Evaluating models of referring expression production on an emerging sign language

Leyla Kursat[1], Brandon Waldon[1], Rabia Ergin[2], & Judith Degen[1]
lkursat@stanford.edu, bwaldon@stanford,edu, rabia.ergin@boun.edu.tr, jdegen@stanford.edu
[1]Department of Linguistics, Stanford University
Stanford, CA 94305, USA
[2] Institute of Social Sciences, Bogazici University
Istanbul, Turkey

Abstract

Redundant modification in referring expression production varies both within language (e.g., English speakers produce more redundant color than size modifiers) and cross-linguistically (e.g., English speakers produce more redundant color modifiers than Spanish speakers). It is an open question whether these asymmetries are the result of asymmetries in the general referential utility of color and size modifiers or of incremental language processing pressures. Cross-linguistic investigations of redundant modification are important to this debate: similar cross-linguistic rates of redundant modification would suggest a strong role for general referential utility. In contrast, lower prevalence of redundant modification in languages with post-nominal modification suggests a strong role for incrementality. Here, we test whether differences in redundant adjective use are systematic for a particularly interesting language: Central Taurus Sign Language. As a language in its infancy, CTSL has no established conventions, and therefore provides us with a unique opportunity to explore how redundancy emerges in the initial stages of language formation. We evaluate different computational models of referring expression that each make different assumptions regarding the source of asymmetries in the production of redundant modifiers.

Keywords: experimental pragmatics; redundancy; village sign language; Rational Speech Act; Bayesian Data Analysis

Introduction

Speakers produce redundant modifiers when referring to objects, and they do so in systematic ways. One of the best-studied patterns of redundant modification is the color/size asymmetry: when mentioning the size of an object is sufficient for establishing unique reference, speakers often include color modifiers in their utterances (e.g., "the small blue button" instead of "the small button" in Fig. 1). In contrast, when it is sufficient to mention color, speakers rarely mention size (Pechmann, 1989; Sedivy, 2003; Rubio-Fernandez, 2016). The tendency to overspecify with color has been attested in English (Degen, Hawkins, Graf, Kreiss, & Goodman, 2020), German (Belke, 2006), Dutch (Koolen, Goudbeek, & Krahmer, 2013) and Spanish (Rubio-Fernandez, 2016). Moreover, cross-linguistic investigations of redundancy found that Spanish speakers are less likely to use color modifiers relative to English speakers (Rubio-Fernandez, Mollica, & Jara-Ettinger, 2021; Wu & Gibson, 2021).

What is the source of these asymmetries? What linguistic, perceptual and communicative factors determine speakers' redundant modification patterns? There are different explanations offered in the literature for how speakers choose referring expressions. The color/size asymmetry has been explained as the result of an asymmetry in the general referential utility of modifiers: e.g., color may be more useful than size because it is perceptually easier to verify (Rubio-Fernandez, 2016; Kursat & Degen, 2021); because it is generally less 'noisy' in communication (Degen et al., 2020); or simply as a matter of brute fact (van Gompel, van Deemter, Gatt, Snoeren, & Krahmer, 2019). In contrast, the Spanish/English asymmetry is commonly interpreted as the result of incremental language processing pressures (Rubio-Fernandez et al., 2021). Here we seek to shed light on the relative importance of these factors in explaining patterns of redundant reference by investigating the production of redundant modifiers in a village sign language: Central Taurus Sign Language (CTSL). CTSL is a young language that arose in the absence of a conventionalized linguistic framework. Having no established conventions for reference – in particular, for the order of nouns and modifiers – it allows us to explore the role of referential utility and incremental processing pressures in guiding signers' production of redundant modifiers.

