 2015) in R (R Core Team, 2017).

<sup>20</sup> In order to address convergence issues with `lmer` when specifying the maximal 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 maximal 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],  $p(\beta > 0) = .96$ ).

two parameters. That is, size modifiers are inferred to be noisier than color modifiers. The high inferred  $\beta_i$  (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 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.

## Discussion

In this section we reported a new dataset of freely produced referring expressions that replicated the well-documented color-size asymmetry in redundant adjective use, 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 cs-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 with color. The effect is that color adjectives are more informative than size adjectives when controlling for the number of distractors that each would rule out under a Boolean semantics. Asymmetries in the cost of the adjectives were not attested, and would 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 cs-RSA: Scene variation leads to a greater increase in overmodification with less noisy modifiers because

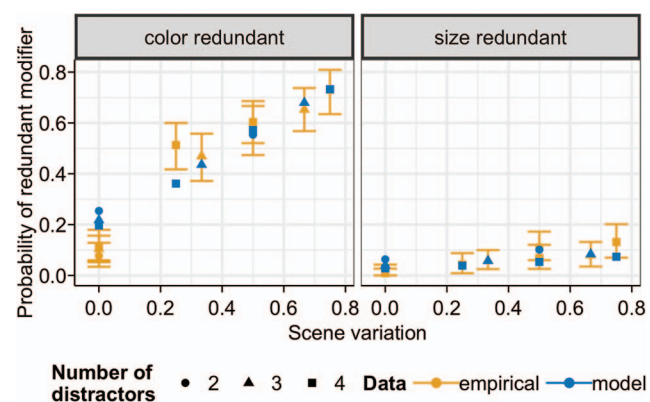


Figure 6. Empirical redundant utterance proportions (orange) alongside point-wise maximum a posteriori (MAP) estimates of the RSA model's posterior predictives for redundant utterance probability (blue) as a function of scene variation in the color redundant (left) and size redundant (right) condition. Here and in all following plots, error bars indicate 95% bootstrapped confidence intervals. See the online article for the color version of this figure.

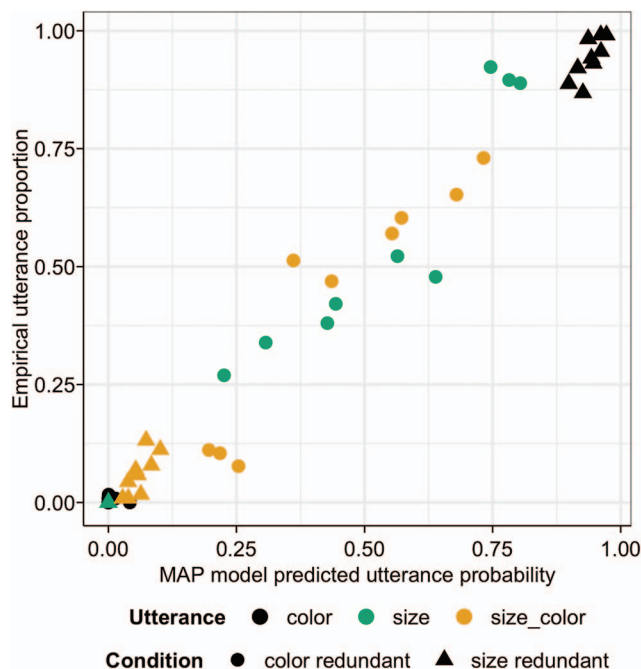


Figure 7. Scatterplot of empirical utterance proportions against point-wise maximum a posteriori (MAP) estimates of the RSA model's posterior predictives. Each dot represents a condition mean. See the online article for the color version of this figure.

these modifiers (colors) on average provide more information about the target.

While we defer a broader discussion of the possible psychological and linguistic interpretations of continuous semantic values to the General Discussion, 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; Pechmann, 1989). 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.<sup>21</sup> In addition, color as a property has been claimed to be inherently salient in a way that size is not (Arts et al., 2011; van Gompel et al., 2019). Finally, color adjectives are rated as less subjective than size adjectives (Scontras, Degen, & Goodman, 2017). All of this evidence suggests that the use of size adjectives may be more likely to vary across speakers and contexts than color.

Critically, our explanation of these phenomena departs from those offered by previous theories. Pechmann (1989) was the first to take the color-size asymmetry as evidence for speakers following an *in-*

<sup>21</sup> 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 discussion of the complex semantics of color terms, see Kennedy and McNally (2010), Rothschild and Segal (2009), and Szabo (2001). For our purposes, it suffices that one can give a color judgment but not a size judgment for an object presented in isolation.

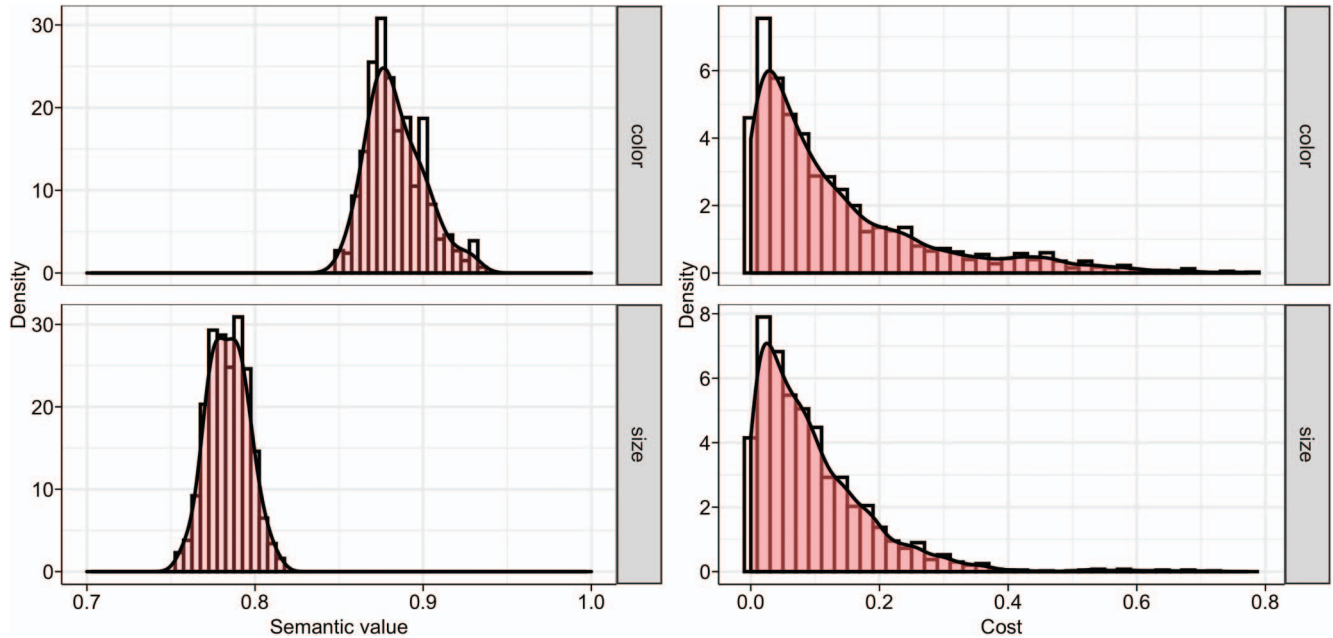


Figure 8. Posterior model parameter distributions for semantic value (left column) and cost (right column), separately for color (top row) and size (bottom row) modifiers. Maximum a posteriori (MAP)  $x_{\text{size}} = 0.79$ , 95% highest density interval (HDI) = [0.76, 0.80]; MAP  $x_{\text{color}} = 0.88$ , 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]. See the online article for the color version of this figure.

*cremental* strategy of object naming. That is, 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. Another explanation appeals to *saliency* considerations: Speakers may 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; Paraboni, van Deemter, & Masthoff, 2007; Rubio-Fernandez, 2016). 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).

Finally, 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’s Quantity maxim, 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 probabilistic referential overspecification model (PRO; van Gompel et al., 2019) 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. In PRO, the uniquely distinguishing property (if there is one) is first selected deterministically. In additional steps, additional properties are 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. This model successfully captures speakers’ overmodification patterns in contexts with one target and two distractors, in the choice of two properties (color, size) and three properties (color, size, border presence). While the PRO model—the most state-of-the-art computational model of human production of modified referring expressions—can capture the basic color-size asymmetry, it does not straightforwardly account for the more subtle systematicity with which the preference to overmodify with color changes based on scene variation or object typicality, which we turn to next.

## Experiment 2: Color Typicality in Modified Referring Expressions

Our modeling results in Experiment 1 raise interesting questions regarding the status of the inferred semantic values: 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 idiosyncrasies such as the typicality effects discussed in the introduction: Speakers are more likely to redundantly produce modifiers that denote atypical rather than typical object features, that is, 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).

A natural first step toward explaining typicality effects is to introduce a more nuanced semantics for nouns in our model. In particular, we could imagine a continuous semantics in which *banana* fits better (i.e., has a semantic value closer to 1 for) the yellow banana than the brown, and fits the brown better than the blue; specific such hypothetical values are shown in the first row of Table 5. Let us further assume that modifying the noun with a color adjective leads to uniformly high semantic values close to 1 for those objects that a simple truth-conditional semantics would return “true” for (see diagonal in Table 5) and a very low semantic value close to 0 for any utterance applied to any object that a simple truth-conditional semantics would return “false” for.

The effect of running the speaker model forward with the standard literal listener treatment of the values in Table 5 for the three contexts in Figure 9, where *banana* is the strictly sufficient utterance for unique reference (i.e., color is redundant under the standard view) is as follows: with  $\beta_i = 12$  and  $\beta_c = 5$ ,<sup>22</sup> the resulting speaker probabilities for the minimal utterance *banana* are .95, .29, and .04, 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 .05, .71, and .96, respectively. That is, redundant color mention increases with decreasing semantic value of the simple *banana* utterance.

