Open Mind: Discoveries in Cognitive Science

Subjectivity predicts adjective ordering preferences

--Manuscript Draft--

Manuscript Number:	OPMI-D-16-00002R1
Full Title:	Subjectivity predicts adjective ordering preferences
Short Title:	Subjectivity predicts adjective ordering preferences
Corresponding Author:	Gregory Scontras Stanford University Stanford, CA UNITED STATES
Other Authors:	Judith Degen
	Noah Goodman
Abstract:	From English to Hungarian to Mokilese, speakers exhibit strong ordering preferences in multi-adjective strings: "the big blue box" sounds far more natural than "the blue big box." We show that an adjective's distance from the modified noun is predicted not by a rigid syntax, but by the adjective's meaning: less subjective adjectives occur closer to the nouns they modify. This finding provides an example of a broad linguistic universal—adjective ordering preferences—emerging from general properties of cognition.
Keywords:	Language; cognition; adjective ordering; subjectivity; semantics

Subjectivity predicts adjective ordering preferences

June 9, 2016

Abstract

From English to Hungarian to Mokilese, speakers exhibit strong ordering preferences in multi-adjective strings: "the big blue box" sounds far more natural than "the blue big box." We show that an adjective's distance from the modified noun is predicted not by a rigid syntax, but by the adjective's meaning: less subjective adjectives occur closer to the nouns they modify. This finding provides an example of a broad linguistic universal—adjective ordering preferences—emerging from general properties of cognition.

1 Introduction

Regularities in the behavior of speakers and speech communities provide a window onto the psychology of language. Here we take up one such regularity: adjective ordering. Speakers and listeners exhibit strong ordering preferences when two or more adjectives are used to modify a noun, as in "the big blue box" or "the good smooth purple plastic chair." Deviate from the preferred order, and the construction becomes odd. Something feels particularly unwieldy about "the blue big box," even more so with "the plastic good purple smooth chair." Why do most strings of adjectives have tightly-constrained order? We investigate the role of adjective meaning, specifically the subjectivity of the properties that the adjectives name, in predicting ordering preferences.

Adjective ordering preferences stand as a particularly striking case of regularity in language. More remarkable than their robustness in English is their cross-linguistic systematicity: we continually find the *same* preferences across the world's languages. Hungarian (Uralic), Telugu (Dravidian), Mandarin Chinese, and Dutch are just a handful of languages with pre-nominal adjectives (i.e., languages where adjectives precede nouns) reported to have the same ordering preferences as English (Martin, 1969b; Hetzron, 1978; Dixon, 1982; Sproat and Shih, 1991; LaPolla and Huang, 2004). In languages like Selepet (Papuan) and Mokilese (Micronesian) with post-nominal adjectives (i.e., where adjectives follow nouns), these preferences are preserved in the reverse (Hetzron, 1978; Dixon, 1982; Sproat and Shih, 1991)—stable preferences determine the linear distance of an adjective from the noun it modifies.

There have been two general approaches to the investigation of adjective ordering preferences. As part of a larger project mapping the syntax and semantics of adjectives, the

linguistics literature advances a universal hierarchy of semantic classes of adjectives. Leading the charge, Dixon (1982) set out to uncover language-internal structure by which to organize ordering preferences. The preferences were assumed to be hard-coded in the grammar; the researcher's job was simply to uncover them. Building on the ordering of semantic classes proposed by Dixon, Cinque (1994) advanced a fully syntactic account of the conventionalization of ordering preferences under which different classes of adjectives populate dedicated syntactic categories which inhabit specialized projections in the syntactic tree. For example, color adjectives project a Color Phrase, shape adjectives project a Shape Phrase. The Shape Phrase syntactically dominates the Color Phrase; with left-branching structure, hierarchical dominance results in linear precedence. The ultimate source of this rigid structure was immaterial; at issue was a comprehensive and deterministic account of the facts (see Scott, 2002, and Laenzlinger, 2005, for similar proposals).

Before the grammatical approaches, which map, as it were, the terrain of adjective *structure*, psychological approaches advanced the idea that aspects of adjectives' *meaning* explain their relative order. The trouble lies in deciding precisely which aspects of meaning are relevant. Kicking off the enterprise in 1898, Sweet proposed that adjectives which are more closely connected with the noun in meaning occur closer to the noun, and that adjectives with a more specialized meaning occur closer to the noun. Similarly, Whorf (1945) proposed that adjectives describing more "inherent" properties occur closer to the noun. Ziff (1960) proposed that adjectives with less context-dependent meaning occur closer to the noun, and that adjectives that felicitously describe a narrower set of nouns occur closer to the noun. Recent compositional approaches have argued that the fundamental factor in predicting adjective ordering is whether or not an adjective forms a new concept with the noun it modifies (McNally and Boleda, 2004; Svenonius, 2008): first you form the concepts (e.g., "wild rice" or "bad apple"), then you modify them (e.g, "Minnesotan wild rice"). Similarly, Truswell (2009) argues that the type of composition an adjective invokes (i.e., intersective vs. subsective) determines its relative order (cf. the "absoluteness" proposal from Sproat and Shih, 1991). These proposals and others like them circle around similar aspects of adjective meaning in their account of ordering preferences; unfortunately, operationalizing metrics like meaning distance, specificity, inherence, and context-dependence is not a trivial task (but see the attempt in Martin, 1969a, as well as our *SI: Comparing subjectivity with alternative accounts of adjective order*). Lacking clear empirical measures of the relevant aspects of adjective meaning, these psychological accounts gave way to the more descriptive, grammatical ones that settle for a deterministic codification of the knowledge that underlies the preferences we observe (e.g., innate syntax).

We revisit the idea that ordering preferences emerge from aspects of adjective meaning, attempting to provide more thorough empirical grounding to these notions; from the grammatical approach we adopt the strategy of using semantic classes of adjectives to structure our investigation and smooth our data. Distilling the psychological proposals that precede us into a single feature, we advance the hypothesis that it is the *subjectivity* of the property named that determines ordering preferences (Hetzron, 1978; Quirk et al., 1985; Hill, 2012). Less subjective adjectives are reliably more useful at communicating the speaker's intended message; the chance of miscommunication decreases with decreasing subjectivity. Perhaps for this reason, less subjective adjectives occur closer to the substantive head of the nominal projection, that is, to the modified noun? In "the big blue box," judgments about bigness

Table 1: Adjectives, noun, and their semantic classes.

Adjective	Class	Adjective	Class	Noun	Class
old	age	good	value	apple	food
new	age	bad	value	banana	food
rotten	age	round	shape	carrot	food
fresh	age	square	shape	cheese	food
red	color	big	dimension	tomato	food
yellow	color	small	dimension	chair	furniture
green	color	huge	dimension	couch	furniture
blue	color	tiny	dimension	fan	furniture
purple	color	short	dimension	TV	furniture
brown	color	long	dimension	desk	furniture
wooden	material	smooth	physical		
plastic	material	hard	physical		
metal	material	soft	physical		

are likely less consistent than judgments about blueness; "blue" is less subjective than "big," and so, according to this theory, it occurs closer to the noun "box."

We believe that subjectivity synthesizes—rather than supplants—the previous psychological approaches, incorporating notions like "inherentness" and "context dependence" into an intuitive psychological construct that readily operationalizes as a behavioral measure. To test the hypothesis that adjective subjectivity predicts ordering preferences, we created and validated empirical measures of the ordering preferences themselves and of an adjective's subjectivity. With reliable estimates of both, we then evaluated the predictive power of subjectivity in adjective ordering preferences.

2 Experiment 1: Establishing the measures

2.1 Ordering preferences

We began by measuring preferences in adjective ordering. We selected a sample of 26 relatively frequent, imageable adjectives from seven different semantic classes (dimension, value, age, physical, shape, color, material). We then elicited naturalness judgments on adjective-adjective-noun object descriptions.

Participants. We recruited 50 participants through Amazon.com's Mechanical Turk crowdsourcing service. Participants were compensated for their participation.

Design and methods. Participants were asked to indicate which of two descriptions of an object sounded more natural. Each description featured a noun modified by two adjectives, for example "the red small chair" or "the small red chair". Descriptions were random combinations of two adjectives and a noun from the list in Table 1, with the constraint that

no description contained adjectives from the same semantic class. Description pairs contained the same words, with relative adjective order reversed. On each trial, participants indicated their choice by adjusting a slider with endpoints labeled with the competing descriptions; an example trial appears in Fig. 1. Participants completed 26 trials. On each trial, we measured the distance of the slider from each endpoint; values ranged between 0 and 1. Only native speakers of English were included in the analyses; we analyzed data from 45 participants.