Accounts of Redundant Modifier Production

According to accounts that ascribe an important role to the general referential utility of different properties, redundant use of modifiers is the result of general principles of cognition and perception (Degen et al., 2020; Kursat & Degen, 2021; Rubio-Fernandez, 2016). For example, one reason why color might have greater referential utility than size is because it is perceptually easier to assess. This type of explanation is formally captured by a recent *continuous semantics* computational model of referring expression production (Degen et al., 2020) that allows modifiers to be informative about objects to varying degrees. Referential utility theories tend to be opaque to cross-linguistic variation in referring expression structure, and thus predict similar patterns across languages. Under these accounts, and assuming that the perceptual mechanisms are shared across people, languages should have similar rates of redundant modification and display similar patterns. However, recent studies conducted in Spanish found evidence contradicting this prediction (Rubio-Fernandez et al., 2021; Wu & Gibson, 2021): redundant color modifier production was more likely in English than in Spanish. This cross-linguistic asymmetry is important evidence that incrementality, in particular sensitivity to word order, needs to be integrated into computational accounts of redundant modifier use.

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society*.

Incremental accounts explain the asymmetry in redundant modifier use in Spanish and English with the different modifier head word orders of the two languages (Rubio-Fernandez et al., 2021; Wu & Gibson, 2021). According to these accounts, it is more useful to produce color modifiers in English than in Spanish because in English, color modification is expressed pre-nominally. Therefore, in visual search tasks, the production of the color modifier facilitates earlier identification of the target via an additional cue to its identity in English, but not in Spanish, where the main cue to the target's identity – the head noun – is produced first. Although there is debate as to whether redundant modification is the result of speaker- or listener-centric behavior, this cross-linguistic asymmetry has been primarily interpreted as pre-nominal language speakers attempting to facilitate listeners' visual search (Jara-Ettinger & Rubio-Fernandez, 2021; Arts, Maes, Noordman, & Jansen, 2011; Rubio-Fernandez et al., 2021; Wu & Gibson, 2021). Cohn-Gordon, Goodman, and Potts (2019) show that the different rates of redundant color modification in English and Spanish can be captured by *incremental* computational models of pragmatic language use that allow for utterances to be planned incrementally. This model predicts different patterns for post-nominal vs. pre-nominal adjective languages because it allows speakers to pragmatically reason about each word's referential utility incrementally, one word at a time. In this model, the set of utterance choices at each point is determined by the word order of the language. The Incremental Communicative Efficiency model (ICE, Jara-Ettinger & Rubio-Fernandez, 2021) is another computational model of referring expression production that analyzes utterances incrementally, but adopts a listener-centric view in formalizing the source of the pressure to be sufficiently informative.

Speaker-centric accounts of redundant modifier production, inspired by availability-based production (Bock, 1987; Ferreira & Dell, 2000), attribute the asymmetries in the production of redundant modifiers to differences in the availability of the modifiers for the speaker. Under this kind of account, size modifiers might be produced less frequently than color modifiers because they are more costly for speakers, perhaps because they are harder to retrieve from memory.

Waldon and Degen (2021) provide a qualitative investigation of computational accounts of referring expression production that formalize the ideas of global referential utility differences vs. incrementality pressures, both independently and jointly, and compare their predictions in varying contexts. They show that the two types of explanations are mutually compatible and propose a new account that computes word utility incrementally using modifiers with referential utility differences. This novel *continuous-incremental* model predicts both the well-documented color/size asymmetry in English, and different rates of color modification in pre-nominal and post-nominal adjective languages.

We proceed by first introducing these computational models of pragmatic language use formalized within the Rational Speech Act (RSA) framework (Frank & Goodman, 2012; Goodman & Frank, 2016; Franke & Jäger, 2016), a computational framework that models pragmatic communication as recursive reasoning between a speaker and a listener. We then introduce Central Taurus Sign Language (CTSL), the language that is our testbed for investigating the source of asymmetries in the production of redundant modifiers. We report two production experiments designed to elicit redundant color and size modifiers in two different languages: CTSL and English. Finally, we evaluate each model on the production data via Bayesian data analysis.