This shows that cs-RSA can predict typicality effects if the semantic fit of the noun (and hence also of color-noun compounds) to an object is modulated by typicality. The reason the typicality

effect arises is that, with the hypothetical values we assumed, the gain in informativeness between using the unmodified *banana* and the modified *COLOR banana* is greater in the blue than in the yellow banana case.

This example is somewhat oversimplified. In practice, speakers sometimes mention an object’s color without mentioning the noun. In the contexts presented in Figure 9 this does not make much sense because there is always a competitor of the same color present. In contrast, in the left contexts of Figure 10, 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.

An additional, more theoretically fraught, simplification concerns where typicality can enter into the semantics and how composition proceeds. In the above, we have assumed that the semantic value of the modified expression is uniformly high, which is qualitatively what is necessary (and, as we will see below, empirically correct) in order for the typicality effects to emerge. However, there is no straightforward way to compositionally derive such uniformly high values from the semantic values of the nouns and the semantic values of the color modifiers, which we have not yet discussed. Indeed, compositional semantics of graded meanings is a well-known problem for theories of modification (Kamp & Partee, 1995; Osherson & Smith, 1981). Rather than try to solve it here, we note that RSA works at the level of whole utterances. Hence, if we can reasonably measure the semantic fit of each utterance to each possible referent, then cs-RSA will make predictions for production without the need to derive the semantic values compositionally. That is, if we can measure the typicality of the phrase *blue banana* for a banana, we do not need to derive it from *blue*, *banana*, and a theory of composition. This separates pragmatic aspects of reference, which are the topic of this paper, from issues in compositional semantics, which are not; hence we will take this approach for experimentally testing the predictions of relaxed semantics RSA for typicality effects.

The stimuli for Experiment 1 were specifically designed to be realistic objects with low color-diagnostics, so they did not include objects with low typicality values or large degrees of variation in typicality. This makes the dataset from Experiment 1 not well-suited for investigating typicality effects.<sup>23</sup> 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 cs-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

Table 5  
Hypothetical Semantic Values for Utterances (Rows) as Applied to Objects (Columns)

| Utterance            | Yellow banana | Brown banana | Blue banana | Other      |
|----------------------|---------------|--------------|-------------|------------|
| <i>banana</i>        | <b>.9</b>     | <b>.35</b>   | <b>.1</b>   | .01        |
| <i>yellow banana</i> | <b>.99</b>    | .01          | .01         | .01        |
| <i>brown banana</i>  | .01           | <b>.99</b>   | .01         | .01        |
| <i>blue banana</i>   | .01           | .01          | <b>.99</b>  | .01        |
| <i>other</i>         | .01           | .01          | .01         | <b>.99</b> |

Note. Values where a Boolean semantics would return “true” are bolded.

<sup>22</sup> The results hold qualitatively for any informativeness weight  $> 1$  and any cost weight  $> 0$ .

<sup>23</sup> We did elicit typicality norms for the items in Experiment 1 and replicated the previously documented typicality effects on the four items that did exhibit variation in typicality. See the online supplementary materials for details.



Figure 9. 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. See the online article for the color version of this figure.

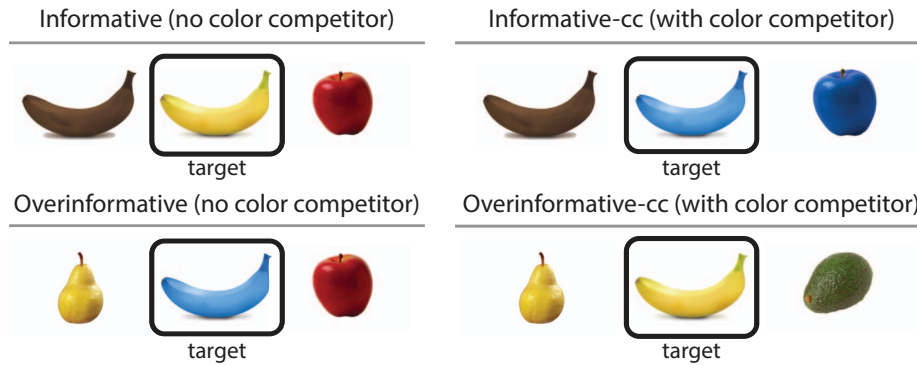


Figure 10. Examples of the four different context conditions in Experiment 2. They differed in the presence of an object of the same type (informative vs. overinformative) and in the presence of another object of the same color as the target (with color competitor vs. without color competitor). The thick border marks the intended referent. See the online article for the color version of this figure.

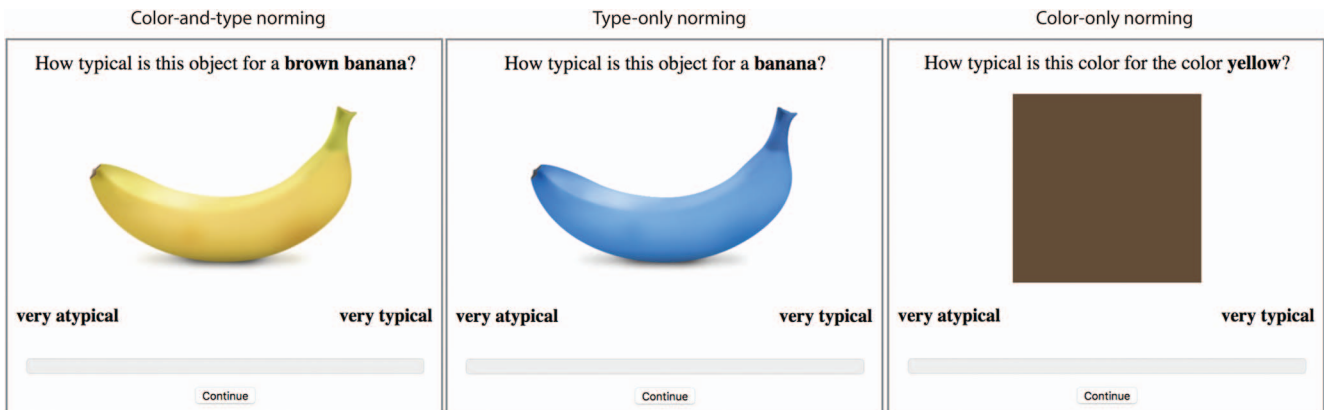


Figure 11. Example stimuli exemplifying the three different typicality norming studies. See the online article for the color version of this figure.

modifier type 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.

## Method

**Participants.** We recruited 61 pairs of participants (122 participants total) over Amazon's Mechanical Turk who were each paid \$1.70 for their participation.

**Procedure.** The procedure of the reference game was identical to that of Experiment 1.

**Materials.** Each participant completed 42 trials. In this experiment, there were no filler trials, because 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.

The stimuli were selected from seven color-diagnostic food items (apple, avocado, banana, carrot, pear, pepper, tomato), which

Table 6  
Overview of the Typicality Norming Studies for Experiment 2

| Utterances | Example       | Images      | Participants | Trials | Items | Excluded participants |
|------------|---------------|-------------|--------------|--------|-------|-----------------------|
| Adj noun   | Yellow banana | Object      | 174          | 110    | 484   | 14                    |
| Noun       | Banana        | Object      | 75           | 90     | 154   | 1                     |
| Adj        | Yellow        | Color patch | 110          | 90     | 176   | None                  |

Note. Column “Items” contains the number of unique utterance-object pairs that we elicited responses for.

all occurred in a typical, midtypical 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 10). A context was “overinformative” when mentioning the type of the item, for example, 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”), color alone was not sufficient. In the informative conditions, color and type mention were necessary for unambiguous reference. Again, one context type did and one did not 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.

**Data preprocessing and exclusion.** We collected data from 1,974 trials. The utterance produced on each trial was classified as belonging to one of the following categories: *type-only* (e.g., *banana*), *color-and-type* (e.g., *yellow banana*), and *color-only* (e.g., *yellow*). Referring expressions that could not be classified were excluded. See the [online supplementary materials](#) for further

details on exclusion criteria and the data preprocessing procedure. Overall, 1,827 utterances entered the analysis.

## Typicality Norming

In order to test for typicality effects on the production data and to evaluate cs-RSA’s performance, we collected empirical typicality values for each utterance/object pair in 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 11, left). 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.; see Figure 11, middle). 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; see Figure 11, right).

On each trial of the *type* or *color-and-type* studies, 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. A summary of the three typicality norming studies is shown in Table 6.<sup>24</sup>

Slider values were coded as falling between 0 (*very atypical*) and 1 (*very typical*). For each utterance-object combination, we computed mean typicality ratings. As an example, the means for the banana items and associated color patches are shown in Table 7. The values exhibit the same gradient as those hypothesized for the purpose of the example in Table 5. The means for all items are visualized in Figure 12. Mean typicality values for utterance-object pairs obtained in the norming studies are used in the analyses and visualizations in the following.

Table 7  
Mean Typicalities for Banana Items

| Utterance            | Banana items |            |            | other |
|----------------------|--------------|------------|------------|-------|
|                      | yellow       | brown      | blue       |       |
| <i>banana</i>        | <b>.98</b>   | <b>.66</b> | <b>.42</b> | .05   |
| <i>yellow banana</i> | <b>.97</b>   | .30        | .15        | .05   |
| <i>brown banana</i>  | .22          | <b>.91</b> | .15        | .04   |
| <i>blue banana</i>   | .16          | .15        | <b>.92</b> | .06   |
| <i>yellow</i>        | <b>.77</b>   | .05        | .06        | .09   |
| <i>brown</i>         | .11          | <b>.87</b> | .01        | .12   |
| <i>blue</i>          | .06          | .06        | <b>.92</b> | .07   |

Note. Combinations where Boolean semantics would return “true” are marked in boldface.

