Modeling cross-linguistic production of referring expressions

Brandon Waldon
Stanford University
bwaldon@stanford.edu

Judith Degen
Stanford University
jdegen@stanford.edu

Abstract

We present a novel probabilistic model of referring expression production, synthesizing recent analyses proposed within the Rational Speech Act (RSA) framework (Frank and Goodman, 2012). Our model makes incremental utterance choice predictions (Cohn-Gordon et al. 2018a; Cohn-Gordon et al. 2018b) and assumes a non-deterministic semantics for adjectives in referring expressions (Degen et al. 2020). The model captures previously attested production patterns in reference game experiments, including English speakers' tendency to produce redundant color adjectives more frequently than redundant size adjectives, as well as Spanish speakers' tendency to employ redundant color adjectives less frequently than English speakers. We report the predictions made by the model under various parameter regimes, motivating future empirical work.[1]

1 Using language to refer

A key communicative use of language is to refer. Understanding the constraints on referring expression production has therefore been a key enterprise in experimental and computational psycholinguistics alike (Pechmann, 1989; Sedivy, 2003; Gatt et al., 2011; van Deemter et al., 2012; Dale and Reiter, 1995). Here we focus on reference to objects presumed to be in visual common ground between speaker and listener.

Figure 1 contains two theoretically-relevant types of referring contexts which will be the focus of this paper. Their respective names—the **size-sufficient** (SS) scene and the **color-sufficient** (CS) scene—derive from expectations of how pragmatically competent speakers can use language to unambiguously establish reference to the target object $o_{\text{small\_blue}}$ (highlighted by green border). Grice (1975) proposed that in order to recover speaker meaning, listeners employ interpretive heuristics that can be formulated in terms of assumptions about how cooperative speakers behave in conversation, including that they should be as informative, but no more informative, than required. This has been interpreted as an expectation that speakers produce the minimally informative expressions that meet the standards of communicative sufficiency in context. In the SS scene, one can establish reference to $o_{\text{small\_blue}}$ using just a size adjective plus a head noun (e.g. *the small pin*). In the CS scene, one can refer to that object using just a color adjective plus a head noun (e.g. *the blue pin*).

Contra what we might expect in light of the above discussion, speakers routinely produce redundant adjectival modifiers in referential contexts (Pechmann, 1989; Nadig and Sedivy, 2002; Maes et al., 2004; Engelhardt et al., 2006; Arts et al., 2011; Koolen et al., 2011). For example, speakers produce *the small blue pin* to refer to $o_{\text{small\_blue}}$ in the SS scene, where the modifier *blue* is an instance of redundant color modification. *The small blue pin* in the CS scene is an instance of redundant size modification, which is much more rarely attested.

In addition, rates of redundant modification appear to vary cross-linguistically. Languages such as Spanish—in which modification tends to occur post-nominally—exhibit lower rates of redundant color modification than does English, in which the canonical adjective placement is pre-nominal (Rubio-Fernández, 2016). As Rubio-Fernández (2016) and Cohn-Gordon et al. (2018b) discuss, this result suggests the need to design theories of referring expression production that are sensitive to the linear order of words within those expres-

[1]We thank the audience at the interActive Language Processing Lab at Stanford (ALPS), Vera Gribanova, and Beth Levin for helpful feedback and discussion. We also gratefully acknowledge three anonymous SCiL reviewers and the Spanish judgments provided by four informants: Evelyn Rocio Fernandez-Lizarraga, Sabrina Grimberg, Adolfo Hermosillo, and Erika Petersen.

Proceedings of the Society for Computation in Linguistics (SCiL) 2021, pages 206-215.
Held on-line February 14-19, 2021

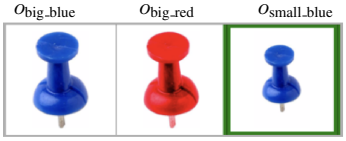
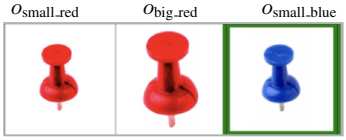
Utterances	Size-sufficient (SS) scene $o_{\text{big\_blue}}$ $o_{\text{big\_red}}$ $o_{\text{small\_blue}}$ 	Color-sufficient (CS) scene $o_{\text{small\_red}}$ $o_{\text{big\_red}}$ $o_{\text{small\_blue}}$ 
English	*blue pin, red pin, big pin, small pin, big blue pin, big red pin,* ***small blue pin***	*blue pin, red pin, big pin, small pin, small red pin, big red pin,* ***small blue pin***
Spanish -postnom.	*pin blue, pin red, pin big, pin small, pin blue big, pin red big,* ***pin blue small***	*pin blue, pin red, pin big, pin small, pin red small, pin red big,* ***pin blue small***
Spanish -split	*pin blue, pin red, pin big, pin small, big pin blue, big pin red,* ***small pin blue***	*pin blue, pin red, pin big, pin small, small pin red, big pin red,* ***small pin blue***
Spanish -postnom. -conj.	*pin blue, pin red, pin big, pin small, big pin blue, big pin red,* ***pin small and blue, pin blue and small***	*pin blue, pin red, pin big, pin small, small pin red, big pin red,* ***pin small and blue, pin blue and small***

Figure 1: Scenes of interest, utterance alternatives in English, and "Englishified" utterance alternatives in our three hypothetical Spanish idiolects. Bolded utterances include a redundant adjectival modifier for the purposes of referring to the target object (highlighted in green border).

sions, and to the inherently incremental nature of linguistic production and comprehension.