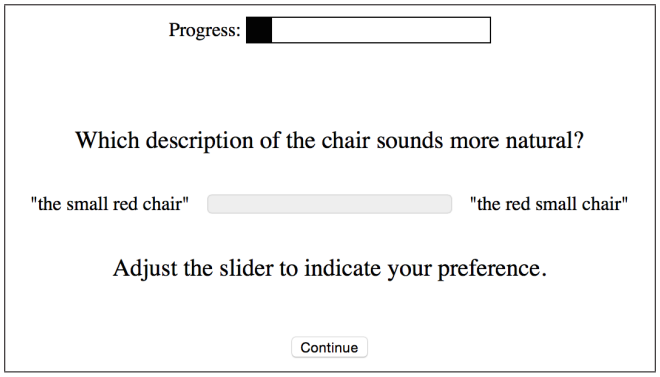


Figure 1: Example trial from *Expt. 1.1 Ordering preferences*; participants indicated the more natural of two adjective-adjective-noun descriptions on a sliding scale.

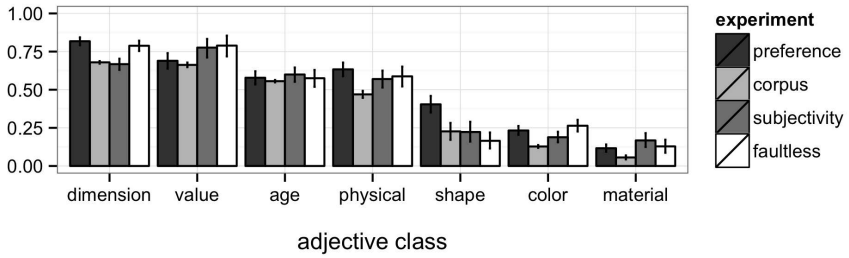


Figure 2: Mean distance from noun inferred from naturalness ratings (*preference*), mean distance from noun calculated from corpus counts (*corpus*), mean subjectivity ratings (*subjectivity*), and mean faultless disagreement ratings (*faultless*) for adjectives grouped by their semantic class. Error bars represent bootstrapped 95% confidence intervals (DiCiccio and Efron, 1996).

Results. For each adjective, we computed its mean naturalness score by averaging ratings of configurations in which it appeared in first position, farthest from the noun. Fig. 2 (*naturalness*) plots these mean naturalness scores by adjective class; greater values signal that a class's adjectives are preferred in first position, farther from the noun. This preferred distance measure closely tracks class-level ordering hierarchies reported in the literature (Dixon, 1982; Sproat and Shih, 1991).

Corpus validation. To validate our behavioral measure of ordering preferences, we conducted a corpus study on the same 26 adjectives and measured their mean distance from

the noun in phrases with two adjectives. We used TGrep2 (Rohde, 2005) and the TGrep2 Database Tools (Degen and Jaeger, 2011) to extract all "A A N" NPs that contained one of the 26 adjectives in Table 1 from the Penn Treebank subset of the Switchboard corpus of telephone dialogues (Godfrey et al., 1992), as well as from the spoken and the written portions of the British National Corpus (BNC, see http://www.natcorp.ox.ac.uk/). For these cases, we computed the distance of each occurrence of our 26 target adjectives from the modified noun, yielding results for a total of 38,418 adjective tokens. For each adjective, mean distance from the noun was computed (where the position directly preceding the noun was coded as 0, and the position preceding that was coded as 1).

Mean distance from the noun for each adjective class is shown in Fig. 2 (*corpus*). The corpus measure closely tracks the qualitative pattern we measured in our naturalness experiment; quantitatively, the two measures are highly correlated ($r^2 = 0.83$, 95% CI [0.63, 0.90]), in spite of the fact that the corpus measure includes cases from a superset of the nouns tested in our naturalness experiment. Our naturalness ratings thus operationalize both immediate ordering preferences and speakers' preferences in natural usage.

Looking for noun effects. We next asked to what extent ordering preferences depend on the modified noun. Analyses of adjective semantics hold that adjectives that form idiomatic concepts or natural kinds (e.g., "bad apple" or "blue cheese") compose first with nouns, before run-of-the-mill intersective adjectives, thus enforcing a strict ordering (McNally and Boleda, 2004; Svenonius, 2008); we might therefore expect to find that adjectives that compose with a noun to yield an idiomatic concept are preferred linearly closer to the modified noun. To evaluate the role of specific noun information in determining ordering preferences, we performed a nested linear model comparison. The models predicted naturalness ratings either by ADJECTIVE (i.e., the adjective farthest from the noun) only, or by ADJECTIVE together with its interaction with NOUN (i.e., the modified noun). The model comparison revealed that noun-specific ratings did not explain any additional variance in ordering preference above and beyond adjective-level ratings ($F(234, 2080) = 1.10, p < 0.15$). Thus, we fail to find evidence of noun-specific effects on ordering preferences in our materials.

2.2 Subjectivity

With clear estimates of ordering preferences, we then measured the subjectivity of the adjectives that were tested in the ordering preferences experiment. We started with a direct measure of "subjectivity."

Participants. We recruited 30 participants through Amazon.com's Mechanical Turk crowdsourcing service. Participants were compensated for their participation.

Design and methods. Participants were shown a series of adjectives and asked to indicate how "subjective" each one was on a sliding scale with endpoints labeled as "completely objective" (coded as 0) and "completely subjective" (coded as 1; Fig. 3). Participants completed a total of 26 trials, one for each adjective in Table 1. The order was randomized

for each participant. Only native English speakers were included in the analyses; we analyzed data from 28 participants.

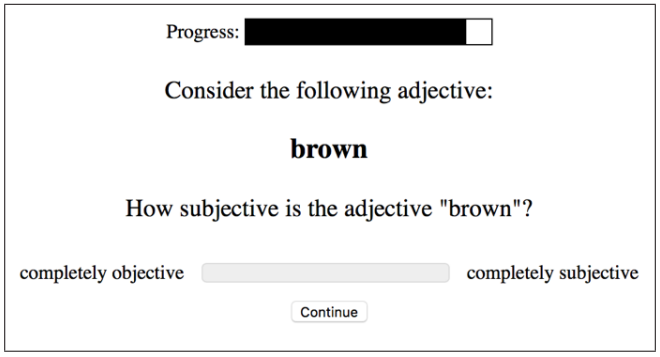


Figure 3: Example trial from *Expt. 1.2 Subjectivity*; participants rated the subjectivity of adjectives.

Results. We averaged the subjectivity scores for each adjective; greater values indicate greater subjectivity. These averages were used in the analyses reported below. Fig. 2 (*subjectivity*) show these scores by adjective class.

Faultless disagreement validation. Because subjectivity may be an ambiguous, or even subjective, property, we explored a second measure that may have greater ecological validity. We operationalized subjectivity as the potential for faultless disagreement between two speakers, which captures potential uncertainty about assessment criteria and assessment outcomes (Kölbel, 2004; Kennedy, 2013; Barker, 2013).[1] We had participants (n=40) evaluate whether two speakers could both be right while the speakers produced conflicting object descriptions. For example, an experimental trial would have Mary assert, "That apple is old," then have Bob counter with "That apple is not old;" participants rated whether both Mary and Bob could be right, or whether one of them must be wrong. This measure, the faultless disagreement potential for the adjective at issue, serves as an empirical estimate of adjective subjectivity. Fig. 2 (*faultless*) plots these scores by adjective class, where a value of 1 signals that a class's adjectives are always amenable to faultless disagreement (i.e., maximally subjective). The results of this method were highly correlated with our direct "subjectivity" scores ($r^2 = 0.91$, 95% CI [0.86, 0.94]), suggesting that they measure a common underlying value: adjective subjectivity.

2.3 Predicting adjective order

To evaluate the power of subjectivity in predicting adjective ordering preferences, Fig. 4 plots mean naturalness ratings (Expt. 1.1) against mean adjective subjectivity scores (Expt. 1.2). Adjective subjectivity scores account for 85% of the variance in the naturalness ratings ($r^2 =$

[1] See MacFarlane (2014) for more discussion of the many factors, both "semantic" and "pragmatic," that contribute to faultless disagreement effects. For a different approach, see Hill (2012), who builds on previous corpus work (Wulff, 2003) to infer adjective subjectivity from surface features of strings.