<sup>24</sup> The typicality elicitation procedure we employed here is somewhat different from that employed by, who asked their participants “How typical is this color for this object?” We did this because the semantic values that enter into the RSA model are best conceptualized as the typicality of an object as an instance of an utterance, rather than a feature-category relation. See the [online supplementary materials](#) for a comparison of our question and the Westerbeek et al. (2015) question as applied to typicality norms for the items in Experiment 1. In general, the TYPE-object values are highly correlated with the Westerbeek et al. (2015) question values.

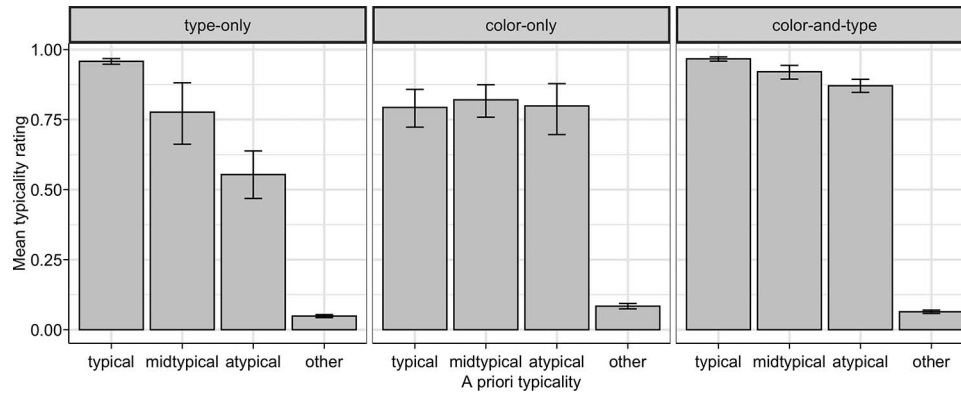


Figure 12. Mean typicality ratings for the three norming studies (type-only, color-only, color-and-type). The results are categorized according to the objects' a priori typicality as determined by the experimenters (yellow banana = typical, brown banana = midtypical, blue banana = atypical). The category *other* comprises all utterance-object combinations where a Boolean semantics would return false (e.g., a pepper). Error bars indicate bootstrapped 95% confidence intervals.

## Behavioral Results

Proportions of type-only (*banana*), color-and-type (*yellow banana*), color-only (*yellow*), and other (*funky carrot*) utterances are shown in Figure 13a 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 invariant to the distractor items?

To address these questions we conducted a mixed effects logistic regression predicting color 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 11, middle) 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 most sophisticated random effects structure that allowed the model to converge.

We found 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 ( $\beta = -4.17$ ,  $SE = 0.45$ ,  $p < .0001$ ), 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 ( $\beta = -5.56$ ,  $SE = 0.33$ ,  $p < .0001$ ). Finally, there was a main effect of color competitor presence, such that color mention was more likely when a color competitor was absent ( $\beta = 0.71$ ,  $SE = 0.16$ ,  $p < .0001$ ). 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.

## Model Evaluation

We evaluated the cs-RSA model on the obtained production data from Experiment 2. In particular, we were interested in using model comparison to address the following issues: First, can RSA using elicited typicality as the semantic values account for quantitative details of the production data? Second, are typicality values sufficient, or is there additional utility in including a noise offset determined by the type of modifier, as was used in the previous section? Third, does utterance cost explain any of the observed production behavior.

While the architecture of the model remained the same as that of the model that treated the data from Experiment 1, we briefly review the minor necessary changes, some of which we already mentioned at the beginning of this section. These changes concerned the semantic values and the cost function.<sup>25</sup>

**Lexicon.** Previously, we considered only three utterance alternatives: *color*, *size*, and *color-size*, collapsing over the precise values these took on. Here, we no longer collapse over these values, including 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 combi-

<sup>25</sup> See Table 10 for an overview of the models reported in the article.



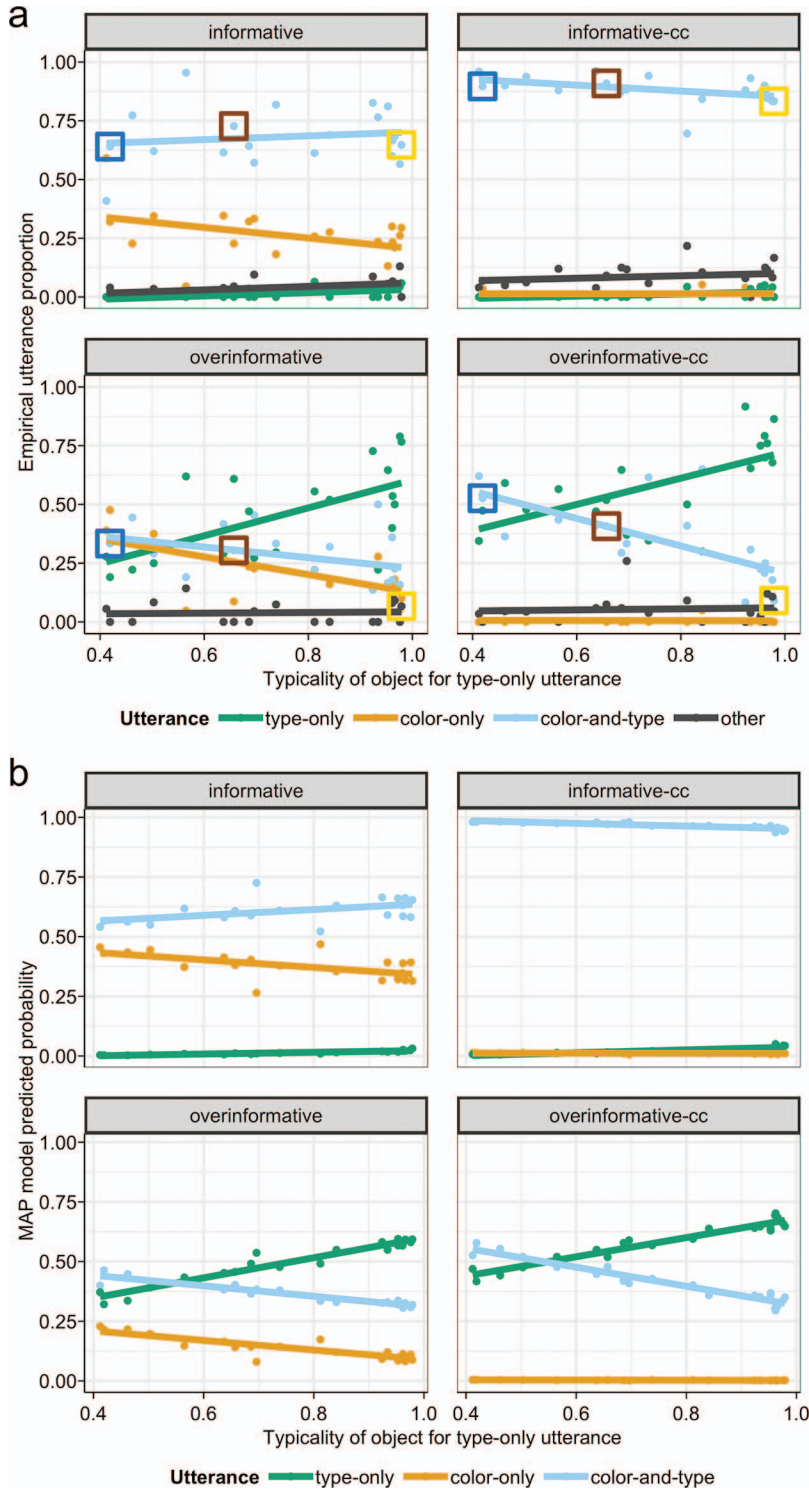


Figure 13. (a) Empirical utterance proportions in Experiment 2 and (b) MAP model predicted utterance probabilities for each target as a function of mean object typicality for the type-only utterance (e.g., *banana*). Color indicates utterance type: type-only (*banana*), color-only (*yellow*), color-and-type (*yellow banana*), and other (*funky carrot*). Facets indicate conditions (*informative* vs. *overinformative*, color competitor present (*cc*) or absent). Modified utterance data points for the banana items are circled in the banana's respective color in (a). See the online article for the color version of this figure.

nation 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 above (rather than inferred as done for Experiment 1, to avoid overfitting). For any given context, we assumed the utterance alternatives that correspond to the individually present features and their combinations. For example, for the “overinformative-cc” context in Figure 10, the set of utterance alternatives was *yellow*, *green*, *pear*, *banana*, *avocado*, *yellow pear*, *yellow banana*, and *green avocado*.

**Semantics.** We compared several choices of semantics for the model. In the full *fixed plus empirical semantics* version, we introduced a parameter  $\beta_{\text{fixed}}$  interpolating between the empirically elicited typicality values ( $\beta_{\text{fixed}} = 0$ ) and the inferred type-level values as employed in the model that treated the data from Experiment 1 ( $\beta_{\text{fixed}} = 1$ ). The type-level values again consisted of one value for color terms and another for type terms, which are multiplied when the terms are composed in an utterance. In a lesioned *empirical semantics* version, we set  $\beta_{\text{fixed}} = 0$  and only used the empirical values. Conversely, in a lesioned *fixed semantics* version, we set  $\beta_{\text{fixed}} = 1$  and only used the inferred type-level values. This allowed us to perform a nested model comparison, because the latter models are special cases of the first.

**Cost function.** For the purpose of evaluating the model that treated the data from Experiment 1 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:

$$c(u) = \beta_F \cdot p(u) + \beta_L \cdot l(u) \quad (7)$$

Here,  $p(u)$  is negative log 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);  $\beta_F$  is a weight on frequency; and  $\beta_L$  is a weight on length. Both  $p(u)$  and  $l(u)$  were normalized to fall into the interval  $[0, 1]$ .<sup>26</sup> 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 with a simpler baseline in which utterances were assumed to have no cost.