In this paper, we present a novel computational model of speakers' choice of referring expression, synthesizing recent analyses proposed within the Rational Speech Act (RSA) framework. In Section 2, we review relevant findings from the experimental literature on linguistic reference, including within-language and cross-linguistic patterns in production choice that inform our desiderata of a successful computational model. In Section 3, we examine the properties of existing models and argue that a synthesis of those models is necessary to meet our desiderata. We present that synthesis in Section 4 and report the cross-linguistic predictions made by the model under various parameter regimes. Section 5 extends the analysis to various possible Spanish idiolects, which vary according to their preferred complex multi-adjectival determiner phrase (DP) structures.

2 Previous experimental findings

A theory of referring expression production should explain observed human production choices. We focus on two phenomena that such a theory should capture: the color/size asymmetry in overmodification observed in English, and cross-linguistic variation in overmodification.

2.1 The color/size asymmetry

Redundant modification is attested in both SS and CS-like contexts. However, in English, Dutch, and German—the prenominal adjective languages which have received the most attention—redundant color modification is much more frequent than redundant size modification (Degen et al., 2020; Gatt et al., 2011; Koolen et al., 2013; Pechmann, 1989; Sedivy, 2003).

2.2 Cross-linguistic variation

Linear ordering of DPs varies cross-linguistically, and there is empirical evidence that this variation patterns with differing rates of overmodification across languages. In particular, Rubio-Fernández and collaborators (Rubio-Fernández, 2016; Rubio-Fernandez et al., in press) have found that speakers of Spanish—a language that obligatorily places color adjectives post-nominally in single-modifier DPs (e.g. *el vestido azul*—'the blue dress')—produce redundant color modifiers less frequently than do speakers of English.

The empirical picture on cross-linguistic redundant modification is far from complete. The status of redundant modification in complex (>1 adjective) DPs is basically unknown beyond English. However, the model should minimally account for the lower rates of postnominal redundant color modification compared to prenominal redundant color modification (consistent with Rubio-Fernández's results for Spanish).

From her Spanish and English results, Rubio-Fernández (2016) argues that redundant modification is positively related to the marginal benefit of producing the modifier in facilitating a listener's search for the intended referent. This marginal benefit is lower in Spanish relative to English because adjectival modification occurs relatively late in the linear order of the DP, after other informative

forms—most importantly, the noun—have been produced. If this is correct, then we might expect an overall dispreference for redundant modification in languages for which modification is obligatorily postnominal. Indeed, this prediction is borne out by our model for one postnominal system we explore.

Given the existing empirical work on redundant modification in Spanish with single-modifier DP structures, a natural next step would be to explore the predictions of the model for that language in complex DP structures. However, there appears to be substantial variation—both in the theoretical literature and among four Spanish native speakers we consulted for this project—regarding linear ordering preferences in complex DPs. We return to this variation in Section 5. For the time being we only consider one type of modification structure: Scontras et al. (2020) report that many Spanish speakers allow for a complex DP structure where the size adjective follows the color adjective, a judgment corroborated by the translation of *the small blue pin* provided by two of our four native Spanish speakers in (1):

(1) La tachuela azul pequeña
det.def.f.sg pin blue small.f.sg
'The small blue pin'

(1) is an instance of fully postnominal modification, with both adjectives coming after the noun *tachuela* (pin). When comparing cross-linguistic predictions of existing models of referring expression choice in Section 3, we will focus on a hypothetical idiolect of Spanish (we'll call it Spanish-postnom.—a shorthand for Spanish-*postnominal*), which prefers this attested pattern of modification in complex DP structures. Comparing English to Spanish-postnom. is sufficient to demonstrate whether existing computational models are capable of making cross-linguistic predictions in general.

3 Previous computational theories

To what extent do previous computational theories of referring expression production predict the color-size asymmetry and cross-linguistic variation in overmodification? We focus on providing a qualitative assessment of the Rational Speech Act model proposed by Frank and Goodman (2012) and extensions thereof.[2]

[2]A non-RSA model of redundant modification is proposed by van Gompel et al. (2019). Their PRO (probabilistic referential overspecification) model captures the color/size asymmetry for languages with prenominal modifiers.

3.1 A note about utterance alternatives

We assume separate sets of utterance alternatives for the SS and CS scenes (displayed in Figure 1), comprised in each case of all possible licit DPs that a) can be composed by combining *big/small*, *blue/red*, and *pin* (or their equivalent Spanish translations); and b) truthfully describe one of images in the scene. We provide the English glosses of the Spanish DPs in Figure 1, omitting definite determiners for readability.

We assume that English exclusively permits prenominal adjectival modification of nouns (e.g. *the blue pin* but not **the pin blue*). We also assume that English permits multiple modifiers in a single DP (e.g. *the big blue pin*), and that speakers of English display robust adjective ordering preferences: complex DPs with both color and size adjectives place color adjectives closest to the head noun (Dixon, 1982; Sproat and Shih, 1991). Recent work provides empirical support for this generalization from the theoretical literature (Scontras et al., 2017; Hahn et al., 2018; Scontras et al., 2019).

Conversely, we will make the simplifying assumption that the Spanish-postnom. idiolect permits postnominal modification exclusively. We assume that the construction with adjectives in the reverse order - *la tachuela pequeña azul*, which flips the order of the size and color adjective - is wholly unavailable; none of our native speaker consultants offered this construction as a possible translation of *the small blue pin*.

