Subjectivity-based adjective ordering maximizes communicative success

Abstract

Adjective ordering preferences (e.g., *big brown bag* vs. *brown big bag*) are robustly attested in English and many unrelated languages (Dixon, 1982). Scontras, Degen, and Goodman (2017) showed that adjective subjectivity is a robust predictor of ordering preferences in English: less subjective adjectives are preferred closer to the modified noun. In a follow-up to this empirical finding, Simonič (2018) and Scontras, Degen, and Goodman (to appear) claim that pressures from successful reference resolution and the hierarchical structure of modification explain subjectivity-based ordering preferences. We provide further support for this claim using large-scale simulations of reference scenarios, together with an empirically-motivated adjective semantics. In the vast majority of cases, subjectivity-based adjective orderings yield a higher probability of successful reference resolution.

Keywords: adjective ordering, subjectivity, reference resolution, hierarchical modification

Introduction

When speakers use two or more adjectives to modify a noun, they exhibit robust preferences in the relative order of the adjectives (e.g., *big brown bag* vs. *brown big bag*). Using a series of behavioral and corpus experiments, Scontras et al. (2017) demonstrated that adjective order in multi-adjective strings is reliably predicted by the subjectivity of the adjectives involved: less subjective adjectives are preferred closer to the modified noun, and the strength of the preference is modulated by the subjectivity differential between the adjectives. Thus, speakers strongly prefer *big brown bag* over *brown big bag*, as *brown* is much less subjective than *big*.

The question that immediately arises is why subjectivity should play the role it does in adjective ordering preferences. The current work follows Simonič (2018) and Scontras et al. (to appear) in advancing the claim that pressures from successful reference resolution deliver subjectivity-based ordering preferences. In certain cases of restrictive modification which proceed incrementally based on syntax-driven meaning composition, adjectives that compose with the nominal later will classify a smaller set of potential referents (e.g., the set of bags vs. the set of brown boxs). We demonstrate that, in order to avoid alignment errors where a listener might mischaracterize the intended referent, it is, when averaging over many contexts of use, a better strategy to introduce the more error-prone (i.e., more subjective) adjectives later in the hierarchical meaning composition; the structure linearizes such that subjectivity decreases the closer you get to the modified noun. We build on the work that precedes ours by making minimal assumptions about online processing (cf. Scontras et al., to appear) and by assuming a more principled implementation of adjective subjectivity within an empirically-motivated semantics (cf. Simonič, 2018).

The paper is structured as follows. First, we review the empirical generalization concerning subjectivity-based preferences, together with the proposals offered to account for this generalization. Then, we consider empirical work on adjective semantics, which serves as inspiration for our own proposal. We demonstrate, using Monte Carlo simulation, how a minimal set of independently-motivated assumptions leads to a ready explanation for subjectivity-based ordering preferences: ordering adjectives with respect to decreasing subjectivity has a higher probability of successful reference resolution, when averaging across many contexts of use.

Background

Given the robustness of adjective ordering preferences within and across languages, there has been no shortage of proposals meant to account for the regularities in adjective ordering. Some have offered grammatical proposals that attend to semantic composition or articulated syntactic hierarchies (e.g., Cinque, 1994; Scott, 2002; McNally & Boleda, 2004; Truswell, 2009). Others have advanced more psychological proposals built around notions like inherentness or accessibility (e.g, Whorf, 1945; Ziff, 1960; Martin, 1969). Recently, Scontras et al. (2017) synthesized several proposals that preceded them and advanced the hypothesis that adjective subjectivity predicts ordering preferences (see also Quirk, Greenbaum, Leech, & Svartvik, 1985; Hetzron, 1978; Dixon, 1982; Tucker, 1998; Hill, 2012).