**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 observed 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 correcting for the flexibility due to extra parameters (the “Bayesian Occam’s Razor”). 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

Table 8  
*Marginal Log Likelihood for Each Model*

| Cost      | Semantics   |             |                      |
|-----------|-------------|-------------|----------------------|
|           | Empirical   | Fixed       | Fixed plus empirical |
| Empirical | −1390.3 (4) | −1295.1 (6) | −1251.4 (7)          |
| Fixed     | −1350.3 (4) | −1272.6 (6) | −1238.5 (7)          |
| None      | −1352.2 (2) | −1223.0 (4) | <b>−1188.4 (5)</b>   |

*Note.* Best model is in bold. Parentheses indicate number of free parameters.

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 8. 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 ( $BF = 3.7 \times 10^{48}$  for fixed costs,  $BF = 2.1 \times 10^{60}$  for empirical costs, and  $BF = 1.4 \times 10^{71}$  for no cost) and fixed-only counterparts ( $BF = 6.5 \times 10^{14}$  for fixed costs,  $BF = 1.0 \times 10^{19}$  for empirical costs, and  $BF = 1.06 \times 10^{15}$  for no cost). In comparison, additional cost-related parameters were not justified, with  $BF = 5.7 \times 10^{21}$  for no cost compared with fixed cost and  $BF = 2.1 \times 10^{27}$  for no cost compared with 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 13. The model successfully reproduces the empirically observed typicality effects in all four experimental conditions, with a reasonably good quantitative agreement. The interpolation weight between the fixed and empirical semantic values  $\beta_{\text{fixed}}$  (see Figure 14) is in the intermediate range: This provides evidence that a noisy truth-conditional semantics as employed in Experiment 1 is justified, but that taking into account graded category membership or typicality in an utterance’s final semantic value is also necessary.

There is one major, and interesting, divergence 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 14, to understand the pattern. In particular, the utterance type level semantic value of type is inferred to be systematically higher than that of color, capturing that type utterances are less noisy than color

<sup>26</sup> Note that we changed the sign on frequency, which means that values closer to 1 in the normalized space reflect greater cost on both the length and the frequency dimension.

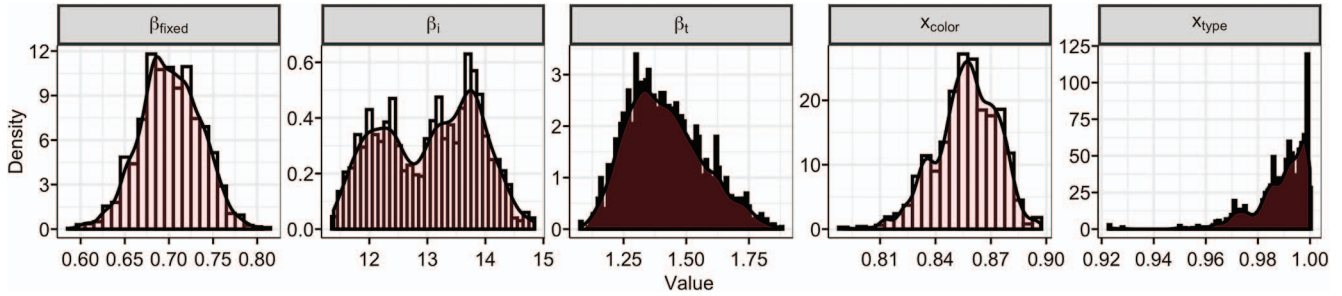


Figure 14. For best model, posterior model parameter distributions for interpolation weight on fixed versus empirical semantics  $\beta_{\text{fixed}}$ , informativity weight  $\beta_i$ , typicality weight  $\beta_t$ , and utterance type level semantic values for color  $x_{\text{color}}$  and type  $x_{\text{type}}$ . Maximum a posteriori (MAP)  $\beta_{\text{fixed}} = 0.69$ , 95% highest density interval (HDI) = [0.64, 0.77]; MAP  $\beta_i = 13.74$ , HDI = [11.58, 14.37]; MAP  $\beta_t = 1.34$ , HDI = [1.19, 1.75]; MAP  $x_{\text{color}} = 0.86$ , HDI = [0.82, 0.89]; MAP  $x_{\text{type}} = 0.998$ , HDI = [0.97, 1.00]. See the online article for the color version of this figure.

utterances.<sup>27</sup> 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 and undesirable 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.

## Discussion

In Experiment 2, we demonstrated that cs-RSA predicts color typicality effects in the production of referring expressions with only minimal extensions to support finer-grained prediction of item-by-item data (see Table 10 in the General Discussion for a more extensive comparison of the model details across sections). This suggests that the dynamics at work in the choice of color versus 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—very atypical—object of the same type present, simply because the literal listener is expected to prefer the typical referent strongly enough.

Model comparison demonstrated the need for assuming a semantics that interpolates between a noisy truth-conditional semantics as employed in Experiment 1 and empirically elicited typicality values. This may reflect semantic knowledge that goes beyond graded category membership, additional effects of compositionality, or perhaps simply differences between our empirical typicality measure and the “semantic fit” expected by RSA models. Perhaps surprisingly, we replicated the result from Experiment 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 cs-RSA

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 Experiment 3

| Domain | Super     | Targets         | Mean sublength (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)           |

<sup>27</sup> Interestingly, the inferred semantic value for color is very similar in absolute terms to that in Experiment 1.

Table 10

Overview of the Best-Performing Models Used for the Three Different Production Datasets Color/Size (Experiment 1), Color Typicality (Experiment 2), and Nominal Choice (Experiment 3)

| Model feature             | Color/size model  | Color typicality model  | Nominal choice model  |
|---------------------------|---|---|---|
| Semantic values           | At type-level (inferred)  | At type-level (inferred) + object-level (elicited)                          | At object-level (elicited)  |
| Size of lexicon $L(u, o)$ | 8 (all combinations of <i>size</i> and <i>color</i> )   | 814 (1 for each utterance-object pair)                                      | 51 (1 for each utterance-object pair)   |
| Set of alternatives       | 8 contextually available feature combinations (size, color)                                   | 8 or 9 contextually available feature combinations (type, color)            | 3 target alternatives (level of reference: <i>sub</i> , <i>basic</i> , <i>super</i> ) |
| Cost                      | Type-level (color and size)   | Not necessary   | empirical (length and frequency)  |
| Free parameters           | $x_{\text{color}}, x_{\text{size}}, \beta_{\text{c(color)}}, \beta_{\text{c(size)}}, \beta_i$ | $x_{\text{color}}, x_{\text{type}}, \beta_l, \beta_r, \beta_{\text{fixed}}$ | $\beta_F, \beta_L, \beta_i, \beta_r$  |

*Note.* Parameter names =  $x_{\text{color}}$ : semantic value of color;  $x_{\text{size}}$  = semantic value of size;  $\beta_{\text{c(color)}}$  = cost of color;  $\beta_{\text{c(size)}}$  = cost of size;  $\beta_i$  = weight on informativity;  $\beta_l$  = weight on cost (as estimated by utterance frequency);  $\beta_r$  = weight on cost (as estimated by utterance length);  $\beta_i$  = weight on elicited typicality values;  $\beta_{\text{fixed}}$  = interpolation weight between fixed type-level and empirical semantic values.

provides a good account of referring expression production more generally.

### Experiment 3: Taxonomic Level in Unmodified Referring Expressions

In this section we investigate whether cs-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. This choice of reference level is 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 differing in word count. Still, we hypothesized that similar factors may contribute: 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).

In order to evaluate cs-RSA for nominal choice, we proceeded as with Experiment 2: 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 15). We also elicited typicality ratings for object-utterance combinations, which entered the model as the semantic values via the lexicon. We then conducted Bayesian data analysis, as in previous sections, for model comparison. Our key insight is that compared to a traditional Boolean semantics where class labels (e.g., *dog*) are strictly true or false of objects, a continuous semantics incorporating knowledge of typicality more successfully predicts preferences for taxonomic level of reference. That is, cs-RSA accurately predicts, from Gricean principles, that speakers will increase their preference for subordinate terms (e.g., *penguin*) when an object is a more atypical example of the basic-level (*bird*) and prefer basic-level terms (e.g., *bird*) when those objects are more atypical examples of the superordinate level (*animal*).

## Method

**Participants.** We recruited 58 pairs of participants (116 participants total, the same participants as in Experiment 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 Experiment 1.<sup>28</sup> Participants proceeded through 72 trials. Of these, half were critical trials of interest and half were filler trials (the critical trials from Experiment 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 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. These target items were randomly assigned distractor items which were selected from three different context conditions, corresponding to different communicative pressures (see Figure 15). 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

<sup>28</sup> A separate earlier data set was reported at the annual meeting of the Cognitive Science Society (Graf et al., 2016), and serves as a close replication of the reported study.

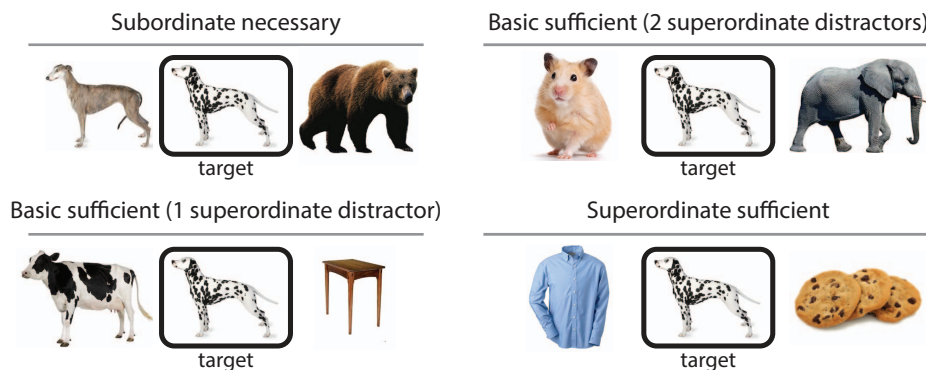


Figure 15. Example contexts in which different levels of reference are necessary for establishing unique reference to the target marked with a thick border: subordinate necessary (*dalmatian*); basic sufficient (*dog*) and subordinate possible (*dalmatian*); superordinate sufficient (*animal*) and basic or subordinate possible (*dog, dalmatian*). See the online article for the color version of this figure.