Our Spanish native speaker judgments are consistent with the observation that ordering preferences in postnominal-modifiying languages tend to 'mirror' the preferences seen in prenominal-modifying languages (Hetzron 1978; Sproat and Shih 1991; for recent experimental evidence from Arabic see Kachakeche and Scontras 2020). Scontras et al. (2020) provide evidence that this holds for Spanish speakers, who disprefer *la tachuela pequeña azul* (which flips the order of the size and color adjective). As with English, we build this preference into the model and rule out *la tachuela pequeña azul* as a possible alternative in Spanish-postnom.

3.2 Standard (S-)RSA

We begin with Frank and Goodman (2012)'s 'standard' (S-)RSA model of referring expression production, which has been shown previously to capture neither redundant modification (Gatt et al., 2014; Degen et al., 2020) nor cross-linguistic vari-

ation (Cohn-Gordon et al., 2018b).

In S-RSA, speakers are modeled as conditional probability distributions over utterances given intended referents. To model speakers in this way, we first define a 'literal' semantic listener L_0 as a conditional distribution over referents R given an observed utterance u from an available set of utterances U. The probability of L_0 inferring r given u is proportional to the output of applying $[[u]]^D$ to r, multiplied by the listener's prior beliefs $P(r)$ about the probability of r being the intended referent. $[[\cdot]]^D$ is a Discrete semantic interpretation function whose outputs are values in $\{0, 1\}$:

$$P_{L_0}(r|u) \propto [[u]]^D(r) \cdot P(r)$$

Formally, we take utterances in S-RSA to be unordered sets of the lexical items i that comprise the utterance. We assume an intersective semantics for adjectives and nouns: that is, we first compute discrete Boolean truth values $\{0, 1\}$ for each lexical item i in u by determining whether i is true or false of the referent r. This is achieved with a discrete lexical interpretation function $\mathcal{L}^D$. We then take the product of values computed for each item in the u to retrieve a truth value for u:

$$\mathcal{L}^D(r, i) = \begin{cases} 1 & \text{if } i \text{ is true of } r \\ 0 & \text{otherwise} \end{cases}$$

$$[[u]]^D(r) = \prod_{i \in u} \mathcal{L}^D(r, i)$$

In what follows, we assume that this literal interpreter has uniform prior beliefs over all and only the visible referents in the context. On this assumption, L_0 assigns zero probability to referents that are truth-conditionally incompatible with the observed utterance and equal probability to all other referents. For example, in the SS context, *blue pin* is true of two referents—$o_{\text{small\_blue}}$ and $o_{\text{big\_blue}}$—so $P_{L_0}(o_{\text{small\_blue}}|\textit{blue pin}) = 0.5$ and $P_{L_0}(o_{\text{big\_blue}}|\textit{blue pin}) = 0.5$. In the CS context, *blue pin* is true only of $o_{\text{small\_blue}}$, so $P_{L_0}(o_{\text{small\_blue}}|\textit{blue pin}) = 1$.

The probability of speaker S_1 producing utterance u given intended referent r is modeled as S_1 soft-maximizing the utility of producing u:

$$P_{S_1}(u|r) \propto e^{\alpha(\ln P_{L_0}(r|u) - C(u))}$$

In particular, the probability of producing u is positively related to the probability that observing u would lead L_0 to infer r and is negatively related to utterance production cost $C(u)$. The conditional distribution over utterances is further modulated by an optimality parameter α: a high α value increases the difference between high-probability and low-probability utterances; an infinite α value corresponds to a utility-maximizing agent.

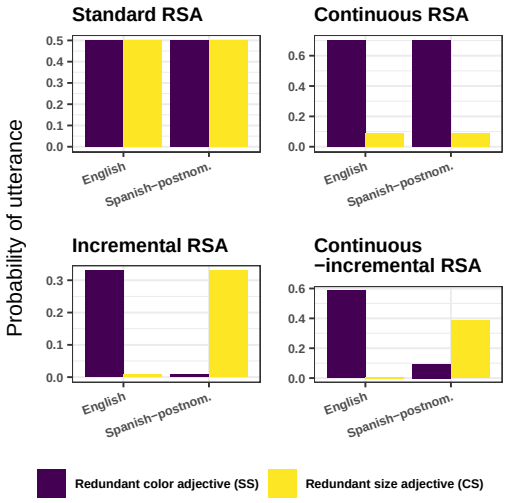


Figure 2: Model predictions for redundant color modification (purple: *small blue pin* or its translation used to refer to $o_{\text{small\_blue}}$ in the SS scene) and redundant size modification (yellow: *small blue pin* or its translation to refer to $o_{\text{small\_blue}}$ in the CS scene). Continuous-incremental (CI)-RSA—our proposal—predicts the color-size asymmetry in English while also predicting overall asymmetries in redundant modification cross-linguistically. For the continuous models, we assume v^{size} = 0.8 and v^{color} = 0.95; for the incremental models, we assume a cost of 0.1 for adjectives. We model predictions with $\alpha = 30$ for utterance-level models and $\alpha = 7$ for incremental models.

Suppose that the speaker finds herself in the SS context—in which case she has at her disposal all of the English utterances provided in the lefthand column corresponding to the SS context in Figure 1—or in the CS context, in which case her English utterance choices are provided in the righthand column. We assume a simple intersective semantics for the words that make up the possible utterances; e.g. *small blue pin* is true of a referent iff that referent is small, blue, and a pin. We assume a cost of 0 for all utterances. Finally, we set α to 30 (though changing cost and alpha does not change the predictions for S-RSA).