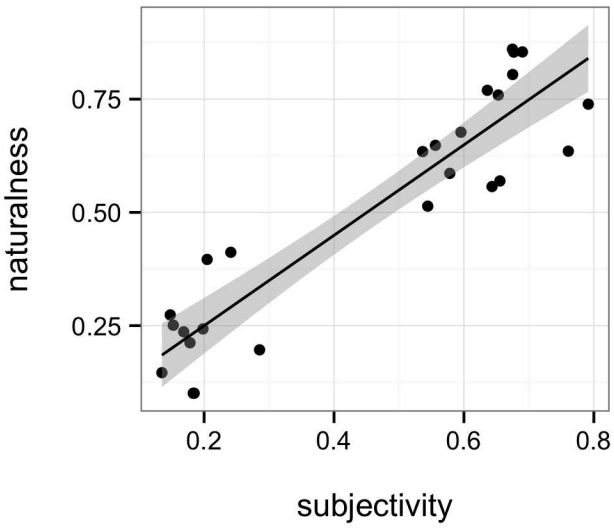


Figure 4: Mean naturalness ratings plotted against mean subjectivity scores for each of the 26 adjectives tested in Expt. 1.

0.85, 95% CI [0.75, 0.90]). The faultless disagreement scores also perform well, accounting for 88% of the variance ($r^2 = 0.88$, 95% CI [0.77, 0.95]). Using either measure, more subjective adjectives are preferred farther from the noun; subjectivity indeed predicts adjective ordering preferences.

One might worry that conducting our analysis at the level of individual adjectives obscures information about the specific adjective-adjective configurations that participants rated in our naturalness experiment. We therefore computed a subjectivity difference score for each adjective class configuration (i.e., an ordered pairing of two adjective classes, CLASS1-CLASS2) by subtracting the mean subjectivity score for CLASS2 from the mean subjectivity score for CLASS1. Higher difference scores indicate that the adjective class closer to the noun is less subjective than the class farther away. Fig. 5 plots mean naturalness ratings for adjective class configurations against these subjectivity difference scores; the two measures are highly correlated ($r^2 = 0.80$, 95% CI [0.68, 0.88]). We also see that as the difference in subjectivity approaches zero, the naturalness ratings approach 0.5 (i.e., chance): ordering preferences weaken for adjectives of similar subjectivity (e.g., "yellow square" or "fresh soft").

2.4 Discussion

We found that adjective subjectivity scores account for almost all of the variance in naturalness ratings, for several different analyses, strongly supporting our hypothesis that less subjective adjectives occur closer to the noun. One might worry that the observed success of subjectivity in predicting ordering preferences is an artifact of the set of 26 adjectives we tested, and might not generalize to a broader set of adjectives. Therefore, we next consider a much larger set of adjectives.

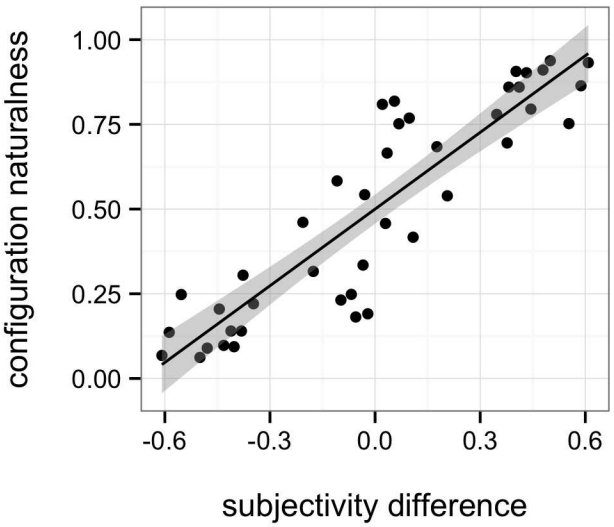


Figure 5: Mean configuration naturalness ratings plotted against subjectivity difference scores for each pair of adjective classes tested in Expt. 1.

3 Experiment 2: Generalizing our findings

To test the generalizability of the findings from Expt. 1, we aimed to construct a set of adjectives that are attested in multi-adjective constructions and that span both semantic classes and a broad spectrum of frequencies and lengths. The set of 78 adjectives we ultimately used includes many adjectives that are traditionally overlooked in investigations of ordering preferences.

3.1 Ordering preferences

Participants. We recruited 495 participants through Amazon.com's Mechanical Turk. Participants were compensated for their participation.

Materials To arrive at a new set of adjectives, we began by extracting every unique adjective appearing in an "A A N" NP from the Penn Treebank subset of the Switchboard corpus; there were 350 unique adjectives. Then, two of the authors (GS and JD) independently coded these 350 adjectives according to semantic classes from the literature (e.g., Dixon, 1982; Sproat and Shih, 1991), indicated in Table 2. We added a miscellaneous category "X" containing non-intersective adjectives (cf. Cinque, 2014). After ignoring adjectives for which there was no obvious semantic class, as well as adjectives on which our classifications did not agree, we were left with 196 unique adjectives from 13 different classes.

To further refine this list of 196, we calculated adjective frequency, and, on the basis of frequency, divided the adjectives within each class into four quartiles. We also calculated adjective length (i.e., number of characters) and classified adjectives as either short or long. Then, for each adjective class with more than eight members, for each frequency quartile, we sampled a random adjective from each length (e.g., one short extremely frequent dimension

Table 2: Adjectives used in Expt. 2.

Adjective	Class	Adjective	Class	Adjective	Class
junior	age	professional	human	sweet	physical
new	age	sad	human	circular	shape
old	age	selfish	human	square	shape
old-time	age	strict	human	fast	speed
senior	age	closest	location	slow	speed
young	age	internal	location	speedy	speed
black	color	overhead	location	current	temporal
blonde	color	corduroy	material	daily	temporal
blue	color	crocheted	material	everyday	temporal
green	color	gold	material	historical	temporal
purple	color	wooden	material	best	value
red	color	brazilian	nationality	exciting	value
white	color	english	nationality	favorite	value
yellow	color	european	nationality	lavish	value
biggest	dimension	hispanic	nationality	plain	value
large	dimension	international	nationality	pleasant	value
long	dimension	japanese	nationality	prestigious	value
mini	dimension	national	nationality	strange	value
narrow	dimension	vietnamese	nationality	designated	X
open	dimension	creamy	physical	different	X
thick	dimension	curly	physical	individual	X
thin	dimension	frozen	physical	last	X
civilized	human	lacy	physical	mixed	X
creative	human	smooth	physical	potential	X
entrepreneurial	human	solid	physical	token	X
playful	human	spicy	physical	unique	X

adjective and one long extremely frequent dimension adjective). There were three cases without any observations from which to sample: "nationality" (Q3 short and Q4 short) and "physical" (Q4 short). We filled in these cases by randomly sampling from the leftovers in those two classes. Finally, we added an additional shape adjective, "square", because previously there was only "circular". This process yielded the 78 adjectives in Table 2.

Nouns were chosen in a similar fashion. From the set of "A A N" NPs in Switchboard, we extracted the 295 unique nouns, then restricted this set by 1) removing all plural or clearly collective nouns, and 2) removing nonsense words or words that might be confused with participles. This process yielded 166 unique nouns.

Design and methods. The design was identical to our previous naturalness rating experiments (Expts. 1.1 and 2.1): participants indicated which of two object descriptions sounded more natural, choosing between adjective-adjective-noun permutations that varied the rela-

tive order of the adjectives. Adjectives were chosen at random from the set in Table 2, with the constraint that adjectives from the same class were not paired together. Participants completed 30 trials. On each trial, participants indicated their choice by adjusting the slider between endpoints labeled with the competing descriptions. Additionally, participants were able to indicate if a particular description did not make sense by checking a box labeled "Neither option makes sense." Only native speakers of English were included in the analyses; we analyzed data from 473 participants.

Results. For each adjective, we computed its mean naturalness score by averaging ratings of configurations in which it appeared in first position, farthest from the noun. Participants demonstrated little preference for adjective order when the descriptions were nonsense. For this reason, we excluded responses to nonsensical descriptions from the analyses of subjectivity below; this exclusion process removed 2295 observations (16% of the total 14,190).