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.

**Data preprocessing and exclusion.** We collected data from 2,193 critical trials. Each referring expression was classified as containing the target's correct *sub*(ordinate, e.g., *dalmatian*), *basic* (e.g., *dog*), or *super*(ordinate, e.g., *animal*) level term, or excluded if classification was not possible. See the [online supplementary materials](#) for details on exclusion criteria and the preprocessing procedure. After exclusions and preprocessing, 1,872 cases entered the analysis.

**Typicality norming.** In order to test for typicality effects on the production data and to evaluate cs-RSA's performance, we again collected empirical typicality values for each utterance/object pair. See the [online supplementary materials](#) for details.

## Behavioral Results and Discussion

Proportions of sub, basic, and super level utterances are shown in [Figure 16](#). Overall, super level mentions are highly dispreferred (<2%), so we focus in this section only on predictors of sub- over basic level mentions. The clearest pattern of note is that sub level mentions are only preferred in the most constrained context that necessitates the sub level mention for unique reference (e.g., target: *dalmatian*, distractor: *greyhound*; see [Figure 15a](#)). Nevertheless, even in these contexts there is a non-negligible proportion of basic level mentions (28%). This includes cases of using just the basic level term (6%, e.g., *dog* for the *German Shepherd* when one of the distractors was a *greyhound*, an atypical dog, akin to the unmodified cases in the informative conditions discussed for Experiment 2) as well as basic level terms with additional modifying material (22%). In the remaining contexts, where the sub and basic level are equally informative, there is a clear preference for the basic level. Finally, 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. Because the maximal model with by-speaker and by-item slopes for all fixed effects did not converge, we simplified the random effects structure, including only by-speaker and by-item random intercepts.

*Frequency* was coded as the difference between the sub and the basic level's log frequency, as extracted from the Google Books Ngram English corpus ranging from 1960 to 2008. Speakers prefer more frequent words over less frequent ones (Oldfield & Wingfield, 1965).

*Length* was coded as the ratio of the subordinate to the basic level term's length in characters. Even among one-word utterances, speakers prefer shorter ones over longer ones (Degen, Franke, & Jäger, 2013; Rohde, Seyfarth, Clark, Jäger, & Kaufmann, 2012). We used the mean number of characters in the utterances participants produced. For example, the *minivan*, when referred to at the subordinate 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."

*Typicality* was coded as the ratio of the target's sub to basic level label typicality.<sup>29</sup> That is, the higher the ratio, the more typical the object was for the sub level label compared with the basic level; or in other words, a higher ratio indicates that the object was relatively atypical for the basic label compared with the sub label. For instance, the *panda* was relatively atypical for its basic level "bear" (mean rating 0.75) compared with the sub level term "panda bear" (mean rating 0.98), which resulted in a relatively *high* typicality ratio. We predicted that subordinate terms may be preferred when an object is a particularly good instance of that term or a particularly bad instance of the basic level term, compared with the other objects in the context.

<sup>29</sup> Typicalities were elicited in a separate norming study that was identical in procedure to that of Experiment 1. See the [online supplementary materials](#) for details about the study.

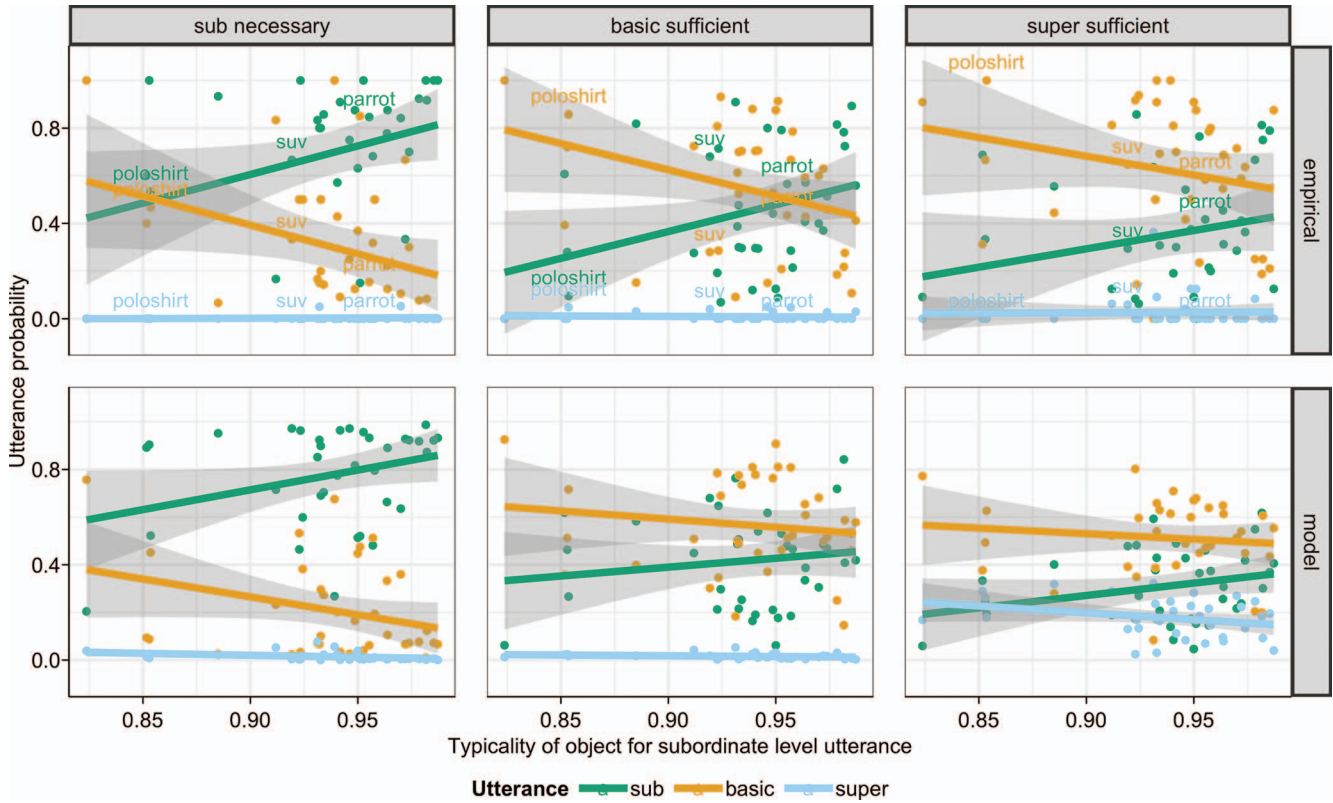


Figure 16. Top: utterance proportions for each target item across different informativeness conditions as a function of the object’s subordinate level typicality. Example target items *polo shirt* (basic: *shirt*, super: *clothes*), *SUV* (basic: *car*, super: *vehicle*), and *parrot* (basic: *bird*, super: *animal*) that were characteristic of relatively low to relatively high sub typicality items are labeled explicitly. Bottom: MAP model predicted utterance probabilities. See the online article for the color version of this figure.

*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 16.

The log odds of mentioning the sub level term were greater in the *sub necessary* condition than in either of the other two conditions ( $\beta = 2.11$ ,  $SE = .17$ ,  $p < .0001$ ), 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.<sup>30</sup> 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 with 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 dis-

preferred compared with “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 with its sub level competitor, mirroring the results from Experiment 2.

<sup>30</sup> 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 do not simply revert to the basic level when no basic-level distractor is in context.

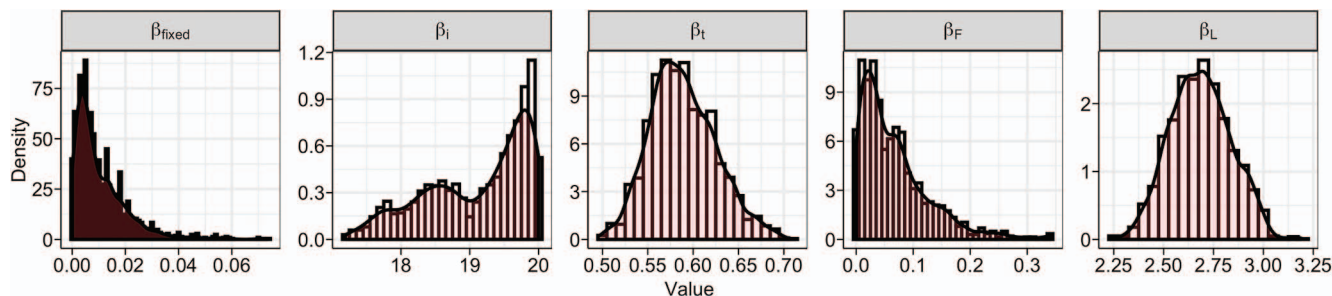


Figure 17. Posterior model parameter distributions for interpolation weight on fixed versus empirical semantics  $\beta_{\text{fixed}}$ , informativity weight  $\beta_i$ , typicality weight  $\beta_t$ , frequency cost weight  $\beta_F$ , and length cost weight  $\beta_L$ . Maximum a posteriori (MAP)  $\beta_{\text{fixed}} = 0.004$ , 95% highest density interval (HDI) = [0.000, 0.03];  $\beta_i = 19.8$ , HDI = [17.71, 20.0]; MAP  $\beta_t = 0.57$ , HDI = [0.53, 0.67]; MAP  $\beta_F = 0.02$ , HDI = [0.00, 0.19]; MAP  $\beta_L = 2.69$ , HDI = [2.42, 2.99]. See the online article for the color version of this figure.

