An incremental RSA model for adjective ordering preferences in referential visual context

Fabian Schlotterbeck and **Hening Wang**
University of Tübingen
Tübingen, Germany
{fabian.schlotterbeck,hening.wang}@uni-tuebingen.de

Abstract

We report data from a preference rating experiment that tested for conflicting effects of subjectivity and discriminatory strength on adjective ordering preferences in referential visual context. Results indicate that, if the communicative efficiency of an adjective is low in a given context, it is preferred later in a multi-adjective expression. To account for qualitative aspects of these data, we propose a novel computational model of incremental processing in the Rational Speech Act framework. What sets the model apart from previous approaches is that it assumes fully incremental interpretation, without the need to anticipate possible sentence completions.

1 Introduction

In noun phrases (NPs) with multiple adjectives, as in (1), the relative order of the adjectives can vary, but at the same time, there are robust cross-linguistic preferences (Sproat and Shih, 1991) such that certain adjective sequences are more common and perceived as more natural than others. For example, the ordering in (1-a) is strongly preferred to that in (1-b).

(1) a. big white bear
b. white big bear

Although adjective ordering preferences have been known and studied for some time, they have resisted a unified explanation. Existing explanations come from different perspectives in linguistics and include semantic hierarchies (Dixon, 1982), syntactic mapping (Cinque, 1993) and psycholinguistic explanations based on as absoluteness (Martin, 1969) or closeness to the meaning of head noun (Whorf, 1945). Here, we focus on two recent hypotheses (Scontras et al., 2017; Fukumura, 2018, see next section for explanation) that have gained support from experimental work and share a common theoretical motivation. In particular, they are both based on the idea that efficiency in communication determines ordering preferences. Despite being based on the same general idea, theses hypotheses may lead us to expect significantly divergent outcomes in certain contexts. To address this tension, we pit these predictions against each other in a preference rating experiment. Furthermore, we implement both hypotheses in a novel computational model of incremental interpretation in the Rational Speech Act (RSA, Frank and Goodman, 2012) framework that not only provides a qualitative explanation of our findings but also sheds light on the relative contribution of the two hypotheses.

2 Two rational explanations of adjective ordering

The first explanation we focus on was proposed by Scontras et al. (2017), who showed that the subjectivity of adjectives is a strong predictor of ordering preferences. We call this the SUBJECTIVITY hypothesis. They operationalized subjectivity as *faultless disagreement*, roughly the degree to which two speakers can disagree about attributing a property to an individual without one of them necessarily being wrong. According to the SUBJECTIVITY hypothesis, (1-a) is preferred over (1-b) because *big* is more subjective than *white* and is also further away from the noun. In fact, gradable dimension adjectives like *big*, *tall* or *heavy* are prime examples of subjective adjectives that have received a lot of attention in previous work. We therefore focus the following discussion on these instances. In subsequent work, Scontras et al. (2019) proposed that the low communicative efficiency of subjective expressions is one possible reason for effects of subjectivity on ordering preferences. The main idea of Scontras et al. (2019) is that more efficient expressions are integrated earlier in the hierarchical structure underlying semantic composition in order to minimize the risk of misidentification of referents, and thus, as a result, these expressions

Proceedings of the Society for Computation in Linguistics (SCiL) 2023, pages 121-132.
Amherst, Massachusetts, June 15-17, 2023

end up closer to the modified noun in the linear sequence (at least in languages with prenominal modification).

The SUBJECTIVITY hypothesis has gained support from corpus studies as well as preference rating experiments in a variety of languages (Scontras et al., 2020b). Furthermore, the idea that communicative efficiency is increased if the more subjective expressions enter later into compositional meaning derivations was corroborated in computational simulations of rational communication (Simonic, 2018; Franke et al., 2019, see section 5 for discussion).

Another explanation of ordering preferences was given by Fukumura (2018), who investigated the impact of the *discriminatory strength* of adjectives. In a given context, a referring expression has greater discriminatory strength if it contains more information about the intended referent. If it singles out the intended referent perfectly, it has maximal strength. The main idea of the DISCRIMINATORY STRENGTH hypothesis is that the more discriminatory an adjective is, the more salient and accessible it will be in a visual context and also the more useful for reference resolution. Consequently, there will be a higher likelihood of early mention in the linear sequence (and thus greater distance from the noun in prenominal modification).

Fukumura (2018) tested the DISCRIMINATORY STRENGTH hypothesis in a production experiment where participants described referents that were marked in visual context. Discriminatory strength was controlled by manipulating the properties of the presented objects. In addition, color adjectives were compared to adjectives describing patterns, e.g. *striped*. As expected based on previous studies, Fukumura (2018) found that color adjectives were preferred before pattern adjectives and she explained this by a high availability of color adjectives in production. In addition, she found that discriminatory strength had the predicted effect and higher discriminatory strength in context led to earlier mention in the participants' productions. However, since there is no strong subjectivity gradient between color and pattern adjectives, her results do not speak to the SUBJECTIVITY hypothesis and the question remains open how these two hypotheses are related to each other.