In order to test the subjectivity hypothesis, Scontras et al. (2017) first had to determine what the ordering preferences were. They established a behavioral measure of the preferences whereby experimental participants indicated the preferred ordering of multi-adjective strings that differed only in the relative order of the adjectives involved (e.g., *the big brown bag* vs. *the brown big bag*). Scontras et al. (2017) then validated their behavioral measure by comparing it with naturalistic productions from corpora. They found a high correlation between the behavioral and corpus measures ($r^2 = .83, 95\%$ CI $[.63, .90]$), suggesting that the behavioral measure was successful in capturing the preferences speakers use when forming multi-adjective strings.

Next, Scontras et al. (2017) measured adjective subjectivity. They started by simply asking participants how "subjective" a given adjective was (e.g., "How subjective is *brown*?"). Wary of how naive participants might interpret the word "subjective," the authors validated their subjectivity measure by comparing it with faultless disagreement scores (Kölbel, 2004; Barker, 2013; Kennedy, 2013; MacFarlane, 2014). In a faultless disagreement task, participants observe a disagreement between two speakers about whether an adjective applies to some object (e.g., whether or not a table is brown). The task is to decide whether the two speakers can both be right while disagreeing, or whether one of them must be wrong; to the extent that both speakers can be right, the adjective admits that degree of faultless disagreement. Scontras et al. (2017) found an extremely high correlation between the raw "subjectivity" scores and the faultless disagreement measure ($r^2 = .91, 95\%$ CI $[.86, .94]$), suggesting that they had a reliable measure of adjective subjectivity.

Comparing the ordering preferences with adjective subjectivity, Scontras et al. (2017) found that subjectivity accounts for 85% of the variance in the ordering preferences ($r^2 = .85, 95\%$ CI $[.75, .90]$) for 26 different adjectives from seven semantic classes. The authors then looked at every multi-adjective string in the Switchboard corpus of English, finding that subjectivity accounts for 61% of the variance in ordering preferences ($r^2 = .61, 95\%$ CI $[.47, .71]$) for 74 unique adjectives from 13 semantic classes. In other words, the authors found strong support for their hypothesis that subjectivity predicts adjective ordering preferences. The question that immediately presents itself, however, is why subjectivity should matter in adjective ordering. Scontras et al. (2017) gesture toward an answer to this question—less subjective adjectives are more useful for establishing reference—but their suggestion is purely speculative.

Using a model of probabilistic utterance choice (e.g., *big brown bag* vs. *brown big bag*), Simonič (2018) systematically explored the idea that subjectivity-based ordering preferences arise under pressure from successful reference resolution. The utterance choice model was formulated within the Rational Speech Act modeling framework (e.g., Franke & Jäger, 2016; Goodman & Frank, 2016).[1] To model adjective subjectivity, the speaker assumes that the listener might have a different lexical meaning for each adjective. If $L^{S,C}_{adj}$ is the speaker's lexical entry for adjective adj in context C, the speaker believes that the listener has lexical entry $L^{L,C}_{adj}$ with probability:

$$P(L^{L,C}_{adj} \mid L^{S,C}_{adj}) \propto \begin{cases} 1 & if L^{S,C}_{adj} = L^{L,C}_{adj} \\ \varepsilon_{adj} & \text{otherwise} \end{cases} \qquad (1)$$

The more subjective the adjective, the higher the error probability ε_{adj}. With these beliefs about lexical divergence, Simonič shows that the subjectivity-based ordering *big brown bag* is a more rational choice for the speaker than *brown big bag* in a wide range of randomly-generated contexts. However, Simonič did not explicitly quantify the extent to which on ordering of adjectives is better than another, when averaging over many contexts.