Bayesian Models of Referring Expression Production

Standard RSA

In the standard RSA model, speakers and listeners recursively reason about each other's mental states to communicate. A pragmatic speaker produces utterances reasoning about a literal listener who observes an utterance, and returns a distribution over intended referents (see Table 1 for formalization). The literal listener computes the probability of an intended referent based on the semantics of the utterance and prior beliefs regarding which referent is most likely to be the target.

On Standard RSA, the semantic value of an utterance is a function from intended referents to binary truth values in {0,1}. These binary truth values are in turn computed from a lexical interpretation function that outputs binary truth values given a lexical item (e.g. an adjective) and an intended referent. Assuming uniform prior beliefs about the referents, a literal listener returns a uniform distribution over all compatible intended referents and assigns zero probability to other referents. Based on this distribution (informativeness of utterances for the literal listener), and the cost of each utterance, the pragmatic speaker calculates the utility of each utterance. As a Bayesian agent, a pragmatic speaker produces an utterance proportional to its utility.

Standard RSA doesn't produce redundant modifiers at the high rates observed empirically because for the literal listener, all utterances that are compatible with the object (e.g. "big" and "big green") are equally informative. In this model, the only way to break symmetry in the empirically-observed direction (i.e. that redundant referring expressions are more frequent than their 'minimal' counterparts) is assuming that the more complex utterance, "big green" is less costly than the simple utterance, "big", and this doesn't seem to be a plausible assumption.

Continuous RSA

Degen et al. (2020) extend the standard RSA model to capture the systematic patterns with which redundant modifiers are produced. By relaxing the deterministic Boolean semantics of utterances to a non-deterministic continuous semantics that returns real values between 0 and 1, they allow utterances to be informative with varying degrees. With this new definition of the semantics, modifiers differ in how noisy, and

Table 1: Summary of models under comparison.

Discrete semantics

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

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

Continuous semantics

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

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

Global production model

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

$$C_u(u) = \sum_{i \in u} C_i(i)$$

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

Incremental production model

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

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

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

Legend: u : an utterance	r : a referent	$[[\cdot]]$: utterance interp. function	$\mathcal{L}$: lexical interp. function
i : a lexical item	c: a partial utterance	v^i : i's continuous semantic value	$P(r)$: prior over r $C_{u/i}$: cost function on u/i

Standard RSA: **Discrete** semantics + **Global** production model
Incremental RSA: **Discrete + Incremental**
Continuous RSA: **Continuous + Global**
Continuous-Incremental RSA: **Continuous + Incremental**

therefore how useful they are for the purpose of establishing reference. Assuming that color modifiers are less noisy, more precise, than size modifiers, the observed asymmetry in speakers' propensity to redundantly use color over size adjectives is predicted. Thus, the symmetry problem that Standard RSA fails to resolve is easily addressed.

The assumption that color is less noisy than size, perhaps because it is easier to perceive or assess, points towards an explanation of the asymmetry in the production of redundant color and size modifiers that assumes a difference in referential utility. This model also accounts for other systematicities, e.g., that redundant color modification is more likely with increased scene variation (Koolen et al., 2013). However, it fails to predict cross-linguistic differences in redundant modification because it is insensitive to word order.

Incremental RSA

In Cohn-Gordon et al. (2019)'s extension of the RSA model, utterances are represented as sequences of words. Agents reason about each word, in the utterance's sequential order. Listeners are modeled as distributions over intended referents, given the word in the context. For the incremental pragmatic speaker, the utility of any partial utterance is a function of the probability that a full-utterance continuation of the partial utterance would truthfully refer to the intended referent. The possible continuations of a sequence of words is determined by language-specific linear constraints on word order.

Waldon and Degen (2021) show that the incremental model captures the color/size asymmetry in English, but also predicts the opposite asymmetry for the production of redundant color and size modifiers in post-nominal languages (more redundant mention of size) in contexts in which redundant modification manifests as multi-adjectival DP constructions (as in Fig. 1). This is a pattern not attested empirically.