As with the previous ordering preference experiment, we performed a nested model comparison to investigate the role of specific noun information in determining ordering preferences. The models predicted naturalness ratings either by ADJECTIVE only, or by ADJECTIVE together with its interaction with NOUN. Once again, the model comparison revealed that noun-specific ratings did not explain any additional variance in ordering preference beyond adjective-level ratings ($F(9676, 11593) = 1.02, p < 0.14$). However, participants were not ignoring the nouns altogether. To see whether nouns had an effect on sense rates, we performed another nested linear model comparison. This time, the models we compared predicted the sense of an object description either with the two ADJECTIVES only, or with the ADJECTIVES and the NOUN. The model comparison revealed that noun information does explain additional variance in sense rates beyond the two adjectives involved ($F(165, 11267) = 2.03, p < 0.001$), suggesting that participants were paying attention to the nouns in the object descriptions they encountered.

3.2 Subjectivity

Next, we evaluated the subjectivity of our new set of adjectives using the direct "subjectivity" task from Expt. 1.2.

Participants. We recruited 198 participants from Amazon.com's Mechanical Turk. Participants were compensated for their participation.

Design and methods. The design was identical to our previous direct "subjectivity" experiment. Participants completed a total of 30 trials. On each trial an adjective was chosen at random from the set of 78 in Table 2. Only native speakers of English were included in the analyses; we analyzed data from 189 participants.

Results. We averaged the subjectivity scores for each adjective; greater values indicate greater subjectivity. To evaluate the power of subjectivity in predicting adjective ordering preferences, we compared subjectivity scores with the naturalness ratings (Fig. 6). Adjective subjectivity scores account for 51% of the variance in the naturalness ratings ($r^2 = 0.51$,

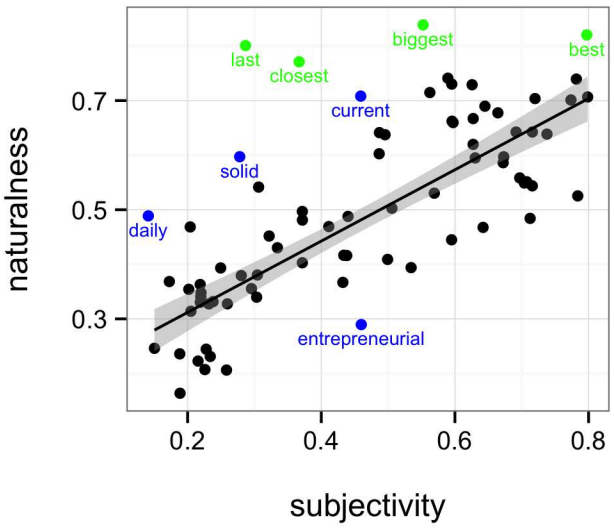


Figure 6: Mean naturalness ratings plotted against mean subjectivity scores for each of the 78 adjectives tested in Expt. 2. Superlatives are labeled in green; outlier adjectives are labeled in blue.

95% CI $[0.32, 0.66]$). Four observations clearly stood out in Fig. 6, corresponding to the superlatives *best*, *biggest*, *closest*, and *last*. Indeed, superlatives have been observed to eschew adjective ordering preferences, occurring farthest from the modified noun regardless of class or subjectivity (Dixon, 1982); our naturalness ratings reflect this fact. Removing superlatives, we find that subjectivity scores perform markedly better, accounting for 61% of the variance ($r^2 = 0.61$, 95% CI $[0.47, 0.71]$). At the level of adjective class configurations, subjectivity difference scores account for 74% of the variance in the configuration ratings ($r^2 = 0.74$, 95% CI $[0.66, 0.79]$; Fig. 7).[2]

A post-hoc look at our data revealed a small number of outlier adjectives (in addition to the four superlatives). To systematically detect these outlier adjectives, we fit a linear regression predicting naturalness ratings by subjectivity scores, then calculated the absolute difference between the actual naturalness ratings and the model's predicted values. Setting the cutoff for this difference score at 3 $\times$ standard deviation, four adjectives stood apart as outliers: *entrepreneurial*, *solid*, *current*, and *daily* (labelled in blue in Fig. 6). Without the four outlier adjectives (and the four superlatives), adjective subjectivity scores account for 70% of the variance in the naturalness ratings ($r^2 = 0.70$, 95% CI $[0.58, 0.78]$).

We also looked at the contribution of frequency and length in predicting ordering preferences. Treating subjectivity, frequency, and length as predictors in a linear regression predicting naturalness ratings (excluding superlatives), the model accounts for 70% of the variance ($r^2 = 0.70$). Nested model comparison reveals that the subjectivity predictor explains significant variance in the extended model ($F(1, 70) = 141.38, p < 0.001$); the frequency and

[2]This analysis and the plot in Fig. 7 exclude superlatives. If we include superlatives in the class configuration analysis, subjectivity difference scores account for 69% of the variance in the naturalness ratings ($r^2 = 0.69$, 95% CI $[0.60, 0.76]$).

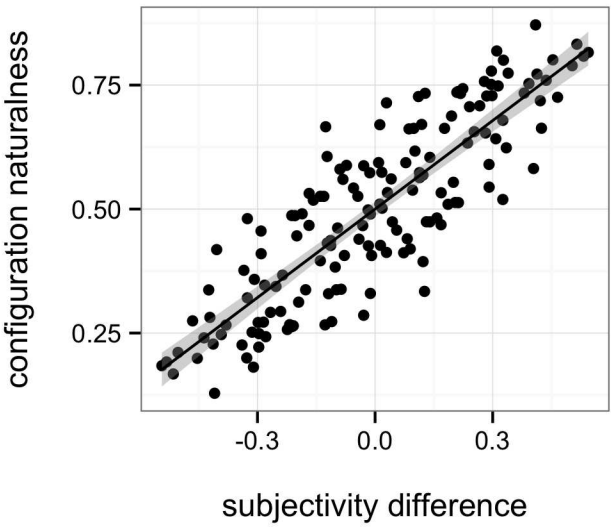


Figure 7: Mean configuration naturalness ratings plotted against subjectivity difference scores for each pair of adjective classes tested in Expt. 2.

length predictors also explain significant variance (*frequency*: $F(1, 70) = 7.71, p < 0.01$; *length*: $F(1, 70) = 9.73, p < 0.01$). If we remove outlier adjectives that fall more than three standard deviations away from the predicted value of the extended model (there were six: *mini*, *frozen*, *solid*, *current*, *daily*, *designated*), the model performs better, accounting for 76% of the variance ($r^2 = 0.76$).

3.3 Discussion

The results of the current experiment demonstrate that subjectivity predicts ordering preferences in a much larger set of materials drawn from naturally-occurring examples. At worst, subjectivity accounts for more than half of the variance in the naturalness ratings for our set of 78 adjectives. Once we exclude superlatives, whose semantics likely dictates their position in strings of nominal modifiers, as well as four outlier adjectives, subjectivity accounts for 70% of the variance in this set of 70 adjectives. While adjective frequency and length contribute to the observed preferences, we saw that subjectivity alone accounts for the vast majority of the variance in our data.

There remains the question of precisely *why* the four outlier adjectives—*entrepreneurial*, *solid*, *current*, and *daily*—performed so poorly with respect to the predictions of subjectivity. Perhaps the most notable feature of this set of adjectives is its heterogeneity: we fail to find clear groupings by semantic class, relative frequency, or length. However, length likely does factor into the observed behavior of *entrepreneurial*, the longest adjective tested, which was the only outlier *under*-predicted by its subjectivity: participants preferred *entrepreneurial* closer to the noun than its subjectivity alone would predict. Indeed, relative length has long been known to affect the order of constituents, even in the domain of adjective ordering (Wulff, 2003): longer constituents appear later. Once we factor length into the equation

predicting ordering preferences, *entrepreneurial* no longer stands out.

4 General discussion

Adjective ordering preferences have received considerable attention throughout the history of generative grammar and cognitive psychology, owing to their remarkable stability within and across languages. Something so robust, the reasoning goes, must evidence a deep principle of the cognitive architecture that shapes language. Yet while descriptions of the phenomenon abound, an explanation has proven elusive. Grammatical theories that posit a rigid syntax of adjective classes offer little more than a codification of the facts, and psychological approaches stumble when it comes to operationalizing the specific aspects of adjective meaning at play.