Scontras et al. (to appear) pursue a similar explanation. They treat adjective subjectivity as potential noise in the semantics of an adjective, similar to Simonič, but they assume that, based on a ground-truth of objective adjective meaning, each agent (speaker or hearer) will incorrectly classify each potential referent in the current context C with an error rate ε_{adj}, which, again, indexes adjective subjectivity:

$$[\![\text{ADJ}]\!]^C = \lambda x \in C.\ \text{if } \texttt{ADJ}(x) \text{ then } \texttt{flip}(1 - \varepsilon_{adj}), \text{ else } \texttt{flip}(\varepsilon_{adj}) \qquad (2)$$

This move allows Scontras et al. to treat deviations from the ground truth as gradient: greater deviation is increasingly less likely. Scontras et al. further assume that each object classification requires some processing cost. As a result, the error probability ε_{adj} is assumed to increase with the size of context C. Based on these assumptions, Scontras et al. demonstrate how subjectivity-based ordering preferences can maximize the probability of correctly classifying the intended referent. The authors explored 103,740 cases of multi-adjective modification and found that subjectivity-based ordering behaved as expected in 93% of those cases.

In sum, both Simonič (2018) and Scontras et al. (to appear) demonstrate how subjectivity-based adjective ordering serves successful referential communication. However, both accounts involve non-trivial and potentially controversial assumptions. Simonič's definition in (1) of the speaker's beliefs about the listener's lexicon are not very intuitive: why would the speaker believe that a small deviation from his own lexicon is equally likely as a massive deviation? Scontras et al. (to appear) likewise merely stipulate that error of classification ε_{adj} in (2) is a function of context size C. It would be much more desirable to derive divergences between the speaker's and listener's semantic classifications from more fundamental assumptions, first and foremost by a more explicit view of what the underlying semantics of adjectives is. Consequently, our aim here is to build on these previous accounts by showing how subjectivity-based ordering serves successful referential communication. However, rather than making what are now rather stipulative assumptions about the misalignment of semantic representations, we will show how these misalignments can arise from a generally plausible context-dependent semantics. It is to one such semantics that we turn next.

Semantic assumptions

Schmidt, Goodman, Barner, and Tenenbaum (2009) built their study of adjective meaning on the observation that gradable adjectives mean different things depending on the nouns they modify: what counts as big for a mouse diverges drastically from what counts as big for an elephant. The question is what serves as the core meaning of a gradable adjective, such that speakers can determine its contextual extension?

To answer this question, Schmidt et al. collected human judgments about what counts as "tall" for different sets of objects. They then compared these judgments with the predictions from a number of semantic models that use various strategies to determine tallness in context. The strategies considered fell into one of two classes. The first class computed the tallness threshold directly, using various parametric and non-parametric procedures to compute a height cutoff above which objects count as tall. The second class inferred the tallness threshold on the basis of category membership, first per-

[1] See Hahn, Degen, Goodman, Jurafsky, and Futrell (2018) for a different approach to modeling adjective ordering within the Rational Speech Act framework. Their model defines speaker utility not in terms of referential success, but rather in terms of communicating subjective opinions about objects.

forming a clustering analysis on the set of objects and then identifying as tall those objects that belonged to the cluster with the tallest object.

Two models outperformed the rest. The simplest was a threshold-computing model that sets the threshold on the basis of relative height by range: any object that fell within the top $k\%$ of the range of heights in context C counts as tall in C. Formally, the set $[\![\text{tall}]\!]^C$ of objects in C that count as tall in C is (where $\texttt{tall}(o)$ is the tallness of object o, $\texttt{max}$ is the tallness of the tallest object in C, and $\texttt{min}$ that of the smallest):

$$[\![\text{tall}]\!]^C = \{o \in C \mid \texttt{tall}(o) \geq \texttt{max} - \theta \cdot (\texttt{max} - \texttt{min})\}, \quad \text{where } \theta = {}^{k}\!/_{100}. \tag{3}$$

So, if the maximum object height is 10 on the relevant scale and the minimum height is 2, a k of 50% would set the tallness threshold at 6; that is, an object with a height greater than 6 would count as tall in that context. Notably, the more complex clustering model performed no better than this threshold model when it came to predicting human judgments. We will therefore use this simple but empirically-motivated threshold semantics in the reasoning that follows, treating the threshold θ as a free model variable.