Continuous-Incremental RSA

Waldon and Degen (2021) consider a novel extension of RSA that combines the continuous and incremental models. The speaker is defined as the incremental decision maker who reasons about a literal listener that computes utterance probabilities based on a continuous semantics for words (see Table 1).

This model both captures the color/size asymmetry observed in English and the lower rates of color modification produced by Spanish speakers. However, similar to the incremental model, this model also predicts Spanish speakers to redundantly mention size more frequently than color.

A Case Study: Central Taurus Sign Language

The conflicting model predictions and empirical results suggest that further cross-linguistic investigations of redundant modification are much needed. So far, studies addressing redundancy have been conducted on a handful of pre-nominal languages (English, Dutch and German) and with the exception of the few studies conducted in Spanish, redundant modification patterns have not been tested in post-nominal languages. Moreover, the only Spanish studies conducted thus far have not investigated cases where the noun is not informative and speakers could in principle mention multiple features via modification. Thus, for instance, whether the color/size asymmetry attested in pre-nominal languages generalizes is unknown. Here we explore this question through a language that is still in the initial stages of language formation: Central Taurus Sign Language (CTSL).

CTSL is a village sign language that emerged naturally within the last half century in a remote village in Southern Turkey (Ergin, 2017; Ergin, Senghas, Jackendoff, & Gleitman, 2020). It is currently used by 30 deaf signers and approximately 100 Turkish speakers, in three neighboring villages. It arose as a result of high incidence of hereditary deafness and the region's geographical isolation and in the absence of a linguistic framework. As a young language, CTSL offers a unique perspective into observing the structure of an emerging communication system.

Previous research on CTSL has focused on finding evidence of conventionalization. Ergin (2017), the first study on this language, explored the expression of sentential argument structure. She found that patterns became more conventionalized across cohorts. Ergin, Kursat, Hartzell, and Jackendoff (2021) focused on the expression of modification and found that the head-modifier order was dominant, but not deterministically so. Observing that signers follow the dominant order even more strictly with increasing semantic complexity, they argued that in CTSL, linear order is used as a minimal syntactic tool to express modification.

Exp. 1 asks whether the color/size asymmetry previously observed in pre-nominal adjective languages generalizes to CTSL. Exp. 2 is a replication of Exp. 1 on an English speaking group on Mechanical Turk. Finally, we conducted a model comparison of the four redundant modification models (standard RSA, continuous RSA, incremental RSA, continuous-incremental RSA) on both datasets.

Experiment 1: Redundant Modifier Production in CTSL

Exp. 1 used an interactive reference game paradigm to assess CTSL signers' redundant color and size modification.

Methods

Participants (n=22, all native CTSL signers, 21 deaf, 1 hearing) were paired into 11 director-guesser dyads to play a real time communication game (Hawkins, 2015).[1] They were asked to sit across from each other and face each other. Each participant had a monitor in front of them and an experimenter sitting next to them to assist with using the computer. In CTSL, only very few colors have lexicalized signs, and color is often expressed by pointing to objects in the immediate environment (clothes, furniture, food etc.). In order to minimize potential data loss, we made pencils of different color available to signers.

On each trial, participants saw four objects in the display. Both participants saw the same images, in different orders. On the director's screen, the target item was marked by a green border. The director's task was to describe the target to the guesser, who in turn was asked to guess the described image from the display by pointing at it. The experimenter sitting next to the guesser clicked on the image the guesser pointed at. After the selection, both participants received feedback about its correctness. Participants were allowed to freely sign to each other during the task.

[1]The small n is due to the generally small number of CTSL signers. Given the large effect sizes previously found in similar paradigms, Type II error is unlikely.

