An Information-Theoretic Explanation of Adjective Ordering Preferences

Michael Hahn[1], **Judith Degen**[1], **Noah Goodman**[1], **Dan Jurafsky**[1], **Richard Futrell**[2]
{mhahn2, jdegen, ngoodman, jurafsky}@stanford.edu, futrell@mit.edu
[1]Stanford University, [2]MIT

Abstract

Across languages, adjectives are subject to ordering restrictions. Recent research shows that these are predicted by adjective subjectivity, but the question remains open why this is the case. We first conduct a corpus study and not only replicate the subjectivity effect, but also find a previously undocumented effect of mutual information between adjectives and nouns. We then describe a rational model of adjective use in which listeners explicitly reason about judgments made by different speakers, formalizing the notion of subjectivity as agreement between speakers. We show that, once incremental processing is combined with memory limitations, our model predicts effects both of subjectivity and mutual information. We confirm the adequacy of our model by evaluating it on corpus data, finding that it correctly predicts ordering in unseen data with an accuracy of 96.2 %. This suggests that adjective ordering can be explained by general principles of human communication and language processing.

Introduction

Across languages, sequences of modifying adjectives show preferences for some orderings over others. In English, 'large wooden table' is preferred to 'wooden large table', and 'beautiful green shirt' is preferred to 'green beautiful shirt'. Such preferences exist across geographically and typologically diverse languages (Dixon, 1982; Sproat & Shih, 1991).

A variety of explanations for these preferences have been offered in the literature, including both semantic and syntactic ones. Syntactic accounts assume a rigid syntactic ordering of projections hosting different kinds of adjectives (Scott, 2002; Cinque, 2010). Semantic accounts have appealed to notions such as specificity (Ziff, 1960), inherentness (Whorf, 1945), absoluteness (Sproat & Shih, 1991), concept-formability (Svenonius, 2008), and subjectivity (Hetzron, 1978; Hill, 2012; Scontras, Degen, & Goodman, 2017).

While not all of these hypotheses have been verified on a broader empirical basis, there is strong empirical support for the idea that adjective subjectivity determines ordering: Scontras et al. (2017) compared order preferences with ratings of subjectivity for individual adjectives in English, and showed that subjectivity explained over 60 % of the variance in order preference ratings. They found that more subjective adjectives tend to occur before less subjective ones.

If these preferences occurred in only a few languages, it would be reasonable to accept this as an arbitrary fact of grammar. But the cross-linguistic stability of the patterns calls for a general explanation: As they occur in languages with widely different grammatical structures, we can expect that such an explanation will make reference to general principles of human communication and cognition. The aim of this paper is to present such an explanation. We first describe a corpus analysis, demonstrating effects of both subjectivity and mutual information on adjective ordering. We then provide an explanatory model of rational adjective use that predicts these effects, and verify that it correctly accounts for the corpus data.

Corpus Analysis: Subjectivity and Mutual Information Effects

While previous hypotheses about adjective ordering such as 'specificity' and 'inherentness' of adjectives to nouns (Ziff, 1960; Whorf, 1945)) suggest that adjective ordering should depend on the noun, Scontras et al. (2017) found no evidence for noun-specific effects. As their study used selected out-of-context noun phrases, one might wonder whether such effects can be shown using corpus data. As a formalization of specificity, we consider *Pointwise Mutual Information*:

$$\mathrm{PMI}(\mathrm{Adj},\mathrm{Noun}) = \log \mathrm{P}(\mathrm{Adj}|\mathrm{Noun}) - \log \mathrm{P}(\mathrm{Adj}) \qquad (1)$$

where $\mathrm{P}(\mathrm{Adj}|\mathrm{Noun})$ is the probability that the adjective Adj occurs, given the noun Noun. This concept is a common measure of collocation (Manning & Schuetze, 1999), and measures the degree to which the two words appear together more frequently than would be expected by chance. If an adjective is specific to a noun, we expect the adjective to appear more frequently with the noun than with most other nouns, which is captured by PMI. Following the specificity theory, our hypothesis is that adjectives with higher mutual information with the noun tend to come closer to the noun. Indeed, words with high mutual information occur closer together in language (Qian & Jaeger, 2012; Gildea & Jaeger, 2015).