3 Relation between SUBJECTIVITY and DISCRIMINATORY STRENGTH

Both the SUBJECTIVITY and the DISCRIMINATORY STRENGTH hypothesis are based on the idea that ordering preferences emerge from pressures towards efficient communication and both of them assume that more informative expressions are in some sense used "**earlier**". However, the two hypotheses take different perspectives and thus arrive at different definitions of what "early" means. In particular, the SUBJECTIVITY hypothesis is derived from the perspective of a listener whereas DISCRIMINATORY STRENGTH assumes a speaker perspective. The listener aims to identify an intended referent by sequentially restricting a set of potential referents in a process that follows the compositional semantic structure of a given expression. Thus, the listener evaluates the adjective that is closer to the noun first (thereby interpreting (1-a) as referring to bears that are big for white bears). As a consequence, the hierarchical structure of the NP determines what counts as "early" in the SUBJECTIVITY hypothesis. The speaker, by contrast, aims to maximize informativity at each step in the word-by-word production of an utterance. In the DISCRIMINATORY STRENGTH hypothesis, the position in the linear sequence of words is thus central. For these reasons, "earlier" translates to either **closer to the noun** or **further away from the noun**, depending on which perspective we take.

This is, in fact, a striking difference between the SUBJECTIVITY and the DISCRIMINATORY STRENGTH hypothesis and it is an interesting empirical question what happens if these two perspectives stand in direct conflict to each other. This could, e.g., be the case in a context in which a less subjective adjective discriminates more strongly than a more subjective one between the intended referent and a set of distractors. This exact question is the main question we addressed in the experiment reported in the next section, in which participants indicated their preferences between multi-adjective expressions like in (1) when referring to a target referent in visual context.

To appreciate the purpose and limitations of our experiment, it may be worthwhile to reflect briefly on the predictions that can be derived from the SUBJECTIVITY hypothesis in the type of contextually-embedded experimental setting underlying the DISCRIMINATORY STRENGTH hypothesis of Fukumura (2018). We acknowledge that, strictly speak-

ing, the SUBJECTIVITY hypothesis, by itself, does not predict how preferences are affected by manipulations of visual context. This is because SUBJECTIVITY does not presuppose that subjective-first expressions are less informative in every setting. There only need to be enough such instances overall for a general preference to "evolv[e] gradually" (Franke et al., 2019; cf. also Scontras, 2023). Thus, the SUBJECTIVITY hypothesis explicitly allows for counterexamples. One such counterexample is the case where a multi-adjective expression like in (1) receives a conjunctive instead of the assumed "sequentially intersective" reading (cf. Franke et al., 2019), such that (1-a) would be understood as referring to bears that are white and big (for bears) rather than big for white bears. In fact, Scontras et al. (2020a) presented empirical evidence that the preference for subjective-first orderings vanishes when adjectives restrict the set of potential referents in conjunction. We cannot exclude the possibility that the specific design of our current experiment constitutes another counterexample, maybe even because conjunctive readings are favored in our design. Be this as it may, a gradual evolution of the SUBJECTIVITY-based preferences that are commonly observed would be extremely challenging to explain based on low informativity of subjective expressions if we find empirically that speakers actually adapt by producing subjective adjectives more often in first position (in the linear sequence) if context renders them more (rather than less) informative.

4 Experimental Data: Preference ratings in visual contexts

4.1 Method

In a web-based experiment, we collected data on adjective ordering preferences in German using preference ratings of multiple adjective sequences in visual referential context. Participants (N=120) were recruited via the platform *prolific.co*. They were instructed at the begin of the experiment by a cover story that they should communicate a target sticker (marked with a red box, see Fig. 1) in a scrapbook to an imagined listener on a telephone call. With this setting, we aimed to rule out the possibilities of using information of relative spatial positions in the context and tried to simulate an online communication situation as closely as possible. In each experimental trial, participants were presented with a visual context and they indicated their preference between two sentences with reversed adjective order using a slider in the middle of the screen (see Fig. 1).

In a mixed factorial design, we manipulated, within participants, the COMBINATION of adjectives from different semantic classes (levels: *dimension & either color or shape* and *color & shape*) and the RELEVANCE of the corresponding properties for reference resolution, i.e. whether the *first*, *second* or *both* properties were needed to identify a referent (cf. Fig. 1). [1] The purpose of these two factors was to test whether the basic findings of Fukumura (2018) replicate also with subjective adjectives and, in particular, whether the preference for subjective adjectives in first position persists if the more subjective adjective has the lesser discriminatory strength.

In addition to this within-participants manipulation, we also manipulated the SIZE DISTRIBUTION of objects (*sharp* vs. *blurred*) between-participants. As in Fig. 1, there were always six objects in the visual context that were either large or small. The large objects had sizes that were randomly sampled from the integers 9 and 10 (in some arbitrary unit of length that effectively depended on the display settings of the experimental participants). If size was the relevant property, the target object was always the biggest, irrespective of SIZE DISTRIBUTIONS. In *sharp* SIZE DISTRIBUTION, sizes of the remaining, small objects were sampled from the integers in the range $[1, 3]$ whereas they were sampled from $[1, 6]$ in *blurred* SIZE DISTRIBUTION. As a result, the small objects in the *blurred* as compared to the *sharp* distribution had greater variance in size among them and a smaller mean distance to the sizes of the big objects. The idea behind this manipulation was to affect the information that size adjectives could convey in such a way that they are more useful in *sharp* vs. *blurred* distributions. In particular, we intended to make size adjectives effectively non-subjective in *sharp* distributions. If any prediction about the effect of this manipulation can be derived from the SUBJECTIVITY hypothesis (see discussion above), the preference

[1] The factor COMBINATION was originally a three-level factor with the levels *dimension & color*, *dimension & shape* and *color & shape*. We aggregated the first two levels here because they did not differ significantly and their distinction is not relevant for our present purpose, in particular for the computational models described in section 6. The complete design and statistical analysis along with a free production experiment in the same general design is described in the unpublished MA thesis of Wang (2022).