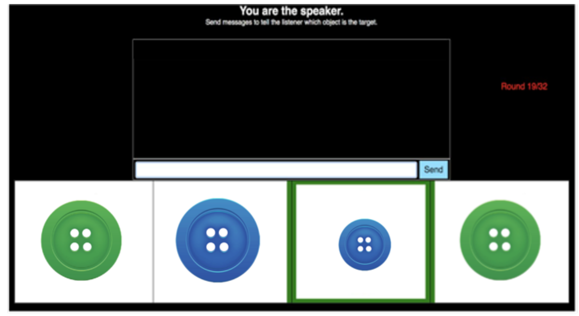

Figure 1: Example display from the director's display on a size sufficient trial.

There were 30 critical trials. On each critical trial (Fig. 1), the four images were of the same object type, and either mentioning the color or size of the target object was sufficient for establishing unique reference. On 15 color sufficient trials, mentioning size was redundant, because a competitor object shared the redundant feature (size) with the target. Similarly, on 15 size sufficient trials, a competitor had the same color as the target item, making color mention redundant. On each trial, there were two additional distractors that shared the sufficient property with the competitor. All items were normed for object and feature nameability.

Both participants were recorded during the task and their signs were transcribed to English for analysis. Different strategies used for expressing modifiers (e.g., pointing at colored pencils, pointing at clothes, mouthing) were also coded for each utterance.

Results and Discussion

We first classified the produced utterances as "color", "size" and redundant "color and size" utterances (see yellow bars in top row of Fig. 2 for empirical utterance proportions). As is clear from the figure, participants frequently produced redundant modifiers. To assess whether signers produced more redundant color than size modifiers, we conducted a mixed effect logistic regression model predicting redundant modifier use from a fixed effect of redundant property. The model included by-participant and by-item random intercepts and slopes. We observed a main effect of redundant property (β= 3.24, SE=0.45, $p<.0001$) such that signers were more likely to redundantly modify with color than size.

Next, to investigate whether modification is expressed pre-nominally or post-nominally in CTSL, for each utterance we coded the order with which the color, size and noun signs were produced (6 unique orders). We found that modification was overwhelmingly expressed post-nominally (87%). Within modification strategies, redundant modifiers were produced more frequently (97%) in post-nominal position: all redundant size modifiers were produced post-nominally and only 0.3% of redundant color modifiers were produced pre-nominally.

Together, these results are at odds with explanations of redundant modification that ascribe a large explanatory role

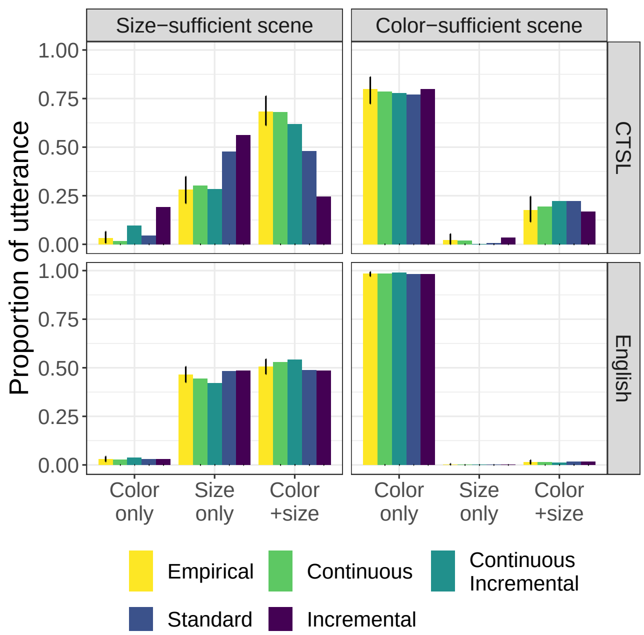

Figure 2: Empirically-observed "color", "size" and redundant "color and size" utterance proportions in CTSL (Exp. 1) and English (Exp. 2), compared against predictions of the four models under evaluation.

to forces of incrementality. Despite modification being expressed post-nominally in CTSL, we did not observe reduced redundancy compared to pre-nominal languages. Instead, the results suggest that global referential utility based principles could be underlying the systematicity with which redundant modifiers are produced.