Following Simonič (2018) and Scontras et al. (to appear), we assume that iterated adjectival modification triggers *sequentially intersective context updates*. Later adjectives (syntactically farther from the modified noun) are interpreted relative to contexts that are already restricted by previous adjectives. For example, the denotation of the phrase "[adj$_\text{i}$ [adj$_\text{j}$ N]]" given a shared context C of potential referents is:

$$[\![[\text{adj}_\text{i}\ [\text{adj}_\text{j}\ N]]]\!]^C = [\![\text{adj}_i]\!]^{[\![\text{adj}_j]\!]^{C \cap [\![N]\!]}} \tag{4}$$

In words, a string like "big brown bag" characterizes the set of all bags in context C that count as brown (in the set of bags in C) and that count as big (in the set of bags that count as brown in the set of bags in C). Each adjective is therefore interpreted relative to its local context of incremental compositional semantic interpretation, so to speak. The effect is that adjectives closer to the noun will operate over a larger context (i.e., one that is less restricted); paired with a context-dependent semantics as in (3), it is conceivable that the ordering of adjectives matters for referential success.

Motivating example

For the discussion that follows, we use "brown" and "big" as mnemonic labels for any two adjectives that are, respectively, less and more subjective. Our goal is to demonstrate why an utterance of "big brown X"—that is, a multi-adjective string ordered with respect to decreasing subjectivity—is communicatively more efficient on average than an utterance of "brown big X"—an utterance not ordered with respect to decreasing subjectivity. An utterance's average communicative success is spelled out here as the *expected utility* in a situation where the speaker wants to refer to an object; this value is specified as the average probability of the listener choosing the intended referent on the basis of that utterance.

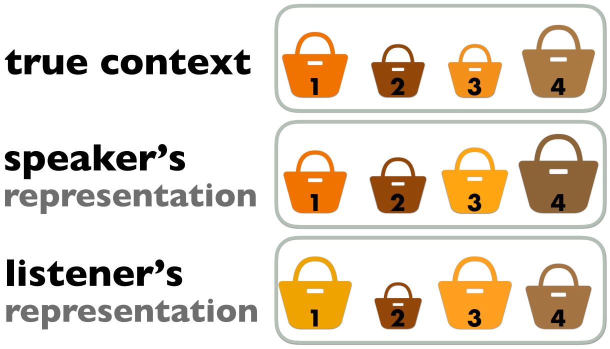

Figure 1: Illustration of subjective agent representations.

We first need to make some assumptions about the effects of adjective subjectivity on our mental representations—representations that will be relevant to referential communication. Figure 1 gives a concrete example to illustrate the main idea. Suppose that the speaker and listener share access to a context of four bags that differ only with respect to color and size. Depending on their different perceptual angles, different background knowledge, or differences in previous experiences, the speaker and listener might represent the context differently: their impressions of object size and object color could deviate from the ground truth.

Here is where subjectivity comes in: more subjective properties are more likely to lead to deviation between the ground truth (i.e., the true context) and an agent's representation of the property. Crucially, by deviating from the ground truth, these more subjective properties are also more likely to lead to deviations between two agent representations (e.g., between the speaker's and listener's representations in Figure 1); these deviations *and our awareness of their potential* are what lead to perceived subjectivity as measured by a faultless disagreement task. Language users are aware that their representations might deviate from each other's, and the potential for deviation is different for different properties. We illustrate this tendency in Figure 1, where the agent representations of size deviate more from the ground truth than their representations of color.