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 effect of frequency. Support for this hypothesis comes from studies comparing written and spoken language, which have 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).<sup>31</sup>

## Model Evaluation

We evaluated cs-RSA on the production data from Experiment 3. The architecture of the model is identical to that of the model that treated the data from Experiment 2. The only difference is the set of alternative utterances.<sup>32</sup> Whereas the models from the previous sections treated all individual features and feature combinations present in the display as utterance alternatives, for computational efficiency we now consider only the three different levels of reference to the target as alternatives, that is, subordinate (*dalmatian*), basic (*dog*), and superordinate (*animal*). So, even when a German Shepherd is present as a distractor, *German Shepherd* is not considered an alternative utterance for the dalmatian target. This has minimal effects on model predictions as long as *German Shepherd* has low semantic fit to the dalmatian target.

In Experiment 2, we tested which of three different semantics was most justified—the empirically elicited typicality semantics, a fixed semantics with type-level semantic values, and one that combined both. Here, the relevant comparison is between a fixed Boolean noun semantics (i.e., 1 if the object belongs to the given class label, and 0 otherwise) and the empirically elicited typicality semantics. Again, we introduced a parameter that interpolated between these semantic values. Additionally, we evaluated which cost function was best supported by the data: The one defined in Equation (7) (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. Priors were again  $\beta_i \sim \mathcal{U}(0, 20)$ ,  $\beta_F \sim \mathcal{U}(0, 5)$ ,  $\beta_L \sim \mathcal{U}(0, 5)$ ,  $\beta_t \sim \mathcal{U}(0, 5)$ . Despite the greater number of parameters associated with adding the cost function, the model that includes nonzero costs was preferred compared to its no-cost counterpart ( $BF = 2.8 \times 10^{77}$ ). Posteriors over parameters are shown in Figure 17. First, we observe that the semantic interpolation value was highly skewed toward 0, strongly indicating that empirical typicality values strongly improve model performance over a Boolean baseline. Second, 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 probability with which it is produced in this paradigm.<sup>33</sup>

MAP model predictions are shown alongside empirical utterance proportions in Figure 16. The correlation between empirical utterance proportions and the model’s MAP predictions at the level of targets, utterances, and conditions was  $r = .86$ . Further collapsing across targets yields a correlation of  $r = .95$ . The model captures the qualitative patterns well, though it somewhat overpredicts subordinate level and underpredicts basic level choices. It also accounts for the strong preference against superordinate level mentions. The reason for this is that the semantics for each utterance (e.g., *dalmatian*, *dog*, *animal*) is taken from the empirically elicited typicality values for each utterance-object pair. The

<sup>31</sup> 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 (subvs. 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],  $p(\beta < 0) = 1$ ).

<sup>32</sup> See Table 10 for an overview of the models reported in the article.

<sup>33</sup> 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.

target images used in this experiment were generally rated as less typical instances of the superordinate level term than of the basic or subordinate level term (see [online supplementary materials](#) for details). This difference is enough to lead to a general bias against using the superordinate level term, especially when coupled with the fact that superordinate terms tend to be costlier than basic level terms.

## General Discussion

In this article we have provided a unified account of referring expression choice that solves a long-recognized puzzle for rational theories of language use: Why do speakers' referring expressions often and systematically exhibit seeming overinformativeness? We have shown here that by allowing contextual utterance informativeness to be computed with respect to a continuous (or noisy) rather than a Boolean semantics, utterances that seem overinformative can in fact be sufficiently informative. This happens when what seems like the *prima facie* sufficiently informative utterance is in fact noisy and may lead a literal listener astray; adding redundancy ensures successful communication. This simple modification to the RSA approach allowed us to capture: The basic well-documented asymmetry for speakers to be more likely to redundantly use color adjectives than size adjectives; the interaction between sufficient dimension and scene variation in the probability of redundancy; and typicality effects in both color modifier choice and noun choice.

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 appropriately informative. 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 straightforwardly extendable to any instance of definite referring expressions of the sort we have examined here.

## Comparison of Model Components Across Experiments

In order to address the possible concern that the different models employed are too different from one another to be comparable, we begin by providing an overview of the parts of the model that remained the same or differed across experiments. While the core architecture with relaxed semantics remained constant throughout the paper, some peripheral components were adjusted to accommodate the aims of the different experiments. These different choices are fully consistent with one another, and many of them were justified against alternatives via model comparison. We have provided an overview of the best-fitting RSA models for each of the three reported production data sets in [Table 10](#).

Most prominently, Experiments 2 and 3 aimed to predict patterns of reference via typicality at the *object-level*; in those cases the model thus required semantic values for each utterance-object pair in the lexicon. While these values could have in principle been

inferred from the data, as we inferred the two type-level values in Experiment 1, it would have introduced a large number of additional parameters (see *size of lexicon*). Instead, we addressed this problem by empirically eliciting these values in an independent task and introducing a single free concentration parameter  $\beta$ , that modulated their strength. In the case of Experiment 2, we found that the best-fitting model smoothly integrated these empirical values with type-level values used in Experiment 1.

The need to make object-level predictions also drove decisions about what to use as the cost function and the set of alternative utterances. For instance, in Experiment 3 we could have inferred the cost of each noun but this again would have introduced a large number of free parameters and risked overfitting. Instead we used the empirically estimated *length* and *frequency* of each word. For Experiment 2, we tested models both using fixed costs for each modifier as in Experiment 1 and empirical length and frequency costs as in Experiment 3, but our model comparison showed that neither sufficiently improved the model's predictions.

Finally, the set of alternative utterances differed slightly across the three experiments for computational reasons. Because Experiment 1 collapsed over the particular levels of size and color, it was practical to consider all utterances in the lexicon for every target. In Experiment 2 and Experiment 3, however, the space of possible utterances was large enough that this exhaustive approach became impractical. We noticed that the probability of using some utterances (e.g., “table” to refer to a Dalmatian) was low enough that we could prune the utterance space to only those that could plausibly apply to the objects in context without substantially altering the model's behavior. Future work must address how predictions may change as more complex referring expressions outside the scope of this paper enter the set of alternatives (e.g., the option of combining adjectives with nominal expressions, as in *the cute, spotted dog*). In the following we discuss a number of intriguing questions that this work raises and avenues for future research it suggests.

## Comparison With PRO

While a detailed comparison of cs-RSA with PRO ([van Gompel et al., 2019](#)), the hitherto most state-of-the-art computational model of human production of modified referring expressions, is outside the scope of this article, we include some comparative remarks here. PRO has the advantage of being computationally more efficient than cs-RSA, partly because it aims to be an algorithmic-level model, which may be of importance for applications. PRO may further have the advantage of having fewer parameters, though this is a bit harder to evaluate in general: While PRO as applied to the choice of color and size in principle involves two parameters, *s* (size preference) and *e* (overspecification eagerness), in the 2019 article the maximum likelihood parameter values are estimated on each of the experimental conditions separately, effectively resulting in six parameters. In the extension to three properties, this results in 14 parameters, and this number increases further as more properties and conditions are added. If the parameter values had been estimated on all conditions jointly, which is what we did in the evaluation of cs-RSA instead of separately, then indeed PRO would have fewer effective parameters than cs-RSA, though it is unclear how this would affect the data fit.



One advantage of the cs-RSA approach is greater generality. For instance, it is not immediately clear how PRO should be extended to contexts that vary in the number and nature of distractors, where empirical overmodification proportions change, but the PRO predictions would not. We see this as one of the great strengths of cs-RSA: Scene variation effects fall out of the model directly. Finally, we have proposed here a way to account for typicality effects. PRO may be able to accommodate typicality effects if its preference parameters can be made to be sensitive to typicality. In general, a systematic comparison and possible combination of these models is an important next step.

### “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.”

This conception of (over-)informativeness assumes that all modifiers are born equal—that is, 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 with 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., Rubio-Fernandez, 2016; Sedivy, 2003).

But this raises a question: what counts as a truly overinformative utterance under RSA with a continuous semantics? Cs-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 6: 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 nonpurple 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.

### Comprehension

While the account proposed in this article 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 article 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, that is, 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.

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

$$P_{L_1}(o|u) \propto P_{S_1}(u|o) \cdot P(o) \quad (8)$$

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 article, 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 Experiments 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.

**Online processing.** The question that has typically been asked about the online processing of redundant utterances is this: Do redundant utterances, compared with 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 nonlinguistic 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 with its mini-

mally 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) (Experiment 2) found that listeners were slower to identify the target referent in response to redundant compared with minimal utterances. It is useful to examine the stimuli they used. In the Engelhardt et al. (2011) study, there was only one distractor that varied in type, that is, 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 and Katsos (2013) study, the “overinformative” conditions contained displays of four objects which differed in type. Stimuli were selected via a production pretest: 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 cs-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.

### Continuous Semantics

The crucial component of the model that allows for capturing “overinformativeness” effects is the continuous semantics. In this section, we consider the nature of these continuous semantic values. Readers already convinced of the utility of a continuous semantics are invited to skip to the next section.

For the purpose of Experiment 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 data sets from Experiments 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 overfitting, and for the dataset from Experiment 2 we found further that a combination of a relaxed (noisy) truth-conditional semantics and the empirically elicited continuous 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 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 versus 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, for example, in color modifiers being more noisy, with concomitant effects on production, that is, redundant color modifiers should become less likely. This is indeed what Viethen, van Vesseem, Goudbeek, and Krahmer (2017) found. Similarly, van Gompel, Gatt, Krahmer, and van Deemter (2014) found that the asymmetry in redundant use of color versus size disappeared when participants were shown displays with very noticeable size contrasts and barely noticeable color contrasts.