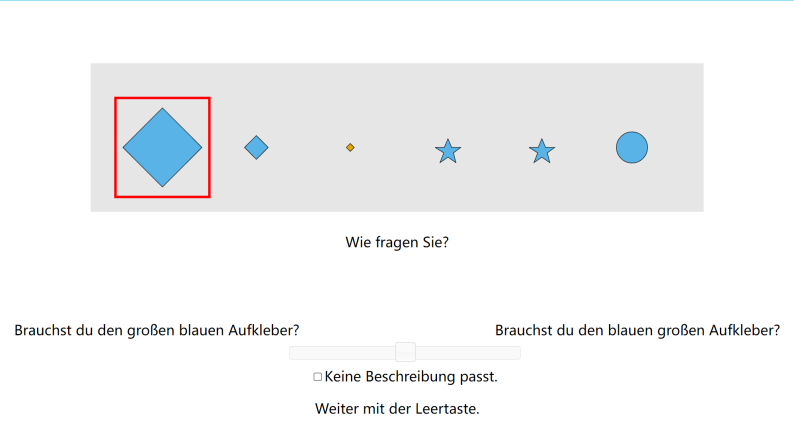


Figure 1: An example item from the current experiment in the condition with COMBINATION of *dimension and color* adjectives and RELEVANCE of the *first* property (i.e. *dimension*) in a *sharp* SIZE DISTRIBUTION. A property was counted as relevant if it was necessary for referent identification. In this example size is relevant but color and shape are not. Glosses for the German linguistic material in the example item are provided in Appendix A.

for subjective-first orders should therefore be weakened in *sharp* SIZE DISTRIBUTION. The reason is that the SUBJECTIVITY hypothesis assumes that less subjective adjectives are integrated earlier into the hierarchical structure.

We generated 27 experimental items in each of the 18 conditions, resulting in a total of 486 items that were distributed across 6 lists (three per SIZE DISTRIBUTION). Each participant saw a total of 81 experimental items. These were combined with 99 filler items that were constructed in a similar way as the experimental items but also included sentences with only one adjective instead of two. Overall, each participant thus completed 180 trials. An experimental session took around half an hour and participants received reimbursement of 5.25 £.

4.2 Results

The mean slider positions are shown in Fig 2. For statistical analysis, we used linear mixed effects models (Bates et al., 2015) that incorporated fixed effects of all manipulated factors and their interactions, along with random intercepts for participants and items. For hypothesis testing, we used model comparisons based on log-likelihood ratio tests. First of all, our results replicate effects of SUBJECTIVITY: There was a strong preference for dimension adjectives in first position which resulted in a significant effect of COMBINATION on slider ratings ($\chi^2(1) = 361.97, p < .001$). Furthermore, we found a significant interaction between RELEVANCE and SIZE DISTRIBUTION ($\chi^2(2) = 21.26, p < .001$). This interaction was due to the fact that there was a preference for orderings with adjectives that are needed (and sufficient) for reference resolution in first position (i.e. an effect of RELEVANCE) and this preference was more pronounced in *sharp* ($\chi^2(1) = 385.91, p < .001$) as compared to *blurred* SIZE DISTRIBUTIONs ($\chi^2(1) = 222.49, p < .001$). Since we had specific expectations concerning the effect of SIZE DISTRIBUTION on the preference for orderings with subjective adjectives in first position, we split the data according to the factor COMBINATION and performed separate analyses on *dimension and X* and *color and form* combinations. In both cases, the interaction between RELEVANCE and SIZE DISTRIBUTION turned out to be significant but for different reasons: In combinations of *dimension and X*, the preference for subjective-first orderings in dimension-relevant contexts was increased in *sharp* as compared to *blurred* DISTRIBUTIONs ($\beta = 0.58, \chi^2(2) = 19.50, p < .001$). In combinations of *color and form* adjectives, *sharp* in comparison to *blurred* distributions led, by contrast, to an increased preference for form-first orderings (the 2nd property in the COMBINATION *color and form*) in form-relevant contexts ($\beta = -0.71, \chi^2(2) = 8.29, p = 0.016$). The former of these two interactions was directly relevant to our hypotheses whereas the latter was completely unexpected and we do not have an explanation for it.

4.3 Discussion

We replicated both the SUBJECTIVITY and DIS-

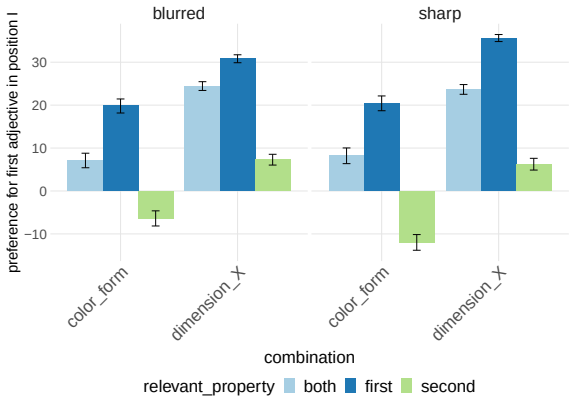


Figure 2: The mean slider positions from the current experiment: The slider had an initial value of 0 and potential values ranged between +50 and -50. A positive value indicates a preference for the first adjective in a COMBINATION (i.e. the color adjective in the combination *color and form* or the dimension adjective in the combination *dimension and x*, where x stands for either color or form) at the first position in the linear sequence. For combinations involving dimension adjectives (labeled *dimension_x*), a positive value indicates the conventional subjective-first order and a negative value shows the opposite.