In our investigation, we established two empirical constructs: the preferences themselves, which we measured using naturalness ratings and validated with corpus statistics; and adjective subjectivity, which we measured directly and corroborated with potential for faultless disagreement. An adjective's semantics predicts its distance from the modified noun, such that less subjective adjectives occur linearly closer to nouns they modify. In our *SI: Comparing subjectivity with alternative accounts of adjective order*, we investigated the predictions of three other hypotheses from the literature: adjective inherentness (i.e., how essential an adjective's meaning is to the noun it modifies; Sweet, 1898; Whorf, 1945), intersective vs. subsective modification (i.e., the mode by which an adjective composes semantically with the noun it modifies; Truswell, 2009), and concept formability (i.e., whether an adjective composes with a noun to form a complex, idiomatic concept; McNally and Boleda, 2004; Bouchard, 2005; Svenonius, 2008). In each case, we found that subjectivity has greater predictive power.

It bears noting that the preference to place less subjective adjectives closer to nouns is not deterministic; non-preferred orderings of adjectives can serve a communicative purpose, for example to establish contrastiveness in discourse (Martin, 1969a, 1970; Hill, 1958; Vendler, 1963). This constrastiveness follows straightforwardly from a manner implicature (Levinson, 2000): marked forms (i.e., non-preferred orderings of adjectives) yield marked interpretations (i.e., atypical modification constituency). The work lies in determining the preferred orderings from which contrastive uses depart. Indeed, many other situational factors are likely to influence ordering (e.g., phonological shape, noun semantics, word and bigram frequencies; cf. Wulff, 2003, and the results of Expt. 2); it is the more general tendencies we are concerned with here.

Adjectives are just one of many elements that may occur in complex nominal constructions. Other classes of elements include demonstratives (e.g., *this* and *that*) and numerals. In his Universal 20, Greenberg observes that the relative order of these higher-order classes is also stable cross-linguistically (Greenberg, 1963; Culbertson and Adger, 2014), suggesting that subjectivity interacts with additional constraints from semantic composition in the determination of word order. Indeed, we saw hints of such interactions in Expt. 2, where superlatives stood apart from run-of-the-mill adjectives. Beyond nominals, adverbs (e.g., *honestly*, *probably*, *carefully*) are reported to exhibit regular orderings cross-linguistically (Cinque, 1999; Ernst, 2002). Understanding these orderings would likely benefit from a systematic empirical treatment similar to the one we have advanced here.

While subjectivity accounts for the regularities we observe in adjective ordering, the deeper explanation for how subjectivity determines the relative order of adjectives remains unsettled. Our results suggest that ordering preferences likely emerge, at least partially, from a desire to place less subjective content closer to the substantive head of a nominal construction (i.e., closer to the modified noun). For now we can only speculate about the ultimate source of this desire. Subjective content allows for miscommunication to arise if speakers and listeners arrive at different judgments about a property description. Hence, less subjective content is more useful at communicating about the world. An explanation along these lines, based on pressures to facilitate successful reference resolution, would have to depend on the hierarchical, not linear, ordering of adjectives: noun phrases are built semantically outward from the noun, and more useful, less subjective content enters earlier in this process (cf. the mirroring of preferences in pre- vs. post-nominal languages). A full explanation must examine not only *why* we observe the preferences that we do, but also *how* and to what extent these preferences get conventionalized via the diachronic processes that shape language—a promising direction for future research.

Whatever its source, the success of subjectivity in predicting adjective ordering preferences provides a compelling case where linguistic universals, the regularities we observe in adjective ordering, emerge from cognitive universals, the subjectivity of the properties that the adjectives name.

References

Barker, C. (2013). Negotiating Taste. *Inquiry 56*(2-3), 240–257.

Bouchard, D. (2005). Sériation des adjectifs dans le SN et formation de concepts. *Recherches linguistiques de Vicennes 34*, 125–142.

Cinque, G. (1994). On the Evidence for Partial N-Movement in the Romance DP. In R. S. Kayne, G. Cinque, J. Koster, J.-Y. Pollock, L. Rizzi, and R. Zanuttini (Eds.), *Paths Towards Universal Grammar. Studies in Honor of Richard S. Kayne*, pp. 85–110. Washington DC: Georgetown University Press.

Cinque, G. (1999). *Adverbs and Functional Heads*. Oxford: Oxford University Press.

Cinque, G. (2014). The semantic classification of adjectives: A view from syntax. *Studies in Chinese Linguistics 35*(1), 1–30.

Culbertson, J. and D. Adger (2014). Language learners privilege structured meaning over surface frequency. *Proceedings of the National Academy of Sciences of the United States of America 111*(16), 5842–5847.

Degen, J. and T. F. Jaeger (2011). The TGrep2 Database Tools.

DiCiccio, T. J. and B. Efron (1996). Bootstrap confidence intervals. *Statistical Science 11*, 189–228.

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

Ernst, T. (2002). *The Syntax of Adjuncts*. Cambridge: Cambridge University Press.

Godfrey, J. J., E. C. Holliman, and J. McDaniel (1992). SWITCHBOARD: A Telephone Speech Corpus for Research and Development. In *Proceedings of ICASSP-92*, pp. 517–520.

Greenberg, J. H. (1963). Some universals of grammar with particular reference to the order of meaningful elements. In *Universals of Human Language*, pp. 73–113. Cambridge, MA: MIT Press.

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

Hill, A. A. (1958). *Introduction to linguistic structures*. New York: Harcourt, Brace.

Hill, F. (2012). Beauty Before Age? Applying Subjectivity to Automatic English Adjective Ordering. In *Proceedings of the 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.

Kölbel, M. (2004). Faultless Disagreement. *Proceedings of the Aristotelian Society 104*, 53–73.

Laenzlinger, C. (2005). French adjective ordering: perspectives on DP-internal movement types. *Lingua 115*, 645–689.

LaPolla, R. J. and C. Huang (2004). Adjectives in Qiang. In R. M. W. Dixon and A. Y. Aikenvald (Eds.), *Adjective Classes: A Cross-Linguistic Typology*, pp. 306–322. Oxford: Oxford University Press.

Levinson, S. C. (2000). *Presumptive Meanings*. Cambridge, MA: MIT Press.

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

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

Martin, J. E. (1969b). Some Competence-Process Relationships in Noun Phrases with Prenominal and Postnominal Adjectives. *Journal of Verbal Learning and Verbal Behavior 8*, 471–480.

Martin, J. E. (1970). Adjective Order and Juncture. *Journal of Verbal Learning and Verbal Behavior 9*, 379–383.

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

Quirk, R., S. Greenbaum, G. Leech, and J. Svartvik (1985). A Comprehensive Grammar of the English Language.

Rohde, D. L. T. (2005). TGrep2 User Manual, Version 1.15.

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.

Sproat, R. and C. Shih (1991). The cross-linguistic distribution of adjective ordering restrictions. In C. Georgopoulos and R. Ishihara (Eds.), *Interdisciplinary approaches to language: Essays in honor of S.-Y. Kuroda*, pp. 565–593. Dordrecht: Kluwer Academic Publishers.

Svenonius, P. (2008). The position of adjectives and other phrasal modifiers in the decomposition of DP. In L. McNally and C. Kennedy (Eds.), *Adjectives and Adverbs: Syntax, Semantics, and Discourse*, pp. 16–42. Oxford: Oxford University Press.

Sweet, H. (1898). *A New English Grammar, Logical and Historical*. London: Clarendon Press.

Truswell, R. (2009). Attributive adjectives and nominal templates. *Linguistic Inquiry 40*, 525–533.

Vendler, Z. (1963). The transformational grammar of English adjectives.

Whorf, B. L. (1945). Grammatical Categories. *Language 21*(1), 1–11.

Wulff, S. (2003). A multifactorial corpus analysis of adjective order in English. *International Journal of Corpus Linguistics 8*(2), 245–282.

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

Gregory Scontras
Judith Degen
Noah D. Goodman
Department of Psychology
Stanford University
Stanford, CA 94305

June 9, 2016

Dear Editor,

We would like to thank you and the three anonymous reviewers for your helpful comments on our paper, "Subjectivity predicts adjective ordering preferences." As you will recall, you highlighted the following three concerns:

1. *The discussion/framing of other approaches needs some attention: R1 mentions some other hypotheses that deserve discussion, and R2 takes issue with the framing of previous approaches as syntactic vs semantic. This latter point is related to another issue mentioned by R2 and R3: the distinction between online decisions about adjective ordering versus conventionalized ones. Clarifying how your proposal relates to these questions would, I think, help address these concerns.*

   We have included reference to and discussion of the other hypotheses that the reviewers mention. We have also taken care in our framing to distinguish the *conventionalization* of ordering preferences from the *explanation* of the observed orderings; it is the latter that we investigate in our paper. Finally, we have also revised our language to emphasize that the subjectivity hypothesis is meant to synthesize—not supplant—previous approaches to adjective ordering.