Experiment 2: Redundant Modifier Production in English

Exp. 2 provided a pre-nominal adjective language baseline against which to interpret the results of Exp. 1. It was identical to Exp. 1, but was conducted on English over the web..

Methods

The task and materials were identical to Exp. 1. 100 participants were recruited through Amazon's Mechanical Turk and each participant was randomly assigned to either the director or the guesser role (50 pairs). Pairs played the online version of the reference game and used a chat-box to send messages.

Results

Proportions of "color", "size" and redundant "color and size" utterances are shown in Fig. 2. A mixed effects logistic regression predicting redundant modifier use from fixed effects of redundant property revealed a main effect of redundant property (β= 2.484, *SE*=0.40, $p<.0001$), such that participants were more likely to redundantly mention color than size, replicating the result of Exp. 1. Note that in comparison to CTSL signers, English speakers never produced redundant size modifiers, consistent with the previous literature.

Model Evaluation

In order to evaluate to what extent each model captures the empirically elicited production data, we conducted Bayesian data analysis with the CTSL and English data. [2] We inferred likely parameter values for each model, conditioned on the observed data. We used maximum estimates of the posteriors to generate model predictions and compared these predictions with to the observed data to assess the performance of each model. We performed full model comparison using Bayes factors to compare our two best performing models.

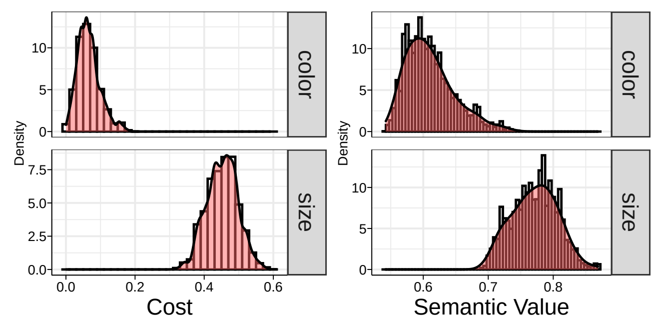

Figure 3: Posterior model parameter values of the continuous model on the CTSL data.

In each model, we assumed uniform prior values for the two separate cost parameters $C(u_{color})$, $C(u_{size}) \sim \mathcal{U}(0,2)$ [3] and the informativeness parameter $\alpha \sim \mathcal{U}(0,40)$. The continuous models included two additional semantic value parameters, v^{color} and $v^{size} \sim \mathcal{U}(0.5,1)$. When running the incremental models on the CTSL data, we constrained the alternative utterance space to only include post-nominal expressions. This was motivated by the results of Exp. 1 that suggested modification in CTSL is overwhelmingly expressed post-nominally (87%). Conversely, the incremental models evaluated on the English dataset only allowed pre-nominal modification.

Table 2 shows maximum likelihood estimates (MLE) of posterior model parameters and correlations for all four models, inferred separately for the CTSL and English datasets. The continuous and continuous-incremental models best fit the data even though predictions from all four models highly correlated with the empirical data.

Fig. 3 shows the posterior over parameter values for the continuous model on the CTSL data. In all four models, size was inferred to have higher cost than color on the CTSL data, a result not observed in previous evaluations of RSA models. In the continuous model, both for the CTSL and English data, the semantic value for size was surprisingly inferred to be higher than the semantic value for color, with overlapping distributions. The same semantic value pattern was observed in the continuous-incremental model on the CTSL data.

The continuous models (unlike the standard and incremental models) correctly predict higher probability of redundant

[2] All code, data, and materials can be accessed at `https://github.com/bwaldon/crossling_reference`

[3] Constraining this model so that the two modifier costs are equal results in almost the identical predictive power as with free costs.

Table 2: Maximum likelihood estimates of model parameters, and correlation with empirical data.