CRIMINATORY STRENGTH effects in our study, which suggests that more than one source can contribute to adjective ordering preferences, especially in visual contexts. We manipulated the communicative efficiency of subjective adjectives by varying discriminatory strength of the size property and varying size distributions of contrast objects in visual contexts. Contrary to the predictions we derived from Scontras et al. (2019), our present results indicate that the robust preference for subjective-first orderings cannot be easily explained by communicative efficiency alone (cf. section 3).

5 Previous modeling approaches

Below, we propose a novel incremental model of interpretation in the RSA framework (Frank and Goodman, 2012; Scontras et al., 2018) in order to account for qualitative aspects of our experimental findings. In doing so, we build on previous models, but also highlight differences between the current and previous approaches.

In order to explain subjectivity-based ordering preferences, computational models of communication were used in recent research. The model we propose in the following section builds on some of these previous proposals (in particular, Simonic, 2018, Scontras et al., 2019 and Franke et al., 2019) that are closely related in spirit to referential communication in the RSA framework (but see also Hahn et al., 2018, for a slightly different approach). The general agreement among these approaches is that less subjective content is more effective in conveying intended meanings because it is more likely to be interpreted in the same way by listeners and speakers. Among the mentioned approaches, Franke et al. (2019) is closest to the standard, vanilla RSA model and it thus serves as a reference point for us.

Furthermore, the model of Cohn-Gordon et al. (2019) is also directly relevant for the current work. In their model a literal listener constructs meanings incrementally at each word by considering all possible completions of the sentence. This type of incremental RSA model was also combined with a continuous semantics (as proposed by Degen et al., 2020) to account for the tendency of English speakers to produce more over-specified expressions with color adjectives than with size adjectives (Waldon and Degen, 2021). However, while these incremental models can address some aspects of the production of referring expressions, they do not directly address ordering preferences for multiple adjectives and, in fact, cannot account for them for reasons we explain below.

6 A fully incremental model of interpretation

Both the SUBJECTIVITY hypothesis and the DISCRIMINATORY STRENGTH hypothesis explain ordering preferences by means of incremental processes. They differ, however, in the perspective they take. The SUBJECTIVITY hypothesis takes the perspective of a listener who performs a sequentially intersective context update in order to identify an intended referent. By contrast, the DISCRIMINATORY STRENGTH hypothesis takes the perspective of an incremental speaker who maximizes information at each step in the word-by-word production of an utterance. In order to see whether these two perspectives (combined or separately) can account for the effects we observed in our preference rating experiment, we implemented a version of an incremental listener as well as an incremental speaker in a fully incremental probabilistic computational model in the RSA framework and compared qualitative modeling results to our empirical

observations. In particular, we compared the listener and speaker perspectives and asked whether one of them or both in combination can account for our qualitative results. In what follows, we focus on the experimental conditions involving dimension adjectives because all relevant effects were found in these conditions. Furthermore, we do not distinguish between color and shape adjectives as we did not find significant differences between them when they were combined with dimension adjectives.

In the vanilla RSA model (Frank and Goodman, 2012; see Scontras et al., 2018 for review), the literal listener, L_0, infers an intended referent r by combining prior expectations, $P(r)$, about what the referent will be with the literal meaning, $[\![u]\!]$, of an utterance u according to the proportionality in (2). The listener thus updates prior expectations by filtering out all potential referents that are incompatible with the literal meaning of the utterance. The speaker, S_1, on the other hand, tries to maximize communicative utility by trading off the information an utterance provides about the intended referent (measured in its surprisal $-\log(L_0(r|u))$) against its production cost, $C(u)$. This is done by choosing utterances according to the soft-max decision rule in (2-b), where α determines how rational a speaker is in choosing between utterances.

(2) a. $L_0(r|u) \propto [\![u]\!](r) \cdot P(r)$
b. $S_1(u|r) \propto \exp(\alpha \cdot (\log L_0(r|u) - C(u)))$

We extend the vanilla RSA model in a number of ways to account for our empirical observations. The main innovations are (i) a fully incremental literal listener, who performs a sequentially intersective context update that respects the hierarchical structure underlying semantic composition (i.e. it interprets German multi-adjective sequences from right to left), and (ii) a fully incremental speaker, who produces one word after the other (from left to right). In principle, these two innovations allow us to capture ordering preferences because they break the symmetry that is usually assumed in the compositional operations used to interpret multi-adjective sequences. In contrast to previous incremental approaches (Cohn-Gordon et al., 2019; Waldon and Degen, 2021), we propose a model that allows for truly incremental processing without the need to anticipate possible sentence completions.

The incremental literal listener is defined in the recursion in the first two rows in Table 1. Applied to a single-word utterance, this is just the standard literal listener from the vanilla RSA model, with the added feature of potentially context-dependent meanings. In particular, it allows for word meanings that vary with the support of the prior probability over possible states (i.e. a distribution over potential referents in our case), $P(r)$. This feature is important for two reasons.