2. *All three reviewers express some dissatisfaction about the lack of direct comparison between your hypothesis and others. As they point out, the fact that your data are well explained by subjectivity tells us little about whether this hypothesis provides a better explanation than any other. Some of the other hypotheses may seem difficult to operationalize, but on the face of it, so is subjectivity. A quantitative comparison to other approaches would considerably strengthen the paper.*

   We agree that a quantitative comparison will strengthen our claims regarding subjectivity by demonstrating not only the absolute, but also the *relative* success of our hypothesis. For this reason, we ran four additional experiments and performed a host of analyses to operationalize three alternative accounts of adjective order and compare their predictions with those made by subjectivity. To satisfy journal length length requirements, we followed your suggestion and included this information in a supplement (*Supporting information: Comparing subjectivity with alternative accounts of adjective order*), summarizing the results in the paper's General Discussion. The new results demonstrate the robust replicability of the success of subjectivity in predicting adjective order. They also demonstrate that the alternative accounts we considered add

little to our understanding of ordering preferences: subjectivity is a better predictor. This exploration has led us to appreciate the intuitive appeal of the subjectivity hypothesis, which targets an easily-accessed, stable psychological construct: adjective subjectivity.

3. *R1 and R2 both express confusion/skepticism about the explanation for *why* subjectivity should determine ordering. I suspect that some readers will be skeptical regardless of the explanation, but it would be good to clarify if possible – or alternatively to admit that the hypothesis is based purely on descriptive adequacy and indeed it is not clear why it holds. (If so, this might suggest that subjectivity is correlated with some other, perhaps as yet undetermined, factor that could provide a more convincing mechanism for ordering.)*

   Reviewers 1 and 2 were uncomfortable with the paper's concluding paragraph, presumably because they took our speculation as a claim about the psychological mechanism. In our revision, we have modified our language to emphasize that these musings are indeed just speculation and suggestions for future research, informed by our experimental results. While we do not claim to have a theoretical explanation for why subjectivity should predict ordering preferences, we do not believe this decreases the significance of the finding. (An extraordinarily high correlation between two *a priori* unrelated but theoretically interesting behavioral measures is itself an important finding.)

In the remainder of this letter, we consider in more detail each of the reviewers' concerns.

Reviewer 1:

1. *I really appreciate the historical framing of the literature, but I do miss some very influential semantic or 'psychological' contemporary hypotheses: The hypothesis that not subjectivity, but subsectivity, determines adjective order (Trueswell 2009); the apparentness hypothesis (Sproat & Shih, 1991); even Laenzlinger (2005). Ignoring these literatures make it seem like the authors are uncomfortable with the idea that others had the same intuition they had, which they shouldn't be. It does, however, relativize the present findings: Subjectivity is one explanation for a whole range of factors that determine adjective order; and as it stands, I am not convinced that subjectivity explains the data any better than the subsective/intersective distinction, or apparentness. All these concepts overlap significantly, and if I am to buy the explanation of subjectivity, I want at the very least some additional data that allow a comparison.*

   We have added the suggested references, and clarified our language to stress that subjectivity synthesizes rather than supplants the work that precedes ours (cf. our response to the Editor's point 1). To address the issue of comparing the predictions of subjectivity with those made by previous hypotheses, we have added the *Supporting information: Comparing subjectivity with alternative accounts of adjective order* (cf. our response to the Editor's point 2).

2. *I am a little bit confused about Experiment 2. If I understand correctly (and I might not!), this study is the same as Experiment 1, but with a wider range of adjectives. Also, the model explains the data much, much less well (88% in Experiment 1 vs. 61% in Experiment 2). Thus, it seems like subjectivity is no better than most models that exist; and certainly worse than Wulff (2003).*

   Expt. 2 uses the same design as Expt. 1 with a much larger set of materials: 78 adjectives and 166 nouns. Many of these adjectives and most of the nouns have been ignored in previous accounts of adjective order. That subjectivity accounts for less variance in Expt. 2 is a direct result of the fact that there is *more* variance in the data, owing to the massive increase in materials. Despite the decrease in predictive power from Expt. 1 to Expt. 2, subjectivity continues to perform remarkably well. As for other models (which we operationalized to the best of our abilities in order to derive comparable predictions), our new SI serves to demonstrate that subjectivity is *much* better at predicting ordering preferences (see Supplement).

   It is impossible to compare the correlations we report directly with the performance of Wulff's LDA model. Wulff uses eight factors to predict a binary outcome: the relative position of two adjectives in a corpus. We used a single factor, subjectivity, to predict a gradient outcome: the strength of the preference for placing that adjective first in a multi-adjective string. (Which is not to say that further exploration of classifier-based models is not warranted.)

3. *The hypothesis that complex concept formation is involved to a significant degree cannot be ruled out, because the set of AAN-combinations in Experiment 1 is limited, and this hypothesis was not tested in Experiment 2.*

We agree that the set of AAN-combinations in Expt. 1 was limited, which is why we in fact did test the concept-formability hypothesis in Expt. 2: in both cases, we failed to find any evidence that ordering preferences depend on the modified noun. However, we also found the concept-formability hypotheses intuitively appealing, so we followed up on this result by replicating our naturalness rating experiment with a new set of nouns chosen to maximize the probability of by-noun effects. Complex concepts will be described using the two-word name, presumably yielding more occurrences of this bigram than would be expected from the unigram frequencies of the noun and adjective. We thus chose new nouns whose co-occurrence probability with our 26 adjectives in the BNC is far greater than one would predict on the basis of their individual word probabilities. We used these new materials in a direct replication of our naturalness ratings experiment, and performed the same nested model comparison predicting naturalness ratings either by ADJECTIVE only, or by ADJECTIVE together with its interaction with NOUN. The model comparison revealed that noun-specific ratings did not explain any additional variance in ordering preference beyond adjective-level ratings. However, adjective subjectivity scores (obtained in *Expt 1.2 Subjectivity*) account for 85% of the variance in the new naturalness ratings (r^2=0.85, 95% CI [0.64, 0.93]). Thus, while we fail to find evidence of noun-specific effects both in our original materials and in materials specifically designed to deliver such effects, we continue to see that subjectivity predicts adjective ordering preferences. We report on the details of this replication in the new supplement to our manuscript.

4. *The proposed psychological mechanism for adjective order seems counterintuitive. If it was indeed the case that more subjective adjectives are ruling out miscommunications, then they should come first in post nominal positions. This is indeed the case for Romance languages, for example. However, then one would expect the same linear order in prenominal positions.*

   Our speculation is that *less* subjective content rules out miscommunications, which is why less subjective content appears closer to the noun in both pre- and post-nominal languages. This idea relies on a notion of hierarchical—rather than simply linear—ordering, and implicates diachronic processes that likely deliver the preferences we observe today. We have added discussion of these points to the manuscript.

5. *The cartographic approach does not seek adjective ordering preferences in their structure (that sounds as if there is something about their morphology that determines their order); it attempts to build a structural model.*

   We agree, and we did not intend to suggest that cartographic approaches were concerned with adjective morphology. To avoid this confusion, we have clarified our language in discussing these approaches, abandoning the term "cartographic" altogether.