We now ask: if the speaker wants to describe a bag that is both big and brown according to her subjective representation of the context, would it be better, on average, to describe it as "big brown bag" or "brown big bag", if the listener would interpret either phrase from his own subjective perspective? Concretely, suppose the speaker wants to refer to bag 4 in Figure 1, which is both brown and big from her subjective point of view. If the listener hears "big brown bag", he tries to find the speaker-intended referent by incrementally restricting the set of possible referents according to the interpretation rule in (4), applying the context-dependent semantics in (3) to his own subjective representation of the objects in question. For the example from Figure 1 and assuming that $\theta = 0.5$ in (3), the phrase "brown bag" would make the listener consider only bags 2 and 4. Of these, only bag 4 is in the top 50% along the range of size in this context set. So, the interpretation of "big brown bag" is successful; the listener recovers the speaker-intended referent uniquely. In contrast, for the expression "brown big bag", the listener first looks at the bags that count as big, which rules out only bag 2, since it is the only bag whose size is in the lower 50% of the range of sizes. Among the remaining bags (1, 3 and 4), bag 3 is clearly not brown. For the sake of this informal example, assume that the listener therefore considers both bags 1 and 4 as possible referents when hearing "brown big bag". The chance of referential success (i.e., choosing bag 4)—neglecting salience or other factors—would be 1/2, lower than the certain communicative success when interpreting "big brown bag".

Computing average communicative success

We use a Monte Carlo simulation to estimate the difference in expected referential success between phrases "big brown bag" and "brown big bag"; we calculate this value by averaging over many different contexts with different numbers of objects and varying degrees of subjectivity for the properties involved. In this way, we are not assuming that agents themselves necessarily reason actively about the stochastic misalignment of semantic judgements, or that they always choose expressions that are optimal with respect to these calculations in each context. (We will come back to this issue in the final discussion.) We merely compute the average communicative success of, say, a fictitious community of agents who would use "big brown bag" (i.e., subjectivity-based ordering) and compare their success to that of a different community that uses "brown big bag" instead. A single run of the Monte Carlo simulation proceeds as follows:

1. We first sample a number n of bags in the current context uniformly at random from 4 to 20.
2. We then sample the degree to which each object is brown and the degree to which it is big. Samples are independent draws from a standard normal distribution. This yields a representation of the *actual context* C as an $n \times 2$ matrix of feature values for the n objects. The probability of sampling context C for fixed n is

$$P(C \mid n) = \prod_{i=1}^{n} \prod_{j=1}^{2} \mathcal{N}(C_{ij} \mid \mu = 0, \sigma = 1)\,.$$

3. Agent X's (speaker's or listener's) subjective representation C^X of C is derived from C by assuming normally distributed noise around the property degrees in C, with a fixed standard deviation for each adjective. The probability of obtaining a subjective representation C^X from true C is

$$P(C^X \mid C) = \prod_{i=1}^{n} \prod_{j=1}^{2} \mathcal{N}(C^X_{ij} \mid \mu = C_{ij}, \sigma = \sigma_j)\,.$$

   The standard deviations $\sigma_{1,2}$ are obtained by sampling two numbers uniformly from the interval $[0; 0.5]$ and assigning the higher number to the more subjective ("tall") and the lower to the less subjective adjective ("brown").

4. A *semantic threshold* θ is sampled uniformly at random from the unit interval. We apply the context-dependent threshold semantics in (3) from Schmidt et al. (2009) with the incrementally intersective context update in (4), using each agent's context representation, to yield each agent's subjective interpretation of each referential phrase.
5. We then sample the *speaker-intended referent object* i^* randomly from the set $[\![\mathrm{adj}_1]\!]^{C^S} \cap [\![\mathrm{adj}_2]\!]^{C^S}$ (i.e., an object that is both brown and big from the point of view of the speaker). If there is no such object, the run is discarded.
6. If the listener's interpretation of the phrase "[adj$_\mathrm{i}$ [adj$_\mathrm{j}$]]" from his subjective point of view is $I = [\![[\mathrm{adj_i}\ [\mathrm{adj_j}]]]\!]^{C^L}$, the probability of recovering the intended referent is $|I|^{-1}$ if $i^* \in I$ and 0 otherwise. We record the probability of recovery for both adjective orders and evaluate their distribution over all samples obtained in this way.