Firstly, gradable adjectives are well-known to have context-dependent interpretations, which have been accounted for in previous computational models in various ways (e.g. Lassiter and Goodman, 2017; Qing and Franke, 2014). Here, we adopt the so-called k%-semantics in (3-a) because it has been shown in previous work (Schmidt et al., 2009; Cremers, 2022) to match speakers' judgments remarkably well and allows for a comparison with Franke et al. (2019), who used this semantics as well. Under this semantics an individual is considered tall if its height exceeds that of k% of the individuals in the comparison class C. The k% semantics was combined with a 'perceptual blur' such that perceived sizes deviated from the ground truth according to the Weber-Fechner law (implemented as in van Tiel et al., 2021). For color adjectives, we assumed the continuous semantics in (3-b) as proposed by Degen et al. (2020). According to (3-b) categorization is imperfect in the sense that blue objects may be judged as non-blue with probability ϵ and vice versa. In the following, a relatively low value of .02 was assumed for ϵ throughout.

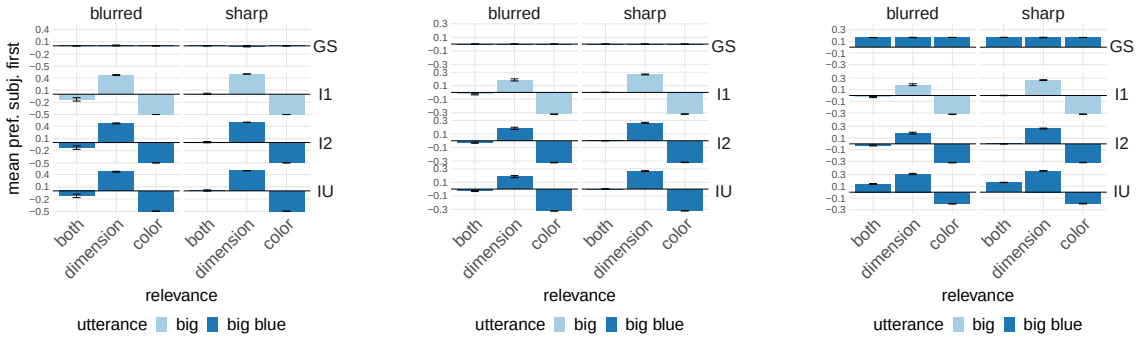
(3) a. $[\![\text{big}]\!]^C = \lambda x.size(x) > max(C) - k/100 * (max(C) - min(C))$
b. $[\![\text{blue}]\!] = \lambda x. \begin{cases} 1-\epsilon & \text{if x is blue,} \\ \epsilon & \text{if x is not blue} \end{cases}$

Secondly, the definition in Table 1 implies that the incremental listener cannot distinguish between different orders if none of the involved meanings depends on the result of the previous step in the sequential update. As a sanity check, we have verified this theoretical result by treating dimension adjectives exactly as color adjectives, using the semantics in (3-b) for them as well.

The global speaker in Table 1 functions as in the vanilla RSA model but produces utterances according to a utility function $\mathbb{U}(\overrightarrow{w}; r)$ (row 7 in Table 1) that is based on the incremental listener. This global speaker contrasts with the incremental sequence speaker, defined in rows 4 and 5 of the table, which maximizes informativity at each word. The latter is a probabilistic speaker that pro-

(1)	Incremental Listener	$L_0^{inc}(r\|w_{1,n})$	$\propto [\![w_1]\!]^{\text{supp}(L_0^{inc}(\cdot\|w_{1,n-1}))}(r) \cdot L_0^{inc}(r\|w_{1,n-1})$
(2)		$L_0^{inc}(r\|w_1)$	$\propto [\![w_1]\!]^{\text{supp}(P)}(r) \cdot P(r)$
(3)	Global Speaker	$S_1(w_{1,n}\|r)$	$\propto \mathbb{U}(w_{1,n}; r) \cdot P(w_{1,n})$
(4)	Incremental Sequence	$S_1^{inc}(w_{1,n}\|r)$	$\propto \mathbb{U}(w_{1,n}; r) \cdot P_{Lang}(w_n\|w_{1,n-1}) \cdot S_1^{inc}(w_{1,n-1}\|r)$
(5)	Speaker	$S_1^{inc}(w_1\|r)$	$\propto \mathbb{U}(w_1; r) \cdot P_{Lang}(w_1\|\varnothing)$
(6)	Incremental Utterance Speaker	$S_1^{inc\_utt}(w_{1,n}\|r)$	$\propto \exp(\alpha \cdot (\log(S_1^{inc}(w_{1,n}\|r)))) \cdot P(w_{1,n})$
(7)	Utility	$\mathbb{U}(\overrightarrow{w}; r)$	$= \exp(\beta \cdot (\log(L_0^{inc}(r\|\overrightarrow{w})) - \text{c}(\overrightarrow{w})))$

Table 1: Model definitions for the Incremental Listener (rows: 1 & 2), the Global Speaker (row: 3; GS in Fig. 3), the Incremental Sequence Speaker (rows: 4 & 5; I1 and I2 in Fig. 3), and the Incremental Utterance Speaker (row: 6; IU in Fig. 3). All speaker models depend on the utility function $\mathbb{U}$ in (7). In all the definitions, r stands for a referent; w_1, $w_{1,n}$ and $\overrightarrow{w}$ stand for the first word in a sequence, a sequence of n words and any sequence of one or more words, respectively; $\text{supp}(\cdot)$ denotes the support of a probability distribution; P denotes prior probabilities over referents and utterances; P_{Lang} assigns prior probabilities to potential next words in a sequence; and, finally, α and β are rationality parameters that govern the soft-max functions defined in rows (6) and (7), respectively. The parameter β was set to 1 in all reported simulations. In addition we used a bias (b in Fig. 3) in the prior $P(w_{1,n})$ of $S_1^{inc\_utt}$. The bias determines how much more likely the subjective-first ordering is *a priori*.