Methods and Results We used the BookCorpus (Zhu et al., 2015), a corpus of 11,038 English novels, encompassing about 74 Million sentences.[1] We estimated mutual information between adjectives and nouns from a randomly selected set of sentences, amounting to about 70 % of the corpus. The conditional probabilities $\mathrm{P}(\mathrm{Noun}|\mathrm{Adj})$ are determined by counting all occurrences where Noun occurred directly after Adj. However, these counts will be impacted by the existing adjective ordering preferences, creating a potential confound. To eliminate this confound, we randomized the order of adjectives occurring in a sequence when counting occurrences.

We then extracted all occurrences of two adjectives between a determiner and a noun from a held-out section amounting to 10 % of the corpus. We retained those occurrences where both adjectives occurred in the experiment of

[1] A reviewer points out that BookCorpus contains duplicate novels, which might result in imprecise MI estimates. Future work should confirm our results on further datasets.

	β	SE	z	p
PMI A_1 – N	-0.501	0.041	-12.2	$< 2.2 \cdot 10^{-16}$
PMI A_2 – N	0.501	0.041	12.2	$< 2.2 \cdot 10^{-16}$
Subjectivity A_1	8.28	1.35	6.12	$9.36 \cdot 10^{-10}$
Subjectivity A_2	-8.28	1.35	-6.12	$9.36 \cdot 10^{-10}$

Table 1: Logistic mixed-effects model predicting whether two given adjectives A_1, A_2 were ordered as A_1A_2 (coded $+1$) or A_2A_1 (coded 0), from mutual information and subjectivity.

Scontras et al. (2017), in order to use their experimentally measured subjectivities. 4699 datapoints remained.

For each corpus, we ran a logistic mixed-effects model predicting the order of each pair of adjectives, including as fixed effects (1) subjectivity of the two adjectives from the data collected by Scontras et al. (2017), (2) mutual information between the noun and each of the two adjectives. The two adjectives were entered as random intercepts. The resulting models are shown in Table 1. We observed main effects of both mutual information and subjectivity, such that more objective adjectives and adjectives with higher mutual information with the noun occurred closer to it. Model comparison with a corresponding model without mutual information predictors (BIC 241, $p < 2.2 \cdot 10^{-16}$) or without subjectivity predictors (BIC 120, p = $2.3 \cdot 10^{-10}$) confirms that both types of predictors contribute independently.

The Function of Subjective Adjectives

Our goal is a formal model of adjective ordering preferences that falls out of considerations of communicatively efficient adjective use. One route is by understanding adjectives as *restricting a set of referents*, picking out from contextually given objects denoted by a noun the one that matches the adjective, as done in much of the literature on content selection in referring expressions (Sedivy, Tanenhaus, Chambers, & Carlson, 1999). However, establishing reference is not the only use of noun phrases, nor do all adjectives simply restrict a set of referents. Adjectives often are used non-restrictively, describing or commenting on some referent. As an example, consider 'Forrest looks at the massive crowd.' (Roth, 1993) – this does not mean that there were two crowds, and Forrest Gump chose to look at the one that was massive – instead, the sentence means that Forrest looked at the crowd, and the narrator (or Forrest) considered the crowd to be massive. In the literature, this use is known as non-restrictive use, as opposed to restrictive use that picks out one of the contextually given elements matching the noun. Our model will thus be centered not around restriction of reference, but around speakers communicating *descriptions* of and *attitudes* to referents.

Scontras et al. (2017) et al. used two operationalizations of subjectivity: In their main experiment, they directly asked participants 'how subjective' a given adjective was. They validated this measure by another experiment in which they described two people disagreeing on a judgment, and asking whether both people could conceivably be right. These measures were highly correlated ($r^2 = 0.91$).

The latter criterion is known as *faultless disagreement*: Adjectives are subjective if people can reasonably disagree, without anyone having to be in error (Kölbel, 2004). In line with the 'faultless disagreement' diagnostic, subjectivity is typically understood to refer to judgments whose truth is relative to individuals (Kölbel, 2004; Lasersohn, 2005).