	CTSL						English					
	α	$C(u_{color})$	$C(u_{size})$	v^{color}	v^{size}	r	α	$C(u_{color})$	$C(u_{size})$	v^{color}	c^{size}	r
Standard RSA	6.22	<0.01	0.20			0.92	10.80	<0.01	0.37			>0.99
Continuous RSA	18.5	0.06	0.47	0.60	0.78	0.99	17.70	0.07	0.70	0.64	0.89	>0.99
Incremental RSA	2.60	<0.01	0.60			0.70	5.43	<0.01	0.06			>0.99
Cont. Incr. RSA	14.4	<0.01	0.35	0.66	0.99	0.84	12.8	0.35	0.15	0.78	0.63	>0.99

modification in the size sufficient condition in both CTSL and English. Empirical utterance probabilities are shown with maximum likelihood estimates of all four models' posterior predictives in Fig. 2.

To compare the continuous and continuous-incremental models quantitatively, we conducted a Bayesian model comparison on these models with both datasets and found a Bayes factor of >100, indicating decisive evidence in favor of the continuous over the continuous-incremental model.

General Discussion

We explored how signers of an emerging sign language use their language to refer. We showed that despite a lack of fully-fledged grammatical conventions, Central Taurus Sign Language exhibits a previously-attested pattern of referring expression production – the color/size asymmetry in redundant adjectival modification. Moreover, despite an overwhelming preference for expressing adjectival modification post-nominally, CTSL signers redundantly modify at rates similar to those observed in English, in which adjectival modification occurs pre-nominally.

These results are consistent with theories of referring expression production that link redundant modification to properties of the lexicon. On speaker-oriented variants of this view, the color/size asymmetry reflects differences in the production cost – broadly construed – of size vs. color adjectives; on the listener-oriented view, size and color are lexicalized concepts that differ fundamentally in referential utility.

Conversely, these results are in tension with theories that center the role of incremental communicative pressures, which may vary cross-linguistically according to morphemic linear ordering constraints. Contra previous findings (Rubio-Fernandez et al., 2021; Wu & Gibson, 2021), we did not observe that redundant modification was attenuated in a language that exhibits linear ordering preferences that mirror those of English. This is unexpected on accounts that posit that redundant adjectives, all else equal, are of lower communicative utility – and hence less likely to be produced – when following rather than preceding sufficiently-informative linguistic units in referring expression production.

We supplement this qualitative evaluation with a quantitative comparison of models that differ in their commitments regarding the role of incremental communicative pressures as well as regarding the relative production cost and referential utility of items in the lexicon. When evaluated against our CTSL data, this comparison favors a model – continuous RSA – whereby speakers compute payoff and cost for entire utterances rather than at the sub-utterance level. However, there are important caveats worth mentioning at this point.

First, the continuous RSA model that best fit our data features parameter values that stand in conflict with previous analyses. In particular, Degen et al. (2020) fit a continuous RSA model to data collected from a similar experimental paradigm and concluded that the semantic noise associated with size adjectives is greater than the noise associated with color adjectives. Moreover, their best-fit continuous RSA model features relatively similar production costs associated with color and size adjectives. In contrast, our best-fit model features the opposite pattern vis a vis semantic noise as well as a marked asymmetry in the production cost of color vs. size adjectives. This result highlights the need to collect production data from language users across a wide array of theoretically-interesting contexts.

Degen et al. (2020) demonstrate, for example, that RSA models that attribute the color/size asymmetry solely to speaker production cost are fundamentally incapable of capturing an equally well-attested pattern in the referring expression production literature, namely that rates of redundant modification increase in the presence of more competitor objects. The trials of our study do not differ from one another in this regard. In the future, intend to collect data from CTSL signers in a wider qualitative array of referring contexts.