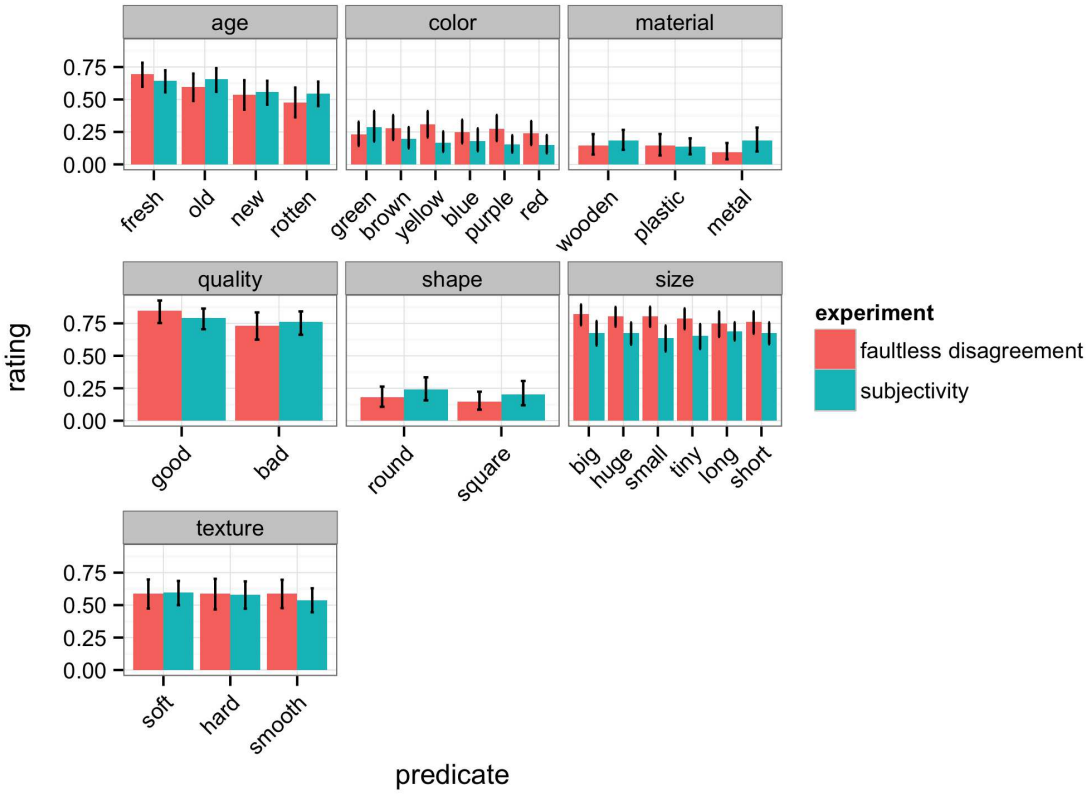
6. *The idea that miscommunication increases with vague adjectives is a very relevant hypothesis to the author's model, and would be one (interesting) way of expanding the current paper.*

   We agree, and have included discussion of this idea in our manuscript. However, we believe expanding the current paper with further exploration of this point would take us

well beyond the scope—and word limits—of the paper. We believe these explorations are better suited to future research.

7. *Table 1: Why are there different numbers of adjectives in each category? Isn't there a big danger that the effects are driven by individual items?*

   There are different numbers of adjectives in each category because we initially selected adjectives that occur relatively frequently while trying to nevertheless cover as many different adjective classes as possible, resulting in unbalanced cells. As for the reviewer's question regarding the danger of outlier-driven effects: almost all of the analyses we report were conducted on the adjective level, not the adjective class level, avoiding the possibility of inducing a class-level bias via an outlier in that class. The concern is greater for the subjectivity difference score analysis: here, we predicted naturalness ratings from a difference score that was computed by subtracting adjective *class* means from each other. However, there was very little variability—and thus, no outliers—among adjectives within each class, either in subjectivity or faultless disagreement scores. We demonstrate this for the 26 Expt. 1 adjectives in the following graph, which plots mean faultless disagreement and subjectivity scores:



8. *It would be an even stronger argument for the authors if there were ordering preferences within the same semantic class, based on subjectivity.*

   This is an excellent point: semantic classes serve merely to structure and smooth our data; subjectivity should apply within—as well as across—the semantic classes we test.

However, we are unable to test this prediction with the current experiments. To avoid nonsensical object descriptions in our experimental items, we avoided pairing adjectives from the same semantic class, so we can't determine ordering preferences within-class.

9. *(naturalness) − > (preference), or vice versa everywhere else*

   We operationalized adjective ordering preferences as naturalness ratings for specific sequences of adjectives, which we obtained in the behavioral experiments. We believe that it is best not to conflate these two notions—naturalness ratings and the preferences they reflect.

Reviewer 2:

1. *The authors suggest a number of alternative accounts of adjective ordering, for example a set of noun classes (as identified by Dixon or used by Cinque), or inherent-ness. They seem to suggest that subjectivity does a better job of operationalizing the relevant semantic notion, and indeed it does explain an impressive amount of the variance in the data. However I would find the paper *much* more convincing if the authors actually tested this by doing some model comparison.*

   We followed this advice, operationalizing alternative hypotheses and comparing their predictions with those made by subjectivity. We included the additional experiments and analyses in the new *Supporting information: Comparing subjectivity with alternative accounts of adjective order* (cf. our response to the Editor's point 2 and Reviewer 1's point 1).

2. *I'd like to see a bit more nuance in the presentation of the general question, in particular how previous work has explained adjective ordering based on the interaction of semantics and syntax. The authors set things up as a debate between whether syntax or semantics explains adjective ordering. But actually I think that's a conflation of two distinct (though related) questions: What is the ultimate explanation for why certain (types of) adjectives occur closer to nouns than others (across languages)? And what is the formalization/representation/conventionalization of these ordering preferences in the syntax of a language? Syntacticians like Cinque are largely interested in the latter. But that doesn't mean that they think the universal structure of the DP is unrelated to semantic features of adjectives. I think this comes up again in the discussion/conclusion, since what the authors end up arguing for is that a particular semantic notion, subjectivity, partially determines order. But of course that leaves open the further question of how this interacts with the syntactic grammar of a language (English or otherwise). Maybe they think that the grammar encodes subjectivity along with frequency, length, etc., together these probabilistically determine order, and that's all there is to it. But it also could be that subjectivity is the ultimate source of a more deterministic set of word order rules, and the latter are encoded in the grammar. The latter seems more in line which how they introduce the issue – there are these very systematic preferences that people show within and across languages. If subjectivity is really doing so much of the work, then it seems to me that ordering might actually be predicted to be quite flexible. In an individual instance, subjectivity will be determined*

largely by context, and thus you'd imagine that certain adjectives could vary quite dramatically in their relative ordering. Is that the case? If not, then something still needs to be said about how subjectivity and conventionalized syntax interact.

We agree that the discussion as we had it conflated why we observe the preferences we do and how the preferences are conventionalized. We have clarified our language, both in the paper's introduction and its general discussion, in order to separate these two issues. Moreover, we have stressed that it is the first question—namely, why we observe the preference we do—that we are after. We leave open this issue of how these preferences are conventionalized—settling this issue would clarify the predictions with respect to shifting contexts. However, it bears noting that the ordering preferences we observed *are* quite gradient and flexible, especially from the perspective of a deterministic syntactic account (cf. our response to Reviewer 2's point 7).

3. *I'm afraid I find the explanation offered about why decreased subjectivity means closer to the noun very unconvincing. (I'm partial to the "inherentness" explanation myself.) Maybe the authors can elaborate, otherwise, I'm not sure what it adds.*

   We have clarified our language so that what previously came across as an explanation is now clearly framed as speculation on the basis of our experimental results. We have also added to this discussion to better communicate the intuition and the directions for future work that it suggests (cf. our response to the Editor's point 2 and Reviewer 1's point 4). As for "inherentness", we tested its predictions in a follow-up experiment and found that adjective inherentness—as operationalized by previous researchers—accounts for little of the variation in the ordering preferences; we included this experiment in the new *Supporting information: Comparing subjectivity with alternative accounts of adjective order*.

4. *"...'cartographic approach", one could say..." I find this an odd comment. On the one hand, Dixon was less interested in claims about syntactic structure than in understanding the semantics types of adjectives and how they are distributed across languages. On the other hand the syntactic theory that Cinque subscribes to is literally called Cartography.*

   We have removed the term "cartographic" (cf. our response to Reviewer 1's point 5).

5. *Any comments about why noun-specific effects are not found? Along the lines of my long-winded comment (2) above, it seems like the subjectivity account might easily predict that particular adjectives might be more subjective when applied to certain nouns than others. But maybe your data just aren't able to actually show that, and with more data per noun (with different adjectives) you'd see it? Are there noun-specific differences in subjectivity ratings? Do you think they are simply ignoring the nouns in the ordering judgment task?*

   We were surprised by the absence of noun-specific effects, which is why we followed up on them with new materials in a new experiment that collected more data per noun. However, once again, we failed to find evidence of noun-specific effects (cf. our response to Reviewer 1's point 3). Our subjectivity ratings did not include noun-specific

information (participants rated adjective subjectivity in isolation, without a noun). However, we followed up on our basic subjectivity experiment (*Expt. 1: Subjectivity*) with a version in which participants rated the subjectivity of adjective-noun object descriptions (e.g., "How subjective is the description 'big banana'?"). There, we failed to find noun-specific differences in subjectivity ratings. However, these new subjectivity scores continue to predict the ordering preferences with remarkable accuracy (r^2=0.81, 95% CI [0.68, 0.89]; cf. the r^2 value of 0.85 for the noun-less subjectivity scores). We chose not to include the null result of this new subjectivity experiment in our revision, given the brevity of the paper.