Under these parameter value assumptions, $P_{S_1}(\textit{small blue pin}|o_{\text{small\_blue}})$ has the same value in the SS context and the CS context (see Figure 2);[3] in other words, S-RSA predicts equal rates of

[3]The values in this and subsequent figures were calculated with WebPPL (Goodman and Stuhlmüller, 2014). All code

redundant color and redundant size modification, which furthermore is never predicted to exceed the rate of non-redundant modification. The only way to break this symmetry in the desired direction is to impose asymmetric costs on color and size adjectives, which is empirically and conceptually unmotivated (see Degen et al. 2020 for discussion).

3.3 Continuous (C-)RSA

To capture that speakers routinely overmodify, and that they do so asymmetrically with color vs. size adjectives, Degen et al. (2020) propose an update to the semantic interpretation function that captures the intuition that certain adjectives are more noisy/less reliable than others. On their proposal, $[[u]]^C(r)$ (a *C*ontinuous semantic interpretation function) returns real values in the interval $[0, 1]$. That value is computed first by considering all of the lexical items that compose u, retrieving values on the interval $[0, 1]$ for each lexical item i in u, then taking the product of those values as before:

$$\mathcal{L}^C(r, i) = \begin{cases} v^i & \text{if } i \text{ is true of } r \\ 1 - v^i & \text{otherwise} \end{cases}$$

$$[[u]]^C(r) = \prod_{i \in u} \mathcal{L}^C(r, i)$$

$$P_{L_0}(r|u) \propto [[u]]^C(r) \cdot P(r)$$

v^i is the continuous semantic value of a lexical item, which is a real number determined by context in the interval $[0, 1]$. Values of v^i close to 1 lead to speaker behavior similar to the binary Boolean semantics assumed by S-RSA: S_1 can make utterance choices with the expectation that the L_0 who encounters utterance u containing i will be highly likely to exclusively consider referents of which i is true. However, the lower we set v^i, the less reliably L_0 exclusively considers referents that truthfully verify i, thus diminishing S_1's expected utility of utterances containing i.

Following Degen et al. (2020), we suppose that v^i is high when i is a color term and relatively low when i is a size term, meaning that size terms are generally lower utility than color terms. We set $v^{\text{blue}} = v^{\text{red}} = 0.95$ and $v^{\text{big}} = v^{\text{small}} = 0.8$, with other parameters unchanged from Section 3.2.

Given this parameterization of the model, $P_{S_1}(\textit{small blue pin}|o_{\text{small\_blue}})$ is higher in the SS context than in the CS context (see Figure 2); that is, C-RSA predicts higher rates of redundant color modification than redundant size modification. This is because color adjectives are less 'noisy' than size adjectives. In the SS scene, the expected utility of *small* in uniquely establishing reference to $o_{\text{small\_blue}}$ outweighs the expectation that the adjective will fail due to its noisiness. The dynamics change in the CS scene, where *small*'s expected utility is lower due to the fact it is true of two objects in the scene.

A similar trade-off obtains for redundant color modification, but there is an overall lower expectation that *blue* will fail, boosting the probability of its production all else equal compared to *small*. This model thus captures the color/size overmodification asymmetry. However, because semantic interpretation is not sensitive to the linear order of lexical items, the model fails to predict any cross-linguistic variability (similar to S-RSA).

3.4 Incremental (I-)RSA

Cohn-Gordon et al. (2018b) propose a different revision to S-RSA, whereby listeners are modeled as conditional distributions over referents given observation of incrementally-produced sentences. Below, c is a context—a possibly empty sequence of words—and i is a lexical item observed after c.

$$L_0^{INCR}(r|c, i) \propto \mathcal{X}^D(c, i, r) \cdot P(r)$$

$\mathcal{X}^D$ is a string interpretation function that returns a semantic value for incomplete strings of the language given an intended referent r. The function considers all grammatical full-utterance continuations u of the string formed by the linear concatenation of a context c and lexical item i and returns the number of such continuations where u is true of r, divided by all possible continuations:

$$\mathcal{X}^D(c, i, r) = \frac{|u{:}[[u]]^D(r){=}1 \wedge u \text{ is a continuation of } c{+}i|}{|u{:}u \text{ is a continuation of } c{+}i|}$$

The outputs of $\mathcal{X}^D$ may be any real value on the interval $[0, 1]$, but we apply the superscript ^{D} because this continuous value is computed using the discrete semantic interpretation function for utterances that was defined for the S-RSA model.

Speakers in turn are modeled as incremental decision-makers, formally as conditional distributions over lexical items given a context c of items already produced and an intended referent r:[4]

$$S_1^{INCR}(i|c, r) \propto e^{\alpha(L_0^{INCR}(r|c,i) - C(i))}$$

can be accessed at https://github.com/bwaldon/crossling_reference.

[4] We discuss the word-level implementation of I-RSA, though see Cohn-Gordon et al. (2018a) for a character-level variant which makes use of continuous semantic values computed from a language model trained on image caption data.