It seems, therefore, that the most natural way of modeling subjective meaning is by explicitly making reference to the opinions of different persons. In our model, we will assume that listeners infer not just properties of objects, but they infer and reason about judgments made by different speakers.

A Model of Adjective Use

In this section, we describe a simple formal model of adjective use. Given that subjectivity essentially refers to the potential for disagreement across speakers, we will explicitly model judgments made by different speakers about objects. Judgments are objective if speakers tend to agree, while they are more subjective if speakers are less likely to agree.

We formalize adjectives as expressing judgments $A \in \{\text{green}, \text{beautiful}, \dots\}$, made by a person s about a referent x. In the case of highly objective adjectives, such as material adjectives, speakers will mostly agree on their judgments, while they may disagree for more subjective adjectives. The possible states of the worlds are truth-value assignments to the set of expressions

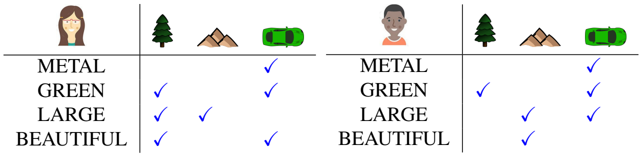
$$\{A(s,x) : A \text{ an adjective}, x \text{ a referent}, s \text{ a person}\}$$

where $A(s,x)$ indicates that person s judges referent x to have property A (e.g., green, beautiful, ...). We assume that there are fixed sets of persons, referents, and properties.

This is illustrated in Figure 1, showing a typical world state: Two speakers mostly agree on more objective judgments, such as material and color, and agree less on more subjective judgments such as size or beauty.

In our model, listeners aim to infer not just judgments of one speaker, but a full world state including multiple persons. This is useful when we consider that a listener might later interact with other persons. For instance, we expect that a listener learning that one of the persons in Figure 1 judges a referent to be 'green' will find it useful to infer that the other person likely applies the same judgment – that is, listeners will generalize objective judgments across people.

World Prior and Inter-Speaker Agreement A world state is a truth value assignment to all the expressions $A(s,x)$, across adjectives, persons, and objects. Speakers and listeners share probabilistic prior beliefs about which world states are more or less likely to be true, formalized by a prior distribution over world states. In our setting, adjectives differ in the *correlation* between judgments by different speakers about the same object. Formally, we assume that for each adjective A, there is a number $\kappa(A)$ such that, under the prior



👩	🌲	⛰	🚗
METAL			✓
GREEN	✓		✓
LARGE	✓	✓	
BEAUTIFUL	✓		✓

👨🏿	🌲	⛰	🚗
METAL			✓
GREEN	✓		✓
LARGE		✓	✓
BEAUTIFUL		✓	

Figure 1: A typical world state: Speakers are likely to agree on more objective judgments, and less likely to agree on more subjective judgments.

over world states, the Pearson correlation between the truth values of $A(s,x)$ and $A(s',x)$ is equal to $\kappa(A)$, whenever s,s' are two different speakers. In the special setting where there are two persons s_1,s_2, this reduces to two Bernoulli variables with fixed means and correlation, and we can write

$$\begin{aligned} &A(s_1,x) \sim Bernoulli(\phi) \\ &A(s_2,x) \sim Bernoulli((1-\kappa(A))\cdot\phi + \kappa(A)\cdot A(s_1,x)) \end{aligned} \tag{2}$$

with $\phi \in [0,1]$. The magnitude of $\kappa(A)$ formalizes correlation of judgments across speakers: Adjectives that show agreement across speakers have $\kappa(A)$ close to 1. For more subjective adjectives, $\kappa(A)$ is smaller.