Results

Based on 10^5 Monte Carlo samples from the process outlined above, we estimate the expected probability of recovering the speaker's intended referent with the subjectivity-based ordering "big brown bag" as 0.54, compared to 0.49 for the reverse ordering "brown big bag". The obtained samples of expected utilities for each ordering appear to indeed be different (paired t-test, $t \approx 19.261$, $p < 10e^{80}$). The direction of this difference lends credence to the general idea that, on average, ordering adjectives by subjectivity does affect average referential success, and that using the less subjective adjective early in sequential interpretation is communicatively beneficial. In other words, ordering adjectives with respect to decreasing subjectivity increases the probability of communicative success.

To understand these results better, Figure 2 shows results from Monte Carlo simulations for a small selection of the parameter values we investigated. We limit our focus to values for standard deviations $\sigma_{\mathrm{brown}} \in \{0.1, 0.2\}$ and $\sigma_{\mathrm{big}} \in \{0.25, 0.3\}$ for the subjective agent representations; we consider semantic threshold values $\theta \in \{0.2, 0.4, 0.6, 0.8\}$. For each combination of these values, we ran 10,000 simulations following the procedure outlined above. The vertical axis in Figure 2 plots two measures. Upward from the 0-mark is the difference in mean communicative success between "big brown bag" and "brown big bag". We see that all mean values are positive, which signals that for all parameter constellations picked out here, the phrase "big brown bag" was indeed estimated to be communicatively more successful in each case. Below the 0-mark in Figure 2, we see the percentage of simulation runs in which the reverse ordering "brown big bag" had a higher expected utility. This shows that the communicative advantage of one adjective ordering over another is not absolute: there are exceptions. However, when averaging over all cases, there is nonetheless a clear communicative benefit of "big brown bag" over "brown big bag".

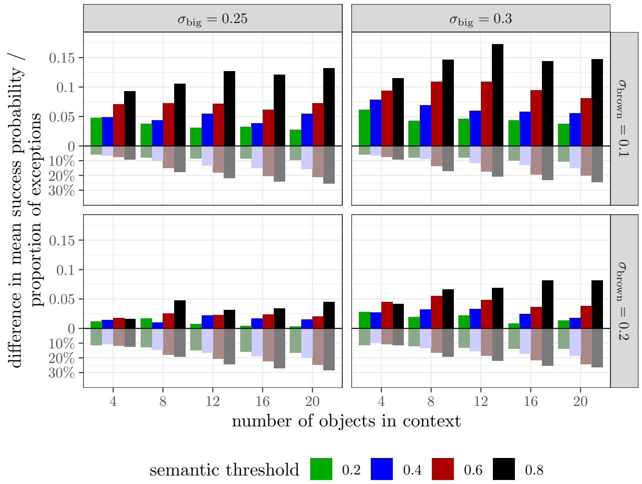

Figure 2: Results from Monte Carlo simulation with fixed values of σ_{brown}, σ_{tall} and θ. Above the 0-mark, the vertical axis shows the mean expected success of "big brown bag" minus that of "brown big bag". Below the 0-mark it shows the percentage of simulation runs where the latter ordering had a higher (however small) expected success.

Discussion

The results of our simulation suggest that a simple, empirically-motivated adjective semantics can lead to increased communicative success when multi-adjective strings are ordered with respect to decreasing subjectivity. We thus have an answer for the question of why subjectivity should matter in adjective ordering: subjectivity matters because ordering adjectives by decreasing subjectivity increases communicative success. Importantly, we arrive at this conclusion without the potentially controversial assumptions from previous work (cf. Simonič, 2018; Scontras et al., to appear). However, our model is not without its own assumptions. In what follows, we revisit the critical assumptions that led to our findings.

From a theoretical standpoint, there are three important assumptions implemented by our model. While each of these assumptions may be challenged, they serve to deliver an articulated hypothesis concerning the interpretation of multi-adjective strings—a hypothesis that offers a plausible explanation for the role of subjectivity.