Utterance-level probabilities are computed by applying the chain rule to the incremental speaker function. Unlike with S-RSA and C-RSA, utterances are ordered sequences of lexical items, and i_j denotes the j-th lexical item in u:

$$S_1(u|r) = \prod_{j=1}^{n} S_1^{INCR}(i_j|c = [i_1...i_{j-1}], r)$$

To compare I-RSA to the other models, we must make some principled modifications to the parameter values assumed thus far. First, because the optimality parameter α operates at the sub-sentential level in the incremental model, high values of α quickly give rise to very extreme distributions. Therefore, we lower α to 7. Lastly, following Cohn-Gordon et al. (2018b), we bookend our utterances with phonologically null *START* and *STOP* tokens. We assume that each mention of an adjective incurs a cost of 0.1, which would not change the qualitative pattern of predictions for S-RSA and C-RSA but allows for S_1^{INCR} to balance informativity against cost at the word level.

Examining the predictions (see Figure 2), we see that the incremental RSA model meets two of our desiderata: first, it correctly predicts the color/size asymmetry in English; second, it predicts that prenominal redundant color modification is more frequent than postnominal redundant color modification.[5] However, the marginal utility of producing a color or size adjective in both languages is purely a function of the alternative options available to the speaker at a given step in utterance production. That is, color and size are assumed to have equal communicative utility all else equal.

Consequently, the model predicts that Spanish-postnom. and English should mirror one another with respect to redundant adjective use in the SS/CS scenes. That is, the probability of a redundant color adjective in Spanish-postnom. is predicted to be the same as the probability of a redundant size adjective in English, and vice versa (the first row of Figure 3 illustrates this symmetry in I-RSA's predictions).

4 Continuous-incremental (CI-)RSA

Our proposal is a simple one: to leverage the incremental architecture proposed by Cohn-Gordon et al. (2018b) with a word-level continuous semantics function following Degen et al. (2020). We use the S_1 speaker definition from I-RSA, but we redefine the string interpretation function employed by the literal listener such that string meanings are computed from continuous rather than discrete utterance meanings, as in C-RSA. To compute this new string meaning, we take the sum of the (continuous) semantic values of all full-utterance continuations and divide by the number of continuations.

$$\mathcal{X}^C(c, i, r) = \frac{\sum [[u]]^C(r) : u \text{ is a continuation of } c+i}{|u : u \text{ is a continuation of } c+i|}$$

$$L_0^{INCR}(r|c, i) \propto \mathcal{X}^C(c, i, r) \cdot P(r)$$

We redefine L_0^{INCR} such that the probability of a referent r given observation of i in context c is proportional to the continuous string meaning of $c + i$ applied to r. Leaving all other parameter values the same as in the incremental RSA implementation above, and assuming semantic values for color/size adjectives identical to those used in the C-RSA implementation, we can now systematically compare our model against its predecessors.

CI-RSA—like I-RSA and C-RSA—predicts the color/size asymmetry in English. It also predicts—like I-RSA—higher rates of redundant color adjective use in English over Spanish-postnom. However, the new model predicts language-level asymmetries in redundant adjective use: that is, across the SS and CS scenes, it predicts a lower overall redundant modification probability in Spanish-postnom. than in English.

This model therefore meets the motivating desiderata: it captures both the observed color/size asymmetry in English as the result of reasoning about noisy modifiers; and it predicts cross-linguistic variation in overmodification that is consistent with the current empirical landscape. Moreover, CI-RSA (like I-RSA) with the above assumed parameter values makes a perhaps surprising prediction: the color/size asymmetry should flip in Spanish (and postnominal languages in general). To the best of our knowledge, it is unknown whether this prediction is borne out empirically due to the absence of empirical data on the relative rates of size and color overmodification in postnominal modifier languages.

5 Modeling variation within Spanish

Spanish speakers appear to be quite diverse with regards to their complex DP preferences. In addition to the postnominal modification assumed above for

[5] That the predicted probability of redundant size modification is so low in English is consistent with results from Degen et al. (2020) and others who report very low rates of size modification in the CS scene displayed in Figure 1 and negligibly higher rates in scenes with more visual variability.