(a) $\alpha = 5; b = 1$ (b) $\alpha = 1; b = 1$ (c) $\alpha = 1; b = 2$

Figure 3: Simulations of preferences for the experimental stimuli (labeling of conditions as in Fig. 2) with different values of α, and the bias for subjective-first orders, b. In each plot, the first row shows results for the global speaker, the second and third row represent the sequence speaker distributions for one- and two-word sequences, respectively, and the fourth row represents the incremental utterance speaker. The y-axes show probabilities shifted to $[-.5, .5]$.

duces n-word sequences by recursively sampling from a sequence speaker for length $n-1$, generating a continuation word and evaluating this, as before, using the utility function $\mathbb{U}(w_{1,n}; r)$. The next word in each step is generated by a language model, $P(w_n|w_{1,n-1})$, that is extremely simple in the present case: It produces either a dimension or color adjective as the first word and then generates the other alternative in the next step. Thus, our two candidate utterances *big blue* and *blue big* are generated with equal frequency prior to factoring in the utility function. Finally, the incremental utterance speaker chooses between alternative utterances by sampling from a prior distribution over candidate utterances (*big blue* and *blue big* in our case) and reweighing their probabilities according to the sequence speaker. In the utterance prior, we used a bias parameter, b, to encode an *a priori* preference for the subjective-first ordering.

6.1 Results and discussion

The model was implemented and simulated using the probabilistic programming language WebPPL (Goodman and Stuhlmüller, 2014). We applied the model to all our stimuli from the conditions that involved dimension adjectives and tested various parameter settings. We report those that best represent the general picture that emerged. We did not find significant deviation from this general pattern for any of the parameter sets we tried. Posterior distributions were inferred using MCMC simulation with 30000 samples (burn-in: 5000, lag: 3) for the incremental listener and sequence speaker and 15000 samples (burn-in: 3000, lag: 3) for the two

utterance speakers. All simulations had an ϵ of .02 for the semantics of color adjectives, a k of 50 for the dimension adjective and a Weber fraction of .5 for the perceptual blur.

In a first simulation, we chose a relatively large value for the rationality parameter, namely $\alpha = 5$, and assumed no bias for the subjective-first order in the utterance speakers (i.e. $b = 1$). The results of this simulation are shown in Figure 3a. We did not find any deviation from uniform preferences in the global speaker (whose preferences are determined by the incremental literal listener alone). In contrast, the other three components (i.e. the sequence speaker for one- and two-word sequences and the incremental utterance speaker) revealed effects of SIZE DISTRIBUTION and also showed the characteristic effects of DISCRIMINATORY STRENGTH. [2] The effect of SIZE DISTRIBUTION was more pronounced in the conditions in which both properties were relevant than in conditions in which only one was relevant. This was because the preferences were at ceiling in the latter four conditions, revealing strong effects of discriminatory strength. Nevertheless, there was still a small effect of SIZE DISTRIBUTION in the conditions in which the dimension adjective was relevant, matching another aspect of our empirical observations.

To attenuate preferences in the conditions in which only one adjective was relevant for reference resolutions, we ran the same simulation with lower α. The results are shown in Fig. 3b. As before, effects are limited to the incremental speaker components of our model and there are again effects of both SIZE DISTRIBUTION and DISCRIMINATORY STRENGTH. As expected, extreme preferences are attenuated compared to the first simulation. This led to a preference pattern in which the effect of SIZE DISTRIBUTION is almost completely restricted to the dimension relevant conditions. Besides this effect, there are still relatively large effects of DISCRIMINATORY STRENGTH. Both of these aspects match our empirical observations. The absolute preferences, on the other hand, do not. This can, e.g., be seen by the negative values in the color-relevant and balanced preferences in the both-relevant conditions.

Absolute preferences were adjusted in a third simulation using a bias of $2 : 1$ ($b = 2$) for the subjective-first order. The resulting preferences are shown in Fig 3c. They matched our empirical observations better but still not perfectly. One notable deviation from our empirical observations consists in preferences for the subjective-last order in the color relevant conditions.

While it would be possible to shrink this deviance further using yet different parameter values, we think that this is beyond the scope of the present qualitative analysis. What our result provide, though, is initial indication concerning the region of the parameter space that may be worth examining further in a quantitative analysis. One first step towards such an analysis would be to specify a linking function between the production preferences of the model and the slider values we observed in the experiment. Their relationship may well be non-linear and could thus lead to compressed slider values in some regions.