As for whether participants were simply ignoring the nouns in our ordering preference experiments, we have clear evidence suggesting that they were not. In *Expt. 2: Ordering preferences* we asked participants to judge whether the adjective-adjective-noun descriptions they saw made sense. To see whether nouns had an effect on sense rates, we performed a nested linear model comparison. The models we compared predicted whether a configuration made sense either with the two ADJECTIVES only, or with the ADJECTIVES and the NOUN. The model comparison revealed that noun information does explain additional variance in sense rates beyond the two adjectives involved ($F(165, 11267) = 2.03, p < 0.001$). We have added this analysis to the writeup of Expt. 2.

6. *Can you say something more about the analyses that use adjective class configurations? Why are the subjectivity ratings sometimes better at accounting for these?*

   We assume that the reviewer is referring to the difference between analyzing the data as visualized in Fig. 6 (where mean naturalness of each adjective as occurring in first position was predicted from that adjective's mean subjectivity) vs. in Fig. 7 (where each class_1-class_2 naturalness mean was predicted from the difference in subjectivity between class_1 and class_2). The class-level difference score analysis yielded a stronger correlation between subjectivity and ordering preferences than the adjective-level one because the noise that is present in the adjective-level ratings was smoothed when collapsing across adjective classes.

7. *Can you show that adjectives that are similar in subjectivity have more flexibility in relative order when they occur together? Would be nice to have some examples of this.*

   We included the configuration analyses in part to show that preferences weaken as the difference in subjectivity between two adjectives approaches zero. We now explicitly mention this point and give examples of relevant adjective pairs (e.g., "yellow square..." and "fresh soft...").

Reviewer 3:

1. *My primary worry is that I don't know how other theories do on this kind of data, so I can't tell how good the author's theory is. In the introduction, the authors outline a number of other features that prior authors thought determined adjective ordering including "inherentness", conceptual combination, specificity, etc. Does subjectivity matter over and above those other factors? Does it explain more variance than they do*

individually? Or overall? In this kind of domain where there's lots of ideas, it's hard to argue for one idea by showing that it works well; better is to show that it outperforms other prior theories. The authors state that there are difficulties in formalizing these ideas, but I don't see why subjectivity is any different. They could pick a reasonable formalization of any of these other factors (ratings or corpus measures) and look to see how well it predicts the ordering preferences. Doing this is particularly important since I'd think that subjectivity will tend to be correlated with some of these other predictors.

We followed this advice, picking reasonable formalizations of three alternative factors suggested by the literature and testing to see how well the formalizations predict ordering preferences (cf. our response to the Editor's point 2, Reviewer 1's point 1, and Reviewer 2's point 1). Subjectivity indeed was highly-correlated with the intersective/subsective distinction. However, subjectivity explained ordering preferences over and above this distinction.

2. *My second (perhaps ignorant) question is why we should think that anything abstract determines ordering preferences. If I just look at the frequency of usage—average distance from an adjective to its noun in a corpus–do I find that predicts ordering preference on new constructions? That could suggest that we just store some ordinal/weighting information about adjective position attached to each adjective. Of course in that analysis the causality could go either way... But in general I'd like to know what kind of evidence rules out those simple frequency/usage-based theories?*

   We explored this prediction of the usage-based account in the corpus validation of our ordering preferences in Expt. 1, which connected naturalness ratings with usage. However, many of the adjectives from Expt. 2 occur infrequently in multi-adjective strings in corpora—our cutoff was adjectives with just a single occurrence in a multi-adjective string—so there are no statistics to track for them. Moreover, a usage-based account leaves unexplained the striking success of subjectivity in predicting ordering preferences, and also begs the question of why these ordering preferences should be so consistent across languages.

3. *What if you gave me a new word and varied the subjectivity of its meaning. A fruit is "paxy" if you tend to like its color (subjective) vs. if it has more than 10 seeds (objective). Would you see ordering preferences for "paxy" depending on which meaning you gave people? If people really in their heads have a subjectivity $->$ order rule, it should apply on novel words like that, no? In general, it would be especially nice in this paper to make and test some novel predictions beyond just ratings for ordering preferences and adjective measures.*

   If subjectivity indeed is operating as an online heuristic while people are stringing together adjectives, we would predict novel adjectives to linearize according to the subjectivity of the properties they name. And if we could teach participants novel adjectives with varying degrees of subjectivity, as suggested, this experiment would be a good test of the prediction. However, we have been careful in our revisions to differentiate the success of subjectivity in accounting for *existing* preferences from the *conventionalization* of those preferences; it is this latter point that the experiment

would be testing, but it is the former point that we focus on in our work (cf. our response to the Editor's point 1 and Reviewer 2's point 2).

4. *In section 3.3 (and perhaps elsewhere), the R^2 values are a little hard to interpret because they include variance from the noise in measuring the subjectivity preferences as well as the naturalness ratings. Because of this measurement error, you could never expect $R^2 = 1$, even for a perfect theory. A useful thing to do might be to do an analysis to see how much of the potentially explainable variance (e.g. subtracting off the measurement error) you actually do explain. Such an analysis would take into account the reliability of the x- and y-values in the correlation and adjust the correlation. Check out Spearman's Prophecy formula.*

   We performed these analyses on our ratings data, but the explainable variance was reliably high (i.e., >0.95), so we decided against including these analyses. (To compute the explainable variance in the naturalness rating data, we first computed the split-half correlation of the data to itself. To do so, we chose a random sample of half of the participants, then correlated their data with the data from the remaining participants. We repeated this process 100 times to compute the mean split-half correlation of our data. To "step up" the split-half correlation to provide an estimated correlation of the data with itself, we entered this value into the Spearman-Brown Prophecy formula.)

5. *I would like to know a little more what these findings tell us in a broader sense beyond adjective ordering. Do the authors expect general subjectivity effects in language? If so, why? I know they avoided saying too much about this due in order to avoid speculating, but I think talking about the bigger picture would help to situate this work for the broader cogsci community.*

   We do indeed expect general subjectivity effects in language, and we discuss this possibility in the concluding section of our paper where we bring up complex nominals and adverb ordering. However, as the reviewer suspects, we have been careful not to over-commit ourselves. The bigger-picture significance of our findings is the strong connection between a clear and robust linguistic universal (i.e., ordering preferences) and a clear and robust cognitive one (i.e., interpersonal subjectivity estimates), which suggests that language is shaped by our social experiences using it to communicate with each other. Understanding the principles of this relation will be an excellent direction for future research.

6. *How many subjects ran in multiple experiments?*

   Expt. 1: One participant ran in both the subjectivity and the order preference experiments, two participants ran in both the faultless disagreement and subjectivity experiments, and three participants ran in both the faultless disagreement and order preference experiments. Expt. 2: Twenty-two participants ran in both the subjectivity and the order preference experiments. Eighteen participants ran in both Expt. 1 and Expt. 2.

7. *When I think about the example of "big blue box" vs. "blue big box", these seem to elicit very different contexts in my imagination. In the first, there are many blue boxes, one*

of which is big; in the latter there are many big boxes, one of which is blue. So it feels like there is some difference in focus/pragmatics/context that is relevant to these effects. With respect to context, wouldn't you expect that different situations (like I just described) would bias people for one preference or the other? Maybe then there is not a single simple factors that matters?

We share the intuition, and agree that subjectivity is likely one of many factors that determine ordering preferences. We discuss other factors in our paper, including word length and frequency, as well as contrastiveness in discourse. It is this last factor that likely operates in the "big blue box" vs. "blue big box" example provided.

Thank you again for the thorough and thoughtful comments on our work. We hope that you will like the new version of the paper. Please let us know if you require additional information. We look forward to hearing from you!

Yours sincerely,

Gregory Scontras, Judith Degen, and Noah D. Goodman