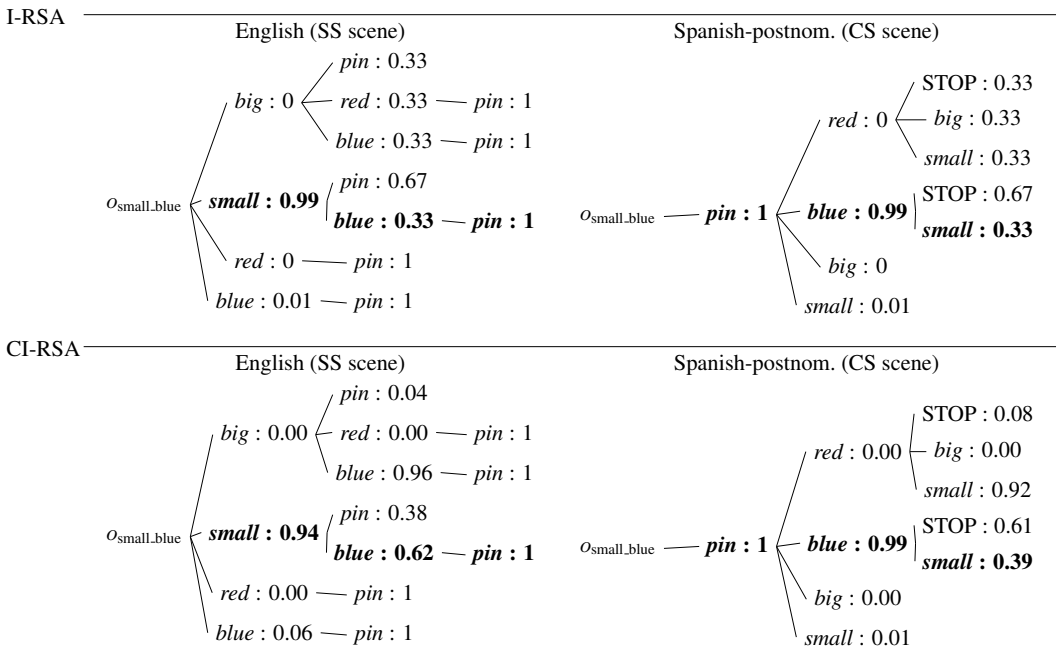


Figure 3: Transitional probabilities from one word to another for the incremental speaker S_1^{INCR}. Multiplying bolded nodes yields probability of producing *small blue pin*. Unlike CI-RSA, I-RSA predicts symmetry in English redundant color modification (left column) and Spanish-postnom. redundant size modification (right column). We omit the *START* and *STOP* tokens when probability $= 1$. In I-RSA, transitions can have exactly 0 probability, in which case later transitions all have equal probability.

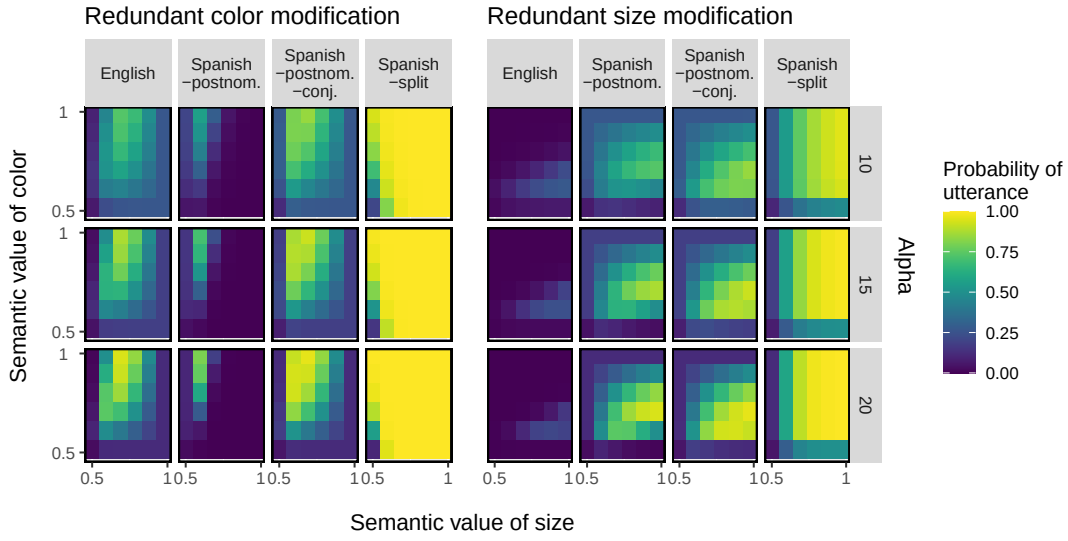


Figure 4: Predicted utterance probabilities for *small blue pin* (and its translations) in the size-sufficient (left panels) and color-sufficient (right panels) scenes across our four languages of focus, under varying semantic values for color and size in CI-RSA. Rows indicate varying α values. For 'Spanish-postnom.-conj.', we report the probability of producing either (2a) *pin small and blue* or (2b) *pin blue and small*. We assume a cost of 0.1 on adjectives.

Spanish-postnom., Spanish also allows for post-nominal modification with conjunction (where adjective ordering preferences are suspended, see Ford and Olson 1975 and Byrne 1979 for evidence from English speakers; Rosales Jr and Scontras 2019 for Spanish). (2a) and (2b) exemplify this strategy:[6]

[6]Conversely, the authors report that English speakers exhibit the same basic ordering preference of size adjectives before color adjectives even when conjunction is present.

(2) a. La tachuela azul y pequeña
det.f.sg pin blue and small.f.sg

b. La tachuela pequeña y azul
det.f.sg pin small.f.sg and blue
'the small blue pin'

Spanish also allows a 'split' structure in which the size adjective precedes the noun and the color adjective follows it as in (3a), which was offered as a possible translation of *the small blue pin* by one of our native speaker consultants (3b is out for this same speaker):[7]

(3) a. La pequeña tachuela azul
det.f.sg small.f.sg pin blue

b. *La azul tachuela pequeña
det.f.sg blue pin small.f.sg

The variation in available strategies suggests that more work is needed to understand the full empirical picture of referring expressions in Spanish. For now, we consider two additional hypothetical idiolects, presented in Figure 1: one ('Spanish-split') in which the preferred complex DP structure is (3a) and another ('Spanish-postnom.-conj.', shorthand for Spanish-*postnominal-conjunctive*) in which (2a) and (2b) are each available.

We present the cross-linguistic predictions of CI-RSA under various possible parameter combinations in Figure 4 (assuming 0.1 cost for adjectives, as above). This illustration demonstrates how CI-RSA expands the prediction space of previous incremental and continuous RSA models, particularly with regards to redundant adjectival modification and its cross-linguistic variation. A notable prediction of the model—robust across many parameter regimes—is that all the explored idiolects of Spanish should exhibit more size overmodification in complex DPs than does English. Another interesting prediction is the overwhelming preference of Spanish-split speakers to overmodify, which can be attributed to the relative utility of producing an informative size adjective versus the uninformative head noun at the beginning of the utterance for the particular referential contexts we consider here.