One surprising result is that we did not find any effects whatsoever in the incremental listener component of the model. We investigated this issue further in two directions. Firstly, we used a different semantics for the dimension adjectives when modeling our experimental stimuli. This semantics was based on the identification of large and small objects based on the optimal breaks algorithm of Jenks (1967) akin to the cluster-based semantics in Schmidt et al. (2009). Secondly, we generated up to 350 random stimuli by sampling sizes from a Gaussian and colors from a Binomial distribution and modeled these stimuli using various sets of parameters (e.g. larger values of α and different values for k). We did, however, not find pronounced preferences in any of these attempts. We see two potential reason for this discrepancy between previous models (Simonic, 2018; Scontras et al., 2019; Franke et al., 2019) and the current results: It could be due to limited sample size in the present simulations or to the fact that previous models implemented different assumptions, (e.g. applying a threshold-based semantics also to color adjectives, as in Franke et al., 2019).

7 General discussion and outlook

We showed that a qualitative account of our data can be given by means of an incremental speaker that maximizes informativity at each word in combination with a general preference for subjective-first sequences. This does not imply that the general preference for subjective-first sequences is not driven by pressures towards efficient communica-

[2] We refrain from reporting statistical analyses because we did not perform a quantitative analysis and existing qualitative effects can be boosted by increasing rationality parameters.

tion in sequential context updates, as was proposed by previous studies. However, as noted in section 3, an explanation along these lines has to acknowledge the type of adaptation we observed in the current preference rating experiments. In particular, participants used subjective expressions earlier in the linear sequence if they were more informative about the intended referent. Based on the current empirical and modeling results, we would like to suggest an alternative explanation of how preferences for subjective-first sequences may emerge, at least for dimension adjectives. Such adjectives are commonly thought of as being used to communicate properties that deviate from the norm. This implies that, when they are used, they tend to have high discriminatory strength and may therefore be produced early in the linear sequence.

What we did not observe in our incremental model are truly incremental effects, i.e. shifts in preferences between one word and the next. Instead, preferences were due to an utterance-level prior in combination with a tendency to start the sequence with an informative word. The reason that incremental effects did not emerge in the current setting was that there were no strong ordering preferences on the listener side that could have modulated any initial biases. Other types of incremental effects may emerge if there are different numbers of continuations depending on how an utterance was started. Such effects were discussed, e.g., by Cohn-Gordon et al. (2019) and they can be reproduced in the current model.

Previous incremental RSA models (Cohn-Gordon et al., 2019; Waldon and Degen, 2021) were based a non-incremental semantics and evaluated all possible sentence completions of a given sentence beginning at each step. This is a natural approach because compositional semantic models often only provide interpretations for complete sentences. In contrast, our listener model evaluates an utterance word-by-word from right to left in line with the assumed sequential context update of multi-adjective strings. The more general idea behind our model is to use a genuinely incremental semantics (as proposed, e.g., by Bott and Sternefeld, 2017) that implements the local evaluation of yet incomplete sentences in a systematic fashion while ensuring that the interpretation of the complete utterance will conform to its standard compositional interpretation. We view our model as an instantiation of this general approach.

An interesting question is how much of our present considerations can be extended to non-definite noun phrases, where ordering preferences seem to persist but the current informativity-based notions do not apply directly because they are tailored to referential communication and the identification of intended referents. [3] Firstly, we see no reason to rule out the possibility that the bias we assumed in order to explain the general (context-independent) preference for subjective-first orders can be extended to non-referential usages right away. Secondly, we think that considerations based on (context-dependent) informativity might also generalize to non-definite noun-phrases. From the perspective of Generalized Quantifier Theory (Barwise and Cooper, 1981), for example, a modified noun in a quantified noun phrase provides the restriction of the quantifier and the meaning of a quantified sentence, like e.g. *many of the big white bears are moving south* is determined by two specific cardinalities: that of the set of elements that are both in the restriction and in the so-called nuclear scope of the quantifier (e.g. the big white bears moving south) and that of the set of elements that are in the restriction but not in the scope (e.g. the big white bears not moving south). Obviously, the relevant sets have to be identified first in order to determine these cardinalities. Informativity-based notions may, in principle, affect the amount of errors that are expected during this process, both from the perspective of a listener performing sequentially intersective updates as well as the perspective of a speaker aiming to provide the most discriminatory information first (see van Tiel et al., 2021, for an RSA model of quantifier interpretation).

Similarly, one might wonder how the present results generalize beyond nominal modification to the modification of verb phrases or even entire propositions (see, e.g., Specht and Stolterfoht, 2023, for an experimental investigation). While some of the present considerations might generalize to such cases, they also pose significant challenges to our present approach. In particular, such modification often involves properties that are fairly abstract or intensional in nature and are, therefore, difficult to control by means of contextual manipulations. Whether the present approach can be extended to cover such cases as well thus remains to be seen.

[3] We would like to thank an anonymous reviewer for raising this question.

Acknowledgements

We would like to thank Britta Stolterfoht, Michael Franke and three anonymous reviewers for helpful discussion and comments. FS received funding from the Baden-Württemberg Ministry of Science (MWK-BW) and the Federal Ministry of Education and Research (BMBF) as part of the Excellence Strategy of the German Federal and State Governments.

References

John Barwise and Robin Cooper. 1981. Generalized quantifiers and natural language. *Linguistics and Philosophy*, 4(2):159–219.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Oliver Bott and Wolfgang Sternefeld. 2017. An event semantics with continuations for incremental interpretation. *Journal of Semantics*, 34(2):201–236.