First, we here operationalize the subjectivity of property *A* as the degree to which, on average, listeners and speakers will have diverging (meaning-relevant) representations of the same object's property *A*. It bears noting that the subjective property representations we assume are not (necessarily) the same as the formal linguist's notion of degree. For us, these representations serve as an abstract way of implementing divergences in truth-value assignments. As modeled here, stochastic misalignments can arise from the particulars of perception in context, but these misalignments could also arise from differing general dispositions toward classifying an object as having the property *A* when paired with random other objects.

Second, we assume that adjectival modification is, at least sometimes (see below), incrementally intersective. Moreover, we assume that meaning composition follows the hierarchical syntactic structure, rather than the linear order of the relevant string. We share this assumption with both Simonič (2018) and Scontras et al. (to appear). This assumption—that the construction of a multi-adjective nominal proceeds outward from the modified noun—ostensibly stands at odds with findings concerning the linear uptake of information in adjectival modification (e.g., Eberhard, Spivey-Knowlton, Sedivy, & Tanenhaus, 1995; Sedivy, Tanenhaus, Chambers, & Carlson, 1999). However, this assumption is common to semantic analyses of modification and necessary in many cases of multi-adjective modification (e.g., "Minnesotan wild rice" or "angry bad apple"; McNally & Boleda, 2004).

The final critical assumption we make is that adjectives have, at least sometimes (see below), a meaning that is determined at least in part by the local context that they modify. In other words, we assume that it is possible to interpret the meaning of "big" in the phrase "big brown bag" as "big for the brown bags". This assumption is the primary driver of the increased communicative success for subjectivity-based orderings: placing more subjective adjectives farther from the modified noun means that they modify a smaller context, which means that there are fewer opportunities for the listener's subjective representation to deviate from the speaker's. While some adjectives are surely less likely to have variable meanings of this sort (e.g., "cardboard", "four-legged"), the presence of any such adjectives in a multi-adjective string will lead to the pressures summarized above, which means that they will lead to pressure toward subjectivity-based orderings.

We conclude by considering the implications of our findings for our understanding of how adjective ordering preferences might develop over time. First, a note on the limitations of our findings. Our simulations, while extensive and systematic, have looked at a narrow sample of properties and scale types. We have begun to explore the predictions for other scale types (i.e., closed scales for adjectives like "full" or "safe"); however, a systematic investigation awaits future research. Still, we have demonstrated a clear communicative benefit of subjectivity-based orderings. Perhaps more importantly, we have demonstrated that this benefit does not apply universally to every possible multi-adjective string. Some parameter settings lead to exceptions where the reverse of subjectivity-based ordering yields a higher probability of communicative success.

The presence of exceptions suggests that speakers' robust, subjectivity-based adjective ordering preferences arise not out of active rational deliberation about the optimal ordering in context, but rather evolved gradually as speakers increasingly took notice of the communicative successes and failures associated with their utterances. In this way, the communicative pressures that favor subjectivity-based orderings in the majority of cases could have strengthened into the robust preferences we observe today. This sort of reasoning calls into question that nature of our knowledge of these preferences. It seems less likely that speakers represent this knowledge as a subjectivity-based heuristic that gets applied in the construction of multi-adjective strings, and more likely that the knowledge is a reflection of the statistical regularities of our linguistic experience.

Other potential explanations for subjectivity-based ordering preferences are conceivable. A prominent example is the recent explanation put forward by Hahn et al. (2018) who, unlike here, focus on non-restrictive uses of multi-adjective strings and communicative benefit related to exchanging subjective opinions about objects, which they show can be related to surface order and its impact on memory. We believe that this approach is perfectly compatible with our approach here. Both factors can play a role in supporting subjectivity-based adjective orderings. Even more usage-types of adjectival modification can and should be considered. Seen in this light, the present contribution is but a first step. It highlights that under one specific kind of use—albeit arguably the most fundamental information conveying mode of language: referential communication—a general benefit accrues for ordering adjectives by subjectivity in the way widely observed in many of the world's languages.