[7] See e.g. Hetzron (1978) and Cinque (1994) for some discussion of this 'split' construction in the Romance languages French, Italian, and Ladin. In these languages, all closely related to Spanish, evaluative and size adjectives may occur prenominally in DPs featuring color modification, for example the French *un joli gros ballon rouge* 'a pretty big red ball' (Hetzron 1978 ex. 8c, cited in Cinque 1994). Cinque argues that the 'split' construction is due to syntactic head movement of the noun over the color adjective to a linear intervening position between the modifiers. Scontras et al. (2020) note, however, that the 'split' construction is evidently not fully productive in Spanish.

6 Discussion and conclusion

This work highlights the empirical gaps that remain regarding our understanding of referring expression production cross-linguistically. More work is needed, for example, to understand when and how redundant modification manifests in Spanish and other postnominal languages, and what complex DP structures are in fact produced by Spanish speakers in referring contexts.

While we explored predictions for three hypothetical Spanish idiolects, it is implausible to assume that real Spanish speakers display an overwhelming preference for only one of the three possible strategies over the other two (e.g., most Spanish speakers are likely to employ both the split structure and at least one of the postnominal structures). Furthermore, one consultant reports that in the SS context, the most natural translation of *the small blue pin* features the diminutive morpheme *-ito/a* affixed to the noun (*la tachuelita azul*). In this paper, of course, we exclusively consider redundant lexical modifiers. More empirical work is needed to understand the factors that may give rise to individual-level and context-specific preferences for particular constructions, and for the time being we have nothing to say about sub-lexical realizations of modification.

Moreover, there is some reason to believe that not all complex DPs express equivalent meanings: Laenzlinger (2000), for example, notes a difference in meaning between the prenominally-modified simple DP *un grand hommme* 'a great man' and the postnominal *un homme grand* 'the tall man' in French, a Romance language closely related to Spanish. More work is needed to understand whether pre-nominal size modification in Spanish gives rise to similar (or more subtle) meaning changes in complex Spanish DPs in referring contexts, and whether this change in meaning is subject to population-level variation.[8]

Further factors that affect the probability of redundant modification that models of referring expression production should capture include scene variation (Davies and Katsos, 2013; Koolen et al.,

[8] For a discussion of meaning shifts between pre and postnominal modification in simple DPs in Spanish, see e.g. Centeno-Pulido (2010), who notes that "prenominal adjectives display particular properties (namely, they are in focus, they are emphasized, or the [sic] express some sort of subjectivity)" (2010: 76). Still to be explored, however, is the distribution of pre-nominal modification in referring contexts, especially in complex DP constructions.

2013; Rubio-Fernandez et al., in press) and feature typicality (Rubio-Fernández, 2016; Sedivy, 2003; Westerbeek et al., 2015; Kreiss and Degen, 2020). Degen et al. (2020) show that C-RSA captures scene variation effects, and this should be true for CI-RSA as well. They also show that an extension of C-RSA captures feature typicality effects; the current formulation of CI-RSA does not.

Cross-linguistic empirical investigations of referring expression production should also investigate redundant modification outside the domains of size and color. There is evidence from English, for example, that redundant material adjectives (e.g. *metal*, *wooden*) are less likely to be produced than are redundant color adjectives (Sedivy, 2003). Kursat & Degen (forthcoming) argue that this asymmetry is explained in part by the relative difficulty of perceiving an object's material compared to perceiving its color. However, the robustness of this effect has yet to be investigated beyond English, and it remains to be seen whether CI-RSA can account for the cross-lingusitic patterns that emerge.

We acknowledge furthermore that the contexts we investigate all feature objects of the same type; thus, we assume that the noun has no communicative value on its own. We present these contexts in the interest of simplicity, though the predictions of CI-RSA should be evaluated in contexts where the noun distinguishes between possible objects.

While the best we can do at present is to provide a qualitative assessment of the model against existing alternatives on the few existing data points as we have done in this paper, an advantage of probabilistic pragmatic theories such as those that extend the RSA framework is that they can be quantitatively evaluated against experimental data. Degen et al. (2020) recently evaluated the C-RSA model in this way and report that the model provides a good fit for English data collected in an interactive reference game study. Their paradigm is a candidate for cross-linguistic replication studies, so that CI-RSA can be rigorously compared against its RSA antecedents. We leave this to future work.

References

Anja Arts, Alfons Maes, Leo Noordman, and Carel Jansen. 2011. Overspecification facilitates object identification. *Journal of pragmatics*, 43(1):361–374.

Brian Byrne. 1979. Rules of prenominal adjective order and the interpretation of "incompatible" adjective pairs. *Journal of Verbal Learning and Verbal Behavior*, 18(1):73 – 78.

Alberto Centeno-Pulido. 2010. *Reconciling generativist and functionalist approaches on adjectival position in Spanish*. Ph.D. thesis, University of Georgia.

Guglielmo Cinque. 1994. On the evidence for partial n-movement in the romance dp. In *Paths towards universal grammar. Studies in honor of Richard S. Kayne*. Georgetown University Press.

Reuben Cohn-Gordon, Noah Goodman, and Christopher Potts. 2018a. Pragmatically informative image captioning with character-level inference. *arXiv preprint arXiv:1804.05417*.

Reuben Cohn-Gordon, Noah D Goodman, and Christopher Potts. 2018b. An incremental iterated response model of pragmatics. *arXiv preprint arXiv:1810.00367*.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive science*, 19(2):233–263.