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, that is, 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 with 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, Degen, and Goodman (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.

What is highlighted by the above discussion is that we have been using the term “semantics” at a fairly high level, to refer to conventional aspects of meaning that are relatively stable across contexts—in RSA these are the representations on which the literal listener performs computations. These real valued representations could be primitives arising in lexical representations and threading through composition, which would constitute a fundamentally different basic semantics than has often been assumed. Alternatively, the necessary real values could arise by adding the right kind of use or world-knowledge related noise to a standard Boolean truth-conditional semantics. Most minimally, intensional parameters of a standard semantics could be set stochastically. Whatever their source, these continuous values provide the right basis for capturing the production choices explored in this article. When to assume a relaxed semantics, and what the implications of such a relaxation are for other semantic and pragmatic phenomena, are questions for future research.

## 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 (Brown-Schmidt & Heller, 2014; Clark & Murphy, 1982; Horton & Keysar, 1996). 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. 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 noncomprehension, when responding to listener questions, or when speaking to strangers (Baker, Gill, & Cassell, 2008). Audience design has played an especially large role in explanations of typicality effects. For example, Huettig and Altmann (2011) 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 like color are rapidly activated and aid visual search. Similarly, Arts, Maes, Noordman, and Jansen (2011) showed that overspecified expressions result in faster referent identification. Nevertheless, such benefits might simply be a happy coincidence and speakers might not, in fact, be deliberately designing their utterances for their addressees.

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 to 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).<sup>34</sup> 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 toward 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.

## Other Factors That Affect Redundancy

RSA with a continuous semantics 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

<sup>34</sup> But see Franke and Degen (2016) for an evaluation of the distribution of listener and speaker types in Quantity inferences.

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, both at the conceptual level of content or property selection and at the level of linguistic realization. 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 color adjectives should occur before size adjectives, thereby inverting the well-documented adjective ordering preferences for English (Bloomfield, 1933; Scontras et al., 2017; Sproat & Shih, 1991). Pechmann (1989) does observe some, but not many, instances of this. An interesting test case for the incrementality hypothesis are cases where adjective ordering preferences are weak. Fukumura (2018) reports one such case in which speakers prefer to order more discriminative and more available properties before less discriminative and less available ones, highlighting incrementality as an important factor affecting the choice of referring expression. On the other hand, it is unclear how incrementality—whether in linguistic realization or in content selection—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 available (i.e., 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.

There are other ways incrementality could play a role in modifier choice. For example, mentioning the color adjective may buy the speaker time when the noun is hard to retrieve. This predicts that in languages with postnominal adjectives, where this delay strategy cannot be used for noun planning, redundant color mention should be less frequent; indeed, this is what Rubio-Fernandez (2016) shows for Spanish.<sup>35</sup> In sum, the incrementality of language production clearly affects the choice of referring expression; the ways in which considerations of incrementality should be incorporated in RSA are to be explored further (see Cohn-Gordon, Goodman, & Potts, 2019, for a step in this direction).

### Extensions to Other Language Production Phenomena

In this article 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 cs-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 a choice in which aspects of an event or entity to mention (content selection) and how to realize that content linguistically, including in the domain of reference (pronouns, names, definite descriptions with postnominal 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.

Similarly, the approach outlined here might be extended to the case of nonrestrictive modifiers. In these cases, the modifier is not used to distinguish a target referent from possible competitors. Instead, the speaker may intend to communicate an aspect of an already contextually established referent, as in *sit by the newly painted table*, where the speaker is warning the listener not to put their elbows on the table (Dale & Reiter, 1995); or *Forrest looks at the massive crowd*, where the speaker is commenting on the extraordinary size of the crowd (Hahn, Degen, Goodman, Jurafsky, & Futrell, 2018). In these cases, the table and the crowd are contextually given referents; the modifiers *newly painted* and *massive* are used to highlight informative aspects of these referents. The approach proposed in this article could be extended to these cases by allowing the speaker to be informative with respect to goals other than getting the listener to infer the intended referent. For instance, the speaker may want to be informative with respect to the goal of highlighting task-relevant properties of contextually given referents, as in the newly painted table case.

More generally, the approach should extend to any phenomenon that affords a choice between a more or less specific utterance. Whenever the more specific utterance adds relevant information compared with the less specific one, it should be produced. This is related to surprisal based theories of production like Uniform Information Density (UID, A. Frank & Jae-

<sup>35</sup> Rubio-Fernández herself argues for an efficiency-based account of the Spanish/English asymmetry in color overmodification: “Color adjectives are a more efficient cue in prenominal position than in post-nominal position because in the latter case the hearer’s visual search is initially guided by the noun” (Rubio-Fernandez, 2016, p. 9). Cs-RSA as presented here does not predict positional asymmetries in modifier use. However, see Cohn-Gordon et al. (2019) for an incremental version of RSA that captures the asymmetry reported by Rubio-Fernandez (2016) as the result of incremental rather than global reasoning about utterance informativeness.

ger, 2008; Jaeger, 2006; Jaeger, 2010; Levy & Jaeger, 2007), where speakers have been found to be more likely to omit linguistic signal if the underlying meaning or syntactic structure is highly predictable to the listener. 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.

### 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. We take this work to provide evidence that, rather than being wastefully overinformative, speakers are usefully redundant.

### References

- Arts, A., Maes, A., Noordman, L., & Jansen, C. (2011). Overspecification facilitates object identification. *Journal of Pragmatics*, *43*, 361–374. <http://dx.doi.org/10.1016/j.pragma.2010.07.013>
- Baker, R., Gill, A. J., & Cassell, J. (2008). Reactive redundancy and listener comprehension in direction-giving. In D. Schlangen & B. A. Hockey (Eds.), *9th SIGdial Workshop on Discourse and Dialogue* (pp. 37–45). Columbus, OH: Association for Computational Linguistics. <http://dx.doi.org/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*, 1–48. <http://dx.doi.org/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*, 237–266. <http://dx.doi.org/10.1080/09541440143000050>
- Bergen, L., Levy, R., & Goodman, N. (2016). Pragmatic reasoning through semantic inference. *Semantics and Pragmatics*, *9*, 1–46. <http://dx.doi.org/10.3765/sp.9.20>
- Bernardy, J.-P., Blanck, R., Chatzikyriakidis, S., & Lappin, S. (2018). A compositional Bayesian semantics for natural language. In M. Sinha & T. Dasgupta (Eds.), *Proceedings of the First International Workshop on Language Cognition and Computational Models* (pp. 1–10). Santa Fe, NM: Association for Computational Linguistics.
- Bloomfield, L. (1933). *Language*. New York, NY: Holt.
- Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology*, *22*, 1482–1493.
- Brown, P., & Dell, G. (1987). Adapting production to comprehension: mention of instruments. *Cognitive Psychology*, *472*, 441–472.
- Brown, R. (1958). *Words and things*. Washington, DC: Free Press.
- 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. <http://dx.doi.org/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*, 1–28. <http://dx.doi.org/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. [http://dx.doi.org/10.1057/9780230524125\\_2](http://dx.doi.org/10.1057/9780230524125_2)
- Clark, H. H., & Murphy, G. L. (1982). Audience design in meaning and reference. *Advances in Psychology*, *9*, 287–299. [http://dx.doi.org/10.1016/S0166-4115\(09\)60059-5](http://dx.doi.org/10.1016/S0166-4115(09)60059-5)
- Cohen, B., & Murphy, G. L. (1984). Models of concepts. *Cognitive Science*, *8*, 27–58.
- Cohn-Gordon, R., Goodman, N. D., & Potts, C. (2019). An incremental iterated response model of pragmatics. *Proceedings of the Society for Computation in Linguistics*, *2*, Article10.
- Dale, R. (1989). Cooking up referring expressions. *Proceedings of the 27th Annual Meeting on Association for Computational Linguistics*, 68–75. <http://dx.doi.org/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. *Journal of Pragmatics*, *49*, 78–106. <http://dx.doi.org/10.1016/j.pragma.2013.01.004>
- Degen, J., Franke, M., & Jäger, G. (2013). Cost-based pragmatic inference about referential expressions. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 376–381). Austin, TX: Cognitive Science Society.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics.
- Emberson, L. L., Loncar, N., Mazzei, C., Treves, I., & Goldberg, A. E. (2019). The blowfish effect: Children and adults use atypical exemplars to infer more narrow categories during word learning. *Journal of Child Language*, *46*, 938–954.
- Engelhardt, P. E., Bailey, K., & Ferreira, F. (2006). Do speakers and listeners observe the Gricean Maxim of Quantity? *Journal of Memory and Language*, *54*, 554–573. <http://dx.doi.org/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*, 304–314. <http://dx.doi.org/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. *Proceedings of The 30th Annual Meeting of the Cognitive Science Society* (pp. 939–944). Austin, TX: 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*, 1–25. <http://dx.doi.org/10.1371/journal.pone.0154854>
- Franke, M., & Jäger, G. (2016). Probabilistic pragmatics, or why Bayes’ rule is probably important for pragmatics. *Zeitschrift für Sprachwissenschaft*, *35*, 3–44. <http://dx.doi.org/10.1515/zfs-2016-0002>
- Fukumura, K. (2018). Ordering adjectives in referential communication. *Journal of Memory and Language*, *101*, 37–50. Retrieved from <http://dx.doi.org/10.1016/j.jml.2018.03.003>
- Gatt, A., van Gompel, R. P. G., Krahmer, E., & van Deemter, K. (2011). Non-deterministic attribute selection in reference production. *Proceedings of the workshop on production of referring expressions: Bridging the gap between empirical, computational and psycholinguistic approaches to reference (PRE-CogSci 2011)*, Boston.
- 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*, 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.

- Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20, 818–829. <http://dx.doi.org/10.1016/j.tics.2016.08.005>
- Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*, 5, 173–184.
- Goodman, N. D., & Stuhlmüller, A. (n.d.). *The design and implementation of probabilistic programming languages*. Retrieved 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, 313–338.
- Hahn, M., Degen, J., Goodman, N., Jurafsky, D., & Futrell, R. (2018). An information-theoretic explanation of adjective ordering preferences. In C. Kalish, M. Rau, J. Zhu, & T. Rogers (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 1766–1771). Austin, TX: Cognitive Science Society.
- Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics* (pp. 159–166). Pittsburgh, PA: Association for Computational Linguistics.
- Hawkins, R. X. D. (2015). Conducting real-time multiplayer experiments on the web. *Behavior Research Methods*, 47, 966–976.
- Hawkins, R. X. D., Stuhlmüller, A., Degen, J., & Goodman, N. D. (2015). Why do you ask? Good questions provoke informative answers. *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 878–883). Austin, TX: Cognitive Science Society.
- Herrmann, T., & Deutsch, W. (1976). *Psychologie der Objektbenennung*. Verlag, Bern: Huber.
- Hoffmann, J., & Ziessler, C. (1983). Objektidentifikation in künstlichen Begriffshierarchien [Object identification in artificial conceptual hierarchies]. *Zeitschrift für Psychologie mit Zeitschrift für angewandte Psychologie*, 191, 135–167.
- Horton, W., & Keysar, B. (1996). When do speakers take into account common ground? *Cognition*, 59, 91–117.
- Huetig, 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*, 64, 122–145. <http://dx.doi.org/10.1080/17470218.2010.481474>
- Jaeger, T. F. (2006). *Redundancy and reduction in spontaneous speech*, Unpublished doctoral dissertation. Stanford University, Stanford, CA.
- Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, 61, 23–62.
- Jeschaniak, 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, 824–843. <http://dx.doi.org/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, 248–277. <http://dx.doi.org/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, 243–275.
- Kamp, H., & Partee, B. (1995). Prototype theory and compositionality. *Cognition*, 57, 129–191.
- 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, 12002–12007.
- Kennedy, C., & McNally, L. (2005). Scale structure, degree modification, and the semantics of gradable predicates. *Language*, 81, 345–381. <http://dx.doi.org/10.1353/lan.2005.0071>
- Kennedy, C., & McNally, L. (2010). Color, context, and compositionality. *Synthese*, 174, 79–98.
- Koolen, R., Gatt, A., Goudbeek, M., & Krahmer, E. (2011). Factors causing overspecification in definite descriptions. *Journal of Pragmatics*, 43, 3231–3250.
- Koolen, R., Goudbeek, M., & Krahmer, E. (2013). The effect of scene variation on the redundant use of color in definite reference. *Cognitive Science*, 37, 395–411. <http://dx.doi.org/10.1111/cogs.12019>
- Krahmer, E., van Erk, S., & Verleg, A. (2003). Graph-based generation of referring expressions. *Computational Linguistics*, 29, 53–72.
- Levinson, S. C. (1983). *Pragmatics*. New York, NY: Cambridge University Press.
- Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106, 1126–1177. <http://dx.doi.org/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 information processing systems* (Vol. 19, pp. 849–856). Cambridge, MA: MIT Press.
- Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. New York, NY: Wiley.
- Maes, A., Arts, A., & Noordman, L. (2004). Reference management in instructive discourse. *Discourse Processes: A Multidisciplinary Journal*, 37, 117–144.
- Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco, CA: Freeman.
- Mitchell, M. (2013). Typicality and object reference. *Proceedings of the 35th Annual Meeting of the Cognitive Science Society* (pp. 3062–3067). Austin, TX: Cognitive Science Society.
- Murphy, G. L., & Smith, E. E. (1982). Basic-level superiority in picture categorization. *Journal of Verbal Learning and Verbal Behavior*, 21, 1–20.
- Nadig, A. S., & Sedivy, J. C. (2002). Evidence of perspective-taking constraints in children’s on-line reference resolution. *Psychological Science*, 13, 329–336. <http://dx.doi.org/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, 273–281.
- Olson, D. R. (1970). Language and thought: Aspects of a cognitive theory of semantics. *Psychological Review*, 77, 257–273.
- Osherson, D. N., & Smith, E. E. (1981). On the adequacy of prototype theory as a theory of concepts. *Cognition*, 9, 35–58. [http://dx.doi.org/10.1016/0010-0277\(81\)90013-5](http://dx.doi.org/10.1016/0010-0277(81)90013-5)
- Paraboni, I., van Deemter, K., & Masthoff, J. (2007). Generating referring expressions: Making referents easy to identify. *Computational Linguistics*, 33, 229–254. <http://dx.doi.org/10.1162/coli.2007.33.2.229>
- Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics*, 27, 89–110.
- Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543). Pittsburgh, PA: Association for Computational Linguistics.
- Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 2227–2237). New Orleans, Louisiana: Association for Computational Linguistics. <http://dx.doi.org/10.18653/v1/N18-1202>
- 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, Switzerland: Springer. <http://dx.doi.org/10.1007/978-3-319-17064-0>
- R Core Team. (2017). *R: A language and environment for statistical computing* [Computer software manual]. Vienna, Austria: R Core Team.
- Rohde, H., Seyfarth, S., Clark, B., Jäger, G., & Kaufmann, S. (2012). Communicating with cost-based implicature: A game-theoretic approach to ambiguity. In S. Brown-Schmidt, J. Ginzburg, & S. Larsson (Eds.), *Proceedings of the 16th Workshop on the Semantics and Pragmatics of Dialogue* (pp. 107–116). Retrieved from [http://events.illc.uva.nl/sem-dial/proceedings/sem-dial2012\\_seinedial\\_proceedings.pdf](http://events.illc.uva.nl/sem-dial/proceedings/sem-dial2012_seinedial_proceedings.pdf)
- Rosch, E. H. (1973). Natural categories. *Cognitive Psychology*, 4, 328–350. [http://dx.doi.org/10.1016/0010-0285\(73\)90017-0](http://dx.doi.org/10.1016/0010-0285(73)90017-0)
- Rosch, E. H., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8, 382–439. [http://dx.doi.org/10.1016/0010-0285\(76\)90013-X](http://dx.doi.org/10.1016/0010-0285(76)90013-X)
- Rothschild, D., & Segal, G. (2009). Indexical predicates. *Mind and Language*, 24, 467–493. <http://dx.doi.org/10.1111/j.1468-0017.2009.01371.x>
- Rubio-Fernandez, P. (2016). How redundant are redundant color adjectives? An efficiency-based analysis of color overspecification. *Frontiers in Psychology*. Advance online publication. <http://dx.doi.org/10.3389/fpsyg.2016.00153>
- Scontras, G., Degen, J., & Goodman, N. D. (2017). Subjectivity predicts adjective ordering preferences. *Open Mind: Discoveries in Cognitive Science*, 1, 53–65. [http://dx.doi.org/10.1162/opmi\\_a\\_00005](http://dx.doi.org/10.1162/opmi_a_00005)
- Sedivy, J. C. (2003). Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. *Journal of Psycholinguistic Research*, 32, 3–23.
- Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, 128, 302–319. <http://dx.doi.org/10.1016/j.cognition.2013.02.013>
- Sproat, R., & Shih, C. (1991). The cross-linguistic distribution of adjective ordering restrictions. In C. Georgopoulos & R. Ishihara (Eds.), *Interdisciplinary approaches to language* (pp. 565–593). Amsterdam, the Netherlands: Springer.
- Szabo, Z. G. (2001). Adjectives in context. In I. Kenesei & R. Harnish (Eds.), *Perspectives on semantics, pragmatics, and discourse* (pp. 119–146). Amsterdam, the Netherlands: John Benjamins. <http://dx.doi.org/10.1075/pbns.90.12gen>
- Tanaka, J. W., & Taylor, M. (1991). Object categories and expertise: Is the basic-level in the eye of the beholder? *Cognitive Psychology*, 23, 457–482. [http://dx.doi.org/10.1016/0010-0285\(91\)90016-H](http://dx.doi.org/10.1016/0010-0285(91)90016-H)
- van Gompel, R. P., van Deemter, K., Gatt, A., Snoeren, R., & Krahmer, E. J. (2019). Conceptualization in reference production: Probabilistic modeling and experimental testing. *Psychological Review*, 126, 345–373. <http://dx.doi.org/10.1037/rev0000138>
- van Gompel, R. P., Gatt, A., Krahmer, E., & van Deemter, K. (2014, September). *Overspecification in reference: Modelling size contrast effects*. Poster presented at AMLaP 2014, Edinburgh, UK.
- 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). Santa Fe, New Mexico: Association for Computational Linguistics.
- Viethen, J., van Vesseem, T., Goudbeek, M., & Krahmer, E. (2017). Color in reference production: The role of color similarity and color codability. *Cognitive Science*, 41, 1493–1514. <http://dx.doi.org/10.1111/cogs.12387>
- Walker, M. A. (1993). *Informational redundancy and resource bounds in dialogue*. Unpublished doctoral dissertation. University of Pennsylvania, Philadelphia, PA.
- Westerbeek, H., Koolen, R., & Maes, A. (2015). Stored object knowledge and the production of referring expressions: The case of color typicality. *Frontiers in Psychology*. Advance online publication. <http://dx.doi.org/10.3389/fpsyg.2015.00935>
- Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological review*, 114, 245–272. <http://dx.doi.org/10.1037/0033-295X.114.2.245>
- Zadeh, L. A. (1965). Fuzzy sets. *Information and control*, 8, 338–353.

Received March 19, 2019

Revision received December 10, 2019

Accepted December 11, 2019 ■