References

Barker, C. (2013). Negotiating Taste. *Inquiry*, *56*(2-3), 240–257. doi: 10.1080/0020174X.2013.784482

Cinque, G. (1994). On the evidence for partial N-movement in the Romance DP. In R. Kayne, G. Cinque, J. Koster, J.-Y. Pollock, L. Rizzi, & R. Zanuttini (Eds.), *Paths towards Universal Grammar. Studies in honor of Richard S. Kayne* (pp. 85–110). Washington DC: Georgetown University Press.

Dixon, R. M. W. (1982). *Where have all the adjectives gone? And other essays in semantics and syntax*. Berlin: Mouton.

Eberhard, K. M., Spivey-Knowlton, M. J., Sedivy, J. C., & Tanenhaus, M. K. (1995). Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research*, *24*(6), 409-436. doi: 10.1007/BF02143160

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.

Hahn, M., Degen, J., Goodman, N. D., Jurafsky, D., & Futrell, R. (2018). An information-theoretic explanation of adjective ordering preferences. In *Proceedings of the 40th annual conference of the Cognitive Science Society.* London: Cognitive Science Society.

Hetzron, R. (1978). On the relative order of adjectives. In H. Seiler (Ed.), *Language universals* (pp. 165–184). Tübingen: Narr.

Hill, F. (2012). Beauty before age? Applying subjectivity to automatic English adjective ordering. In *NAACL HLT 2012 Student Research Workshop* (pp. 11–16).

Kennedy, C. (2013). Two sources of subjectivity: Qualitative assessment and dimensional uncertainty. *Inquiry*, *56*(2-3), 258–277. doi: 10.1080/0020174X.2013.784483

Kölbel, M. (2004). Faultless Disagreement. *Proceedings of the Aristotelian Society*, *104*, 53–73. doi: 10.1111/j.0066-7373.2004.00081.x

MacFarlane, J. (2014). *Assessment Sensitivity*. Oxford: Clarendon Press.

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

McNally, L., & Boleda, G. (2004). Relational adjectives as properties of kinds. *Empirical Issues in Formal Syntax and Semantics*, *5*, 179–196.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language.* London: Longmans.

Schmidt, L. A., Goodman, N. D., Barner, D., & Tenenbaum, J. B. (2009). How tall is *tall*? Compositionality, statistics, and gradable adjectives. In *Proceedings of the 31st annual conference of the Cognitive Science Society.*

Scontras, G., Degen, J., & Goodman, N. D. (2017). Subjectivity predicts adjective ordering preferences. *Open Mind: Discoveries in Cognitive Science*, *1*(1), 53-65. doi: 10.1162/opmi_a_00005

Scontras, G., Degen, J., & Goodman, N. D. (to appear). On the grammatical source of adjective ordering preferences. *Semantics and Pragmatics*.

Scott, G.-J. (2002). Stacked adjectival modification and the structure of nominal phrases. In G. Cinque (Ed.), *The cartography of syntactic structures, Volume 1: Functional structure in the DP and IP* (pp. 91–120). Oxford: Oxford University Press.

Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G., & Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, *71*, 109-147. doi: 10.1016/S0010-0277(99)00025-6

Simonič, M. (2018). *Functional explanation of adjective ordering preferences using probabilistic programming* (Unpublished master's thesis). University of Tübingen.

Truswell, R. (2009). Attributive adjectives and nominal templates. *Linguistic Inquiry*, *40*, 525-533. doi: 10.1162/ling.2009.40.3.525

Tucker, G. (1998). *The lexicogrammar of adjectives: A systemic functional approach to lexis*. London: Cassell Academic.

Whorf, B. L. (1945). Grammatical Categories. *Language*, *21*(1), 1–11. doi: 10.2307/410199

Ziff, P. (1960). *Semantic Analysis*. Ithaca, NY: Cornell University Press.