Catherine Davies and Napoleon Katsos. 2013. Are speakers and listeners 'only moderately gricean'? an empirical response to engelhardt et al. (2006). *Journal of Pragmatics*, 49(1):78 – 106.

Kees van Deemter, Albert Gatt, Ielka van der Sluis, and Richard Power. 2012. Generation of referring expressions: Assessing the incremental algorithm. *Cognitive science*, 36(5):799–836.

Judith Degen, Robert D Hawkins, Caroline Graf, Elisa Kreiss, and Noah D Goodman. 2020. When redundancy is useful: A bayesian approach to "overinformative" referring expressions. *Psychological Review*.

Robert MW Dixon. 1982. *Where have all the adjectives gone?: and other essays in semantics and syntax*, volume 107. de Gruyter.

Paul E Engelhardt, Karl GD Bailey, and Fernanda Ferreira. 2006. Do speakers and listeners observe the gricean maxim of quantity? *Journal of memory and language*, 54(4):554–573.

William Ford and David Olson. 1975. The elaboration of the noun phrase in children's description of objects. *Journal of Experimental Child Psychology*, 19(3):371 – 382.

Michael C Frank and Noah D Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.

Albert Gatt, Roger van Gompel, Emiel Krahmer, and Kees van Deemter. 2011. Non-deterministic attribute selection in reference production. In *Proceedings of CogSci*.

Albert Gatt, Emiel Krahmer, Kees Van Deemter, and Roger PG Van Gompel. 2014. Models and empirical data for the production of referring expressions. *Language, Cognition and Neuroscience*, 29(8):899–911.

Roger PG van Gompel, Kees van Deemter, Albert Gatt, Rick Snoeren, and Emiel J Krahmer. 2019. Conceptualization in reference production: Probabilistic modeling and experimental testing. *Psychological review*, 126(3):345.

Noah D Goodman and Andreas Stuhlmüller. 2014. The Design and Implementation of Probabilistic Programming Languages. http://dippl.org. Accessed: 2021-1-14.

H. P. Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics: Vol. 3: Speech Acts*, pages 41–58. Academic Press, New York.

Michael Hahn, Judith Degen, Noah D Goodman, Dan Jurafsky, and Richard Futrell. 2018. An information-theoretic explanation of adjective ordering preferences. In *CogSci*.

Robert Hetzron. 1978. *On the relative order of adjectives*, pages 165–184. Narr Tübingen.

Zeinab Kachakeche and Gregory Scontras. 2020. Adjective ordering in arabic: Post-nominal structure and subjectivity-based preferences. *Proceedings of the Linguistic Society of America*, 5(1):419–430.

Ruud Koolen, Albert Gatt, Martijn Goudbeek, and Emiel Krahmer. 2011. Factors causing overspecification in definite descriptions. *Journal of Pragmatics*, 43(13):3231 – 3250.

Ruud Koolen, Martijn Goudbeek, and Emiel Krahmer. 2013. The effect of scene variation on the redundant use of color in definite reference. *Cognitive Science*, 37(2):395–411.

Elisa Kreiss and Judith Degen. 2020. Production expectations modulate contrastive inference. In *CogSci*.

Christopher Laenzlinger. 2000. French adjective ordering: Perspectives on dp-internal movement types. *Lingua*, 115(5):645–689.

Alfons Maes, Anja Arts, and Leo Noordman. 2004. Reference management in instructive discourse. *Discourse Processes*, 37(2):117–144.

Aparna S Nadig and Julie C Sedivy. 2002. Evidence of perspective-taking constraints in children's online reference resolution. *Psychological Science*, 13(4):329–336.

Thomas Pechmann. 1989. Incremental speech production and referential overspecification. *Linguistics*, 27(1):89 – 110.

Cesar Manuel Rosales Jr and Gregory Scontras. 2019. On the role of conjunction in adjective ordering preferences. *Proceedings of the Linguistic Society of America*, 4(1):32–1.

P. Rubio-Fernandez, F. Mollica, and J. Jara-Ettinger. in press. Speakers and listeners exploit word order for communicative efficiency: A cross-linguistic investigation. *Journal of Experimental Psychology*.

Paula Rubio-Fernández. 2016. How redundant are redundant color adjectives? an efficiency-based analysis of color overspecification. *Frontiers in Psychology*, 7:153.

Gregory Scontras, Galia Bar-Sever, Zeinab Kachakeche, Cesar Rosales Jr, and Suttera Samonte. 2020. Incremental semantic restriction and subjectivity–based adjective ordering. In *Proceedings of Sinn und Bedeutung 24*.

Gregory Scontras, Judith Degen, and Noah D Goodman. 2017. Subjectivity predicts adjective ordering preferences. *Open Mind*, 1(1):53–66.

Gregory Scontras, Judith Degen, and Noah D Goodman. 2019. On the grammatical source of adjective ordering preferences. *Semantics and Pragmatics*, 12.

Julie C Sedivy. 2003. Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. *Journal of psycholinguistic research*, 32(1):3–23.

Richard Sproat and Chilin Shih. 1991. The cross-linguistic distribution of adjective ordering restrictions. In *Interdisciplinary approaches to language*, pages 565–593. Springer.

Hans Westerbeek, Ruud Koolen, and Alfons Maes. 2015. Stored object knowledge and the production of referring expressions: the case of color typicality. *Frontiers in Psychology*, 6:935.