Guglielmo Cinque. 1993. On the evidence for partial N-movement in the Romance DP. *Working Papers in Linguistics, 3.2, 1993, pp. 21-40*.

Reuben Cohn-Gordon, Noah Goodman, and Christopher Potts. 2019. An incremental iterated response model of pragmatics. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, pages 81–90.

Alexandre Cremers. 2022. Interpreting gradable adjectives: rational reasoning or simple heuristics? In *Empirical Issues in Syntax and Semantics 14*, pages 31–61, Paris.

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*, 127(4):591–621.

Robert M.W. Dixon. 1982. *Where have all the adjectives gone?: and other essays in semantics and syntax (Vol. 107)*. Janua Linguarum. Series Maior. Walter de Gruyter, Berlin, New York.

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

Michael Franke, Gregory Scontras, and Mihael Simonic. 2019. Subjectivity-based adjective ordering maximizes communicative success. In *Proceedings of the 41st Annual Meeting of the Cognitive Science Society*, pages 344–350.

Kumiko Fukumura. 2018. Ordering adjectives in referential communication. *Journal of Memory and Language*, 101:37–50.

Noah D. Goodman and Andreas Stuhlmüller. 2014. The Design and Implementation of Probabilistic Programming Languages. http://dippl.org. Accessed: 2023-2-22.

Michael Hahn, Judith Degen, Noah D Goodman, Dan Jurafsky, and Richard Futrell. 2018. An information-theoretic explanation of adjective ordering preferences. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society*.

G. F. Jenks. 1967. The data model concept in statistical mapping. *International Yearbook of Cartography*, 7:186–190.

Daniel Lassiter and Noah D. Goodman. 2017. Adjectival vagueness in a Bayesian model of interpretation. *Synthese*, 194:3801–3836.

James E. Martin. 1969. Semantic determinants of preferred adjective order. *Journal of Verbal Learning and Verbal Behavior*, 8(6):697–704.

Ciyang Qing and Michael Franke. 2014. Gradable adjectives, vagueness, and optimal language use: A speaker-oriented model. In *Proceedings of the 24th Semantics and Linguistic Theory Conference*, volume 24, pages 23–41. Linguistic Society of America.

Lauren A. Schmidt, Noah D. Goodman, David Barner, and Joshua B. Tenenbaum. 2009. How tall is tall? compositionality, statistics, and gradable adjectives. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, pages 2759–2764.

Gregory Scontras. 2023. Adjective ordering across languages. *Annual Review of Linguistics*, 9(1):357–376.

Gregory Scontras, Galia Bar-Sever, Zeinab Kachakeche, Cesar Manuel Rosales Jr, and Suttera Samonte. 2020a. Incremental semantic restriction and subjectivity-based adjective ordering. In *Proceedings of Sinn und Bedeutung*, volume 24, pages 253–270.

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*, 12:7.

Gregory Scontras, Z. Kachakeche, A. Nguyen, C. Rosales, S. Samonte, E. Shetreet, Y. Shi, Elli N. Tourtouri, and N. Trainin. 2020b. Cross-linguistic evidence for subjectivity-based adjective ordering preferences. Talk presented at the workshop on Theoretical and Experimental Approaches to Modification (TExMod2020), held at the University of Tübingen.

Gregory Scontras, Michael H. Tessler, and Michael Franke. 2018. Probabilistic language understanding: An introduction to the Rational Speech Act framework. https://www.problang.org. Accessed: 2023-2-22.

Mihael Simonic. 2018. Functional explanation of adjective ordering preferences using probabilistic programming. Master's thesis, University of Tübingen.

Larissa Specht and Britta Stolterfoht. 2023. Processing word order variations with frame and sentence adjuncts in German: Syntactic and information-structural constraints. *Glossa: a journal of general linguistics*, 8(1).

Richard Sproat and Chilin Shih. 1991. The cross-linguistic distribution of adjective ordering restrictions. In Carol Georgopoulos and Roberta Ishihara, editors, *Interdisciplinary Approaches to Language: Essays in Honor of S.-Y. Kuroda*, pages 565–593. Springer Netherlands, Dordrecht.

Bob van Tiel, Michael Franke, and Uli Sauerland. 2021. Probabilistic pragmatics explains gradience and focality in natural language quantification. *Proceedings of the National Academy of Sciences*, 118(9):e2005453118.

Brandon Waldon and Judith Degen. 2021. Modeling cross-linguistic production of referring expressions. In *Proceedings of the Society for Computation in Linguistics 2021*, pages 206–215, Online. Association for Computational Linguistics.

Hening Wang. 2022. Subjektivität vs. diskriminatorische Stärke: Eine experimentelle Untersuchung zur Adjektivreihenfolgenpräferenz im visuellen Kontext. Master's thesis, University of Tübingen.

Benjamin Lee Whorf. 1945. Grammatical categories. *Language*, 21(1):1–11.

A Glosses for example item

(4) The question below visual contexts as part of the cover story in the current experiment (see. Fig. 1)

a. Wie fragen Sie?
how ask you
'How do you ask?'

(5) ...and questions on both sides of the slider for rating

a. Brauchst du den großen blauen Aufkleber?
need you the big blue sticker
' Do you need the big blue sticker?'

b. Brauchst du den blauen großen Aufkleber?
need you the blue big sticker
' Do you need the blue big sticker?'