Communication: Rational Listeners and Speakers In our model, speakers aim to communicate judgments about objects by uttering three-word phrases consisting of two adjectives and a noun. An utterance A_1A_2N is true for the speaker in a world if the speaker judges that the adjectives A_1,A_2 both apply to the referent of the noun. That is, the truth value depends on those parts of the world state that relate to the speaker, but not on those that relate to other persons. In this model, we assume that there is a known mapping from nouns to entities, though this assumption can be relaxed. Our model is couched in the framework of Bayesian models of communication (Franke, 2010; Frank & Goodman, 2012; Goodman & Frank, 2016), consisting of a literal listener and a speaker reasoning about the listener.

We will start with a listener who hears an utterance, and incrementally updates her belief about the world. While incrementality will not be necessary for deriving the core prediction of our model, we want to make explicit how the model fits with the known psycholinguistic fact that adjectives are processed incrementally (Sedivy et al., 1999). When hearing a sequence A_1A_2N, the listener maintains a buffer of the words heard so far, and conditions her belief by restricting to those worlds compatible with possible continuations of the buffer:

$$\begin{aligned} P^0_{listener}(w) &:= P_{prior}(w) \\ P^1_{listener}(w) &\propto P^0_{listener}(w)\cdot\delta_{\exists u = A_1A'_2N' \,:\, w\models_s u} \\ P^2_{listener}(w) &\propto P^1_{listener}(w)\cdot\delta_{\exists u = A_1A_2N' \,:\, w\models_s u} \\ P^3_{listener}(w) &\propto P^2_{listener}(w)\cdot\delta_{w\models_s A_1A_2N} \end{aligned} \tag{3}$$

where $w \models_s u$ is a shorthand for 'the utterance u is true for the speaker in the world state w', and $\delta_{\dots}$ is 1 if the condition in the subscript is true, else 0. In Figure 2, we visualize the incremental updates for the utterance 'big green tree'.

When choosing which utterance u to utter, speakers trade off communicative utility $U(u)$ and the cost of production $C(u)$ using a softmax decision rule:

$$P_{speaker}(u) \propto \exp(\alpha\cdot(U(u)-\beta\cdot C(u))) \tag{4}$$

Here, $\alpha > 0$ indicates the degree of rationality, while $\beta > 0$ trades off utility and cost. When the speaker has perfect knowledge of the world state, a natural choice for $U(u)$ is the negative surprisal of the true world state under the posterior belief of the listener (Frank & Goodman, 2012). In our case, the speaker does not have full information about the world, as she might not know the judgments made by other speakers. Therefore, we take for $U(u)$ the *expected* negative posterior surprisal of the ground truth about the other speakers: This quantity is equal (up to a constant independent of u) to the negative KL divergence between the speaker's belief and the listener's posterior belief after hearing the utterance – a common utility function in rational models of language use (Goodman & Stuhlmüller, 2013; Regier, Kemp, & Kay, 2015):

$$\begin{aligned} U(u) &:= -\,\mathrm{KL}(P_{speaker}||P_{listener}(\cdot|u)) \\ &= \sum_w P_{speaker}(w)\log\frac{P_{listener}(w|u)}{P_{speaker}(w)} \end{aligned} \tag{5}$$

We assume that $P_{speaker}$ is equal to the prior conditioned on the ground truth judgments of the speaker.

For the cost $C(u)$, we take the *surprisal* of the utterance $u = A_1A_2N$ according to a general language model – e.g., describing the statistics of a community's language use.

$$C(A_1A_2N) = -\log P(A_1A_2N) \tag{6}$$

Unlike the utility function, this cost function is purely a property of the surface string A_1A_2N in the statistics of the language, without reference to meaning. We assume that the speaker also computes these probabilities incrementally word-by-word. We assume that the language model encodes no prior ordering preference – both orderings of an adjective pair will have the same probability and thus the same cost as long as this probability $P(A_1A_2N)$ is evaluated exactly.

Adding Noise

So far, while sequences such as 'beautiful green tree' and 'green beautiful tree' result in different sequences of belief updates in the listener, the final result will be identical. Thus, both sequences so far have the same communicative utility. Similarly, in a setting where no prior order preferences are encoded in the language model, they have the same cost.

We now show how processing and specifically memory limitations break this symmetry, predicting both subjectivity and mutual information effects. It has by now been established that, during language production and language comprehension, linguistic material further in the past becomes

Timesteps	T_0	T_1	T_2	T_3
New Input		big	green	tree
Buffer		big	big green	big green tree
Compatible continuations	big beautiful tree beautiful green car . . .	big beautiful tree big green car . . .	big green tree big green car . . .	big green tree

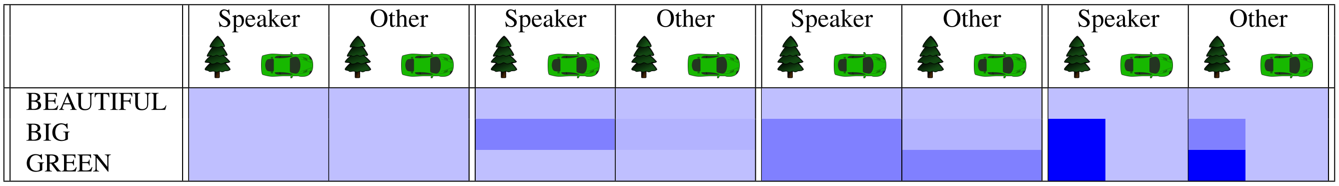
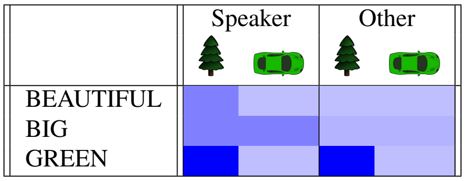


Figure 2: Simulated incremental inference in a listener hearing 'big green tree', about judgments made by the speaker and another person about two objects. The listener maintains a buffer of words received so far, and in each step, considers all possible continuations (top). The bottom part shows the listener's incremental posterior belief about the world. For expository purposes, we assume a simple setting where there are two persons, two objects, and three properties, with $\kappa(\textit{beautiful}) = 0.3$, $\kappa(\textit{big}) = 0.5$, $\kappa(\textit{green}) = 0.95$. The strength of the color in each cell indicates the listener's degree of belief that the given person (column) would judge a given property (row) to apply the given object (column). In each step, the listener considers all world states that are compatible with potential continuations of the buffer, and accordingly updates her belief about the speaker's judgments. To the extent that persons tend to agree about properties, the listener can infer that the other person likely has the same judgments. This effect is strong for the objective property ('green'), and weak for the subjective property ('big').

Timesteps	T_3
New Input	tree
Buffer	??? green tree
Compatible continuations	beautiful green tree big green tree



Timesteps	T_3
New Input	tree
Buffer	??? big tree
Compatible continuations	beautiful big tree green big tree

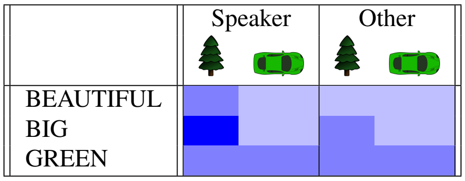


Figure 3: Simulated posterior listener belief if the first adjective is lost when the noun is reached, for input 'big green tree' (top) and 'green big tree' (bottom), in the same setting as Figure 2. If the objective adjective is retained (top), information generalizes across speakers. Retaining the subjective adjective (bottom) is less useful due to potential for disagreement between speakers.

harder to access and integrate with new material. Classical families of examples is include dependency locality effects (Gibson, 1998) and models of cue retrieval in sentence processing (McElree, 2001; Lewis & Vasishth, 2005).

To formally integrate such memory limitations into our model, we follow Futrell and Levy (2017), assuming that during incremental processing, previous words in the input may be deleted stochastically. Crucially, the probability of a word being deleted increases as one goes further back in the sequence (Futrell & Levy, 2017).

In our model, there are two places where incremental processing can be affected by noise: the listener's incremental belief updates, and the computation of cost.

Noisy Belief Updates First, let us consider what happens when the listener's buffer is affected by progressive noise. Let us consider the simple case where, at each step, at most the two last words were integrated: The belief updates after hearing the two adjectives are as before. When encountering the noun, the first adjective is the furthest away from the current input word, and – in this case – deleted from the buffer.

When computing the posterior, only the last two words are available, and the listener considers the possible completions of the now incomplete buffer (compare Equation 3):

$$\widehat{P}^3_{listener}(w) \propto P^2_{listener}(w) \cdot \delta_{\exists A'_1 : w \models_s A'_1 A_2 N} \qquad (7)$$

where, as before, $w \models_s u$ is a shorthand for 'the utterance u is true for the speaker in the world state w'. As noise is stochastic, utility $U(u)$ is now the *expected* KL-divergence, where the expectation is taken over the possible noise patterns.

In Figure 3, we illustrate the listener's state when reaching the noun, for the two possible orderings of the more subjective adjective 'big' and the less subjective adjective 'green'. Depending on which adjective was subject to deletion, the listener has different posterior beliefs not just about the speaker, but also about the other person: Due to the objective nature of 'green', integrating this adjective (top) provides information that generalizes across speakers. As loss is progressive, the first adjective is more likely to be lost when the noun is reached. Thus, placing the objective adjective closer to the noun is predicted to, on average, result in lower levels of uncertainty about the full state of the world.

Noise in the Cost The second place that involves incremental computation and will thus be affected by progressive noise is the cost term. If the first adjective is lost when computing the conditional probability $P(N|A_1A_2)$ of the noun in context, the calculation marginalizes out the first adjective and the resulting quantity will be $P(N|A_2)$. Thus, cost will be estimated as $-\log P(A_1)P(A_2|A_1)P(N|A_2)$ in this case. Using the definition of PMI, we can write

$$C(A_1A_2N) - C(A_2A_1N) = \lambda \cdot (\mathrm{PMI}(A_1,N) - \mathrm{PMI}(A_2,N)) \tag{8}$$

More generally, Futrell and Levy (2017) show that the estimated surprisal will be biased towards this value when loss is progressive. Thus, we predict that putting the adjective with higher PMI with the noun closer to it results in lower cost.

Simulations

We implemented the model in the probabilistic programming language WebPPL (Goodman & Stuhlmüller, 2014). We constructed contexts with 20 objects, four properties, and two persons (one speaker and one other person). For simplicity, we consider the case where only the first adjective is subject to loss, at loss rate $\lambda \in [0,1]$. For inference, we randomly sample 10,000 worlds from the world prior and compute the listener model by exact enumeration of these samples.

We have described how the model predicts subjectivity and mutual information effects, but one might wonder how robust these effects are to changes in parameter values. We considered the predictions the model makes for different values of the inter-speaker correlations $\kappa(A_1),...,\kappa(A_{n_{adj}})$, the loss probability λ, the rationality parameters $\alpha,\beta > 0$, and the prior probability ϕ in (2). We sampled $\alpha \sim \Gamma(5,1)$, $\beta \sim \Gamma(5,1)$, $\lambda \sim \mathrm{Uniform}(0,1)$, $\phi \sim \sigma(\mathcal{N}(0,0.5))$.

We first considered the setup where A_1 is more subjective than A_2 – that is, $\kappa(A_1) < \kappa(A_2)$, while taking $\mathrm{PMI}(A_1,N) = \mathrm{PMI}(A_2,N)$. The correlations for other adjectives are uniformly random. We sampled 10,000 parameter settings subject to this constraint. For every single setting, we found $U(A_1A_2N) > U(A_2A_1N)$ – placing adjectives with higher inter-speaker correlation closer to the noun increased utility. In Figure 4, we plot utility difference as a function of $\kappa(A_1) - \kappa(A_2)$. Utility difference is directly proportional to the difference in inter-speaker correlations.

We then carried out the same with $\kappa(A_1) = \kappa(A_2)$ and $\mathrm{PMI}(A_1,N) < \mathrm{PMI}(A_2,N)$ – that is, assuming that A_2 is more predictive of the noun. In this case, as shown in Equation 8, difference in cost is proportional to difference in PMI (Figure 4). Thus, assuming noisy memory, both subjectivity and MI are predicted to affect order preferences of the speaker.

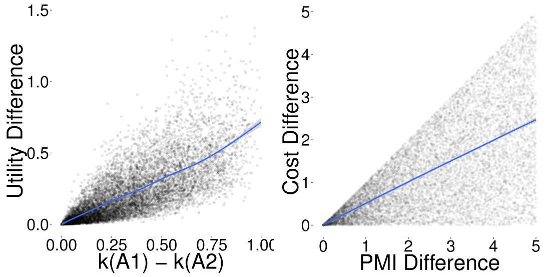


Figure 4: Left: Utility Difference between Orderings, as a function of the difference between inter-speaker correlations $\kappa(A_1)$, $\kappa(A_2)$. Across parameter settings, putting the subjective adjective earlier results in higher utility. Right: Cost difference, as a function of PMI difference. Placing the adjective with higher PMI closer to the noun results in lower cost. Both plots show LOESS-smoothed means.

Testing against Corpus Data

We tested the speaker model against the data from our corpus analysis described above. For the inter-speaker correlations $\kappa(A)$, we took $\kappa(A)$ to be one minus the average subjectivity score from Scontras et al. (2017). For the cost term, we used mutual information data from the corpus analysis.

We used Bayesian data analysis to infer the numerical parameters of our model (rationality parameters α,β, loss rate λ, prior probabilty ϕ) from the BookCorpus data we used in the beginning. We specified priors $\alpha \cdot \lambda \sim \mathcal{N}(0,10)$, $\alpha \cdot \beta \cdot \lambda \sim \mathcal{N}(0,10)$, $\phi \sim \sigma(\mathcal{N}(0,2))$, where σ is the inverse-logit function.[2] To obtain approximate posterior distributions, we used variational inference with minibatches in Pyro (`http://pyro.ai/`). We obtained posterior means $\alpha \cdot \lambda = 5.07$ ($\sigma^2 = 0.243$), $\alpha \cdot \beta \cdot \lambda = 0.39$ ($\sigma^2 = 0.033$), and $\phi = 0.095$ ($logit(\phi) = -2.253$, $\sigma^2 = 0.13$). The fitted values suggest that utility is weighted much more strongly than cost, and that most judgments are relatively unlikely a priori.

Plugging in the posterior means for these parameters, the model achieves a classification accuracy of 93.7 % on the task of predicting adjective order on the dataset.[3] A model with only the cost term would achieve an accuracy of 69 %, while a model with only the utility term achieves an accuracy of 93.3 %, very close to the accuracy of the full model.[4] This highlights the central role of the utility term – and thus sub-

[2] Our model does not make it possible to obtain independent estimates of α,β and λ.

[3] Logistic regression models with surprisal and PMI predictors would achieve the same accuracy. However, note that our model is an explanatory cognitive model, as opposed to a data analysis.

[4] While the cost term does not contribute much in terms of accuracy, a mixed-effects analysis analogous to the corpus analysis above confirms that it contributes significantly ($p < 2.2 \cdot 10^{-16}$).

jectivity – for ordering. To test whether results generalize to unseen data, we used a further held-out 20 % of the corpus. Classification accuracy was 93.1 %.

As the prior probability that an adjective is applied to objects might not be uniform, we also considered the setting where ϕ in (2) varies with the adjective. We assumed a hierarchical model with hyperparameters $\phi_0 \sim \mathcal{N}(0,2)$, $S^2 \sim \mathcal{N}(0,1)$, and parameters $\phi(A) \sim \sigma(\mathcal{N}(\phi_0, S^2))$ for each adjective A. We obtained similar estimates: $\alpha \cdot \lambda = 5.6$ $(\sigma^2 = 0.25)$,$\alpha \cdot \beta = 0.36$ $(\sigma^2 = 0.088)$, $\phi_0 = -2.1$ $(\sigma^2 = 0.12)$, $S^2 = 0.31$ $(\sigma^2 = 0.069)$. Classification accuracy increases to 97.3 % on the original dataset, and to 96.2 % on the held-out set, which shows that the improvement obtained from the increase in model complexity generalizes to unseen data. Future research should test the prediction that the inferred values for $\phi(A)$ correspond to the prior probability that a speaker would apply a given adjective to an object (Equation 2).

Discussion

We provided an explanatory cognitive model of adjective ordering, building on the insight that subjectivity and specificity, formalized by mutual information, impact adjective ordering. We first conducted a corpus study and showed that ordering is impacted independently by subjectivity and mutual information. We then presented a model of adjective use in which listeners infer judgments made by speakers and other persons. We integrated this model with a recent model of memory limitations in language processing, and showed that it predicts both subjectivity and mutual information effects. We evaluated the model on corpus data, finding that it predicts adjective ordering with an accuracy of 96.2 %. In the following, we discuss some of the implications of this work, and questions that it raises.

Research has shown that subjective material more generally tends to appear at the periphery of phrases and clauses, and that diachronic meaning change towards more subjective meanings correlates with movement to the periphery (Traugott, 2010). This is in line with our proposal: Our analysis should equally apply to other types of subjective material, predicting that memory limitations favor placing them further away from the head. Future research should test our model on other types of subjective content.

We have assumed that the speaker's communicative goal is communicating descriptions and attitudes, rather than establishing reference. This was motivated by the observation that adjectives are often not used for establishing reference. Future research should compare our account with accounts of adjective ordering preferences that rely on the assumption that adjectives are used primarily for reference resolution.

In languages where adjectives follow the noun, such as Spanish or Arabic, typically the reverse order is observed (Dixon, 1982). Our account seems to make the correct prediction: In such languages, the noun is more likely to be lost when the second (subjective, in this case) adjective is reached. We furthermore make the prediction that, in such languages, adjectives with higher mutual information with the noun will also be more likely to come closer to the noun.

Recently, Dye, Milin, Futrell, and Ramscar (2017) interpreted prenominal modifiers as smoothing entropy, making nouns more equally predictable, and speculated that this may account for adjective ordering preferences. A notable difference between this theory and ours is that theirs predicts major differences between ordering patterns of prenominal and postnominal adjectives, whereas ours is symmetric.

In conclusion, the work reported here suggests that adjective ordering preferences are plausibly the result of efficiently trading off cost and informational utility of utterances for the purpose of communicating maximally generalizable information about objects.

References

Cinque, G. (2010). *The syntax of adjectives*. MIT Press.

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

Dye, M., Milin, P., Futrell, R., & Ramscar, M. (2017). Cute Little Puppies and Nice Cold Beers: An Information Theoretic Analysis of Prenominal Adjectives. In *39th Annual Meeting of the Cognitive Science Society, London, UK*.

Frank, M. C., & Goodman, N. D. (2012). Predicting Pragmatic Reasoning in Language Games. *Science*, *336*, 998–998.

Franke, M. (2010). *Signal to Act*. Unpublished doctoral dissertation.

Futrell, R., & Levy, R. (2017). Noisy-context surprisal as a human sentence processing cost model. In *Proceedings of EACL*.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, *68*(1), 1–76.

Gildea, D., & Jaeger, T. F. (2015). Human languages order information efficiently. *arXiv:1510.02823 [cs]*. (arXiv: 1510.02823)

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

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and Implicature: Modeling Language Understanding as Social Cognition. *Topics in Cognitive Science*, *5*, 173–184.

Goodman, N. D., & Stuhlmüller, A. (2014). *The Design and Implementation of Probabilistic Programming Languages*.

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

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).

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

Lasersohn, P. (2005). Context Dependence, Disagreement, and Predicates of Personal Taste. *Linguistics and Philosophy*, *28*(6), 643–686.

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*(3), 375–419.

Manning, C., & Schuetze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

McElree, B. (2001). Working Memory and Focal Attention. *Journal of experimental psychology. Learning, memory, and cognition*, *27*(3), 817–835.

Qian, T., & Jaeger, T. F. (2012). Cue Effectiveness in Communicatively Efficient Discourse Production. *Cognitive Science*, *36*(7), 1312–1336.

Regier, T., Kemp, C., & Kay, P