Second, we evaluated 'incremental' models of referring expression production on a single, specific notion of incrementality, one that assumes that speakers a) plan referring expressions word-by-word; and b) compute the referential utility of a word by evaluating the utility of full-utterance continuations of that word (and its preceding linguistic context). Even within the sparse existing literature on 'incrementalized' variants of the RSA framework, neither of these assumptions are universally adopted (Cohn-Gordon, Goodman, & Potts, 2018), nor is there a principled reason to favor them a priori over alternatives, e.g., that incremental production proceeds at the sub-morphemic level.

Finally, we acknowledge that the two experiments were conducted in different modalities (signed vs. written), with possibly different reference resolution pressures applying. One avenue of future work is to test control groups in spoken, signed and written language to explore the extent to which modality differences affect redundant modification.

In sum, this work demonstrates the utility of investigating under-studied languages for the purpose of informing theories of redundant referring expression production. It also highlights the importance of using a wide variety of referential contexts for informing rigorous model comparison.

References

Arts, A., Maes, A., Noordman, L., & Jansen, C. (2011). Overspecification facilitates object identification. *Journal of pragmatics*, *43*(1), 361–374.

Belke, E. (2006). Visual determinants of preferred adjective order. *Visual Cognition*, *14*(3), 261–294.

Bock, K. (1987). An effect of the accessibility of word forms on sentence structures. *Journal of memory and language*, *26*(2), 119–137.

Cohn-Gordon, R., Goodman, N., & Potts, C. (2018). Pragmatically informative image captioning with character-level inference. *arXiv preprint arXiv:1804.05417*.

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*(1).

Degen, J., Hawkins, R. D., Graf, C., Kreiss, E., & Goodman, N. D. (2020). When redundancy is useful: A bayesian approach to "overinformative" referring expressions. *Psychological Review*. doi: 10.1037/rev0000186

Ergin, R. (2017). *Central taurus sign language: A unique vantage point into language emergence* (Doctoral dissertation). Psychology Department, Tufts University.

Ergin, R., Kursat, L., Hartzell, E., & Jackendoff, R. (2021). Central taurus sign language: On the edge of conventionalization.

Ergin, R., Senghas, A., Jackendoff, R., & Gleitman, L. (2020). Structural cues for symmetry, asymmetry, and non-symmetry in central taurus sign language. *Sign Language & Linguistics*, *23*(1-2), 171–207.

Ferreira, V. S., & Dell, G. S. (2000). Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive psychology*, *40*(4), 296–340.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Franke, M., & Jäger, G. (2016). Probabilistic pragmatics, or why bayes' rule is probably important for pragmatics. *Zeitschrift für sprachwissenschaft*, *35*(1), 3–44.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, *20*(11), 818–829.

Hawkins, R. X. (2015). Conducting real-time multiplayer experiments on the web. *Behavior Research Methods*, *47*(4), 966–976. doi: 10.3758/s13428-014-0515-6

Jara-Ettinger, J., & Rubio-Fernandez, P. (2021). The social basis of referential communication: Speakers construct physical reference based on listeners' expected visual search. *Psychological review*.

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

Kursat, L., & Degen, J. (2021). Perceptual difficulty differences predict asymmetry in redundant modification with color and material adjectives. *Proceedings of the Linguistic Society of America*, *6*(1), 676–688.

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

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

Rubio-Fernandez, P., Mollica, F., & Jara-Ettinger, J. (2021). Speakers and listeners exploit word order for communicative efficiency: A cross-linguistic investigation. *Journal of Experimental Psychology: General*, *150*(3), 583.

Sedivy, J. C. (2003). Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. *Journal of Psycholinguistic Research*, *32*(1), 3–23. doi: 10.1023/A:1021928914454

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*(3), 345.

Waldon, B., & Degen, J. (2021). Modeling cross-linguistic production of referring expressions. *Proceedings of the Society for Computation in Linguistics*, *4*(1), 206–215. doi: 10.7275/vsfn-t057

Wu, S. A., & Gibson, E. (2021). Word order predicts cross-linguistic differences in the production of redundant color and number modifiers. *Cognitive Science*, *45*(1), e12934. doi: 10.1111/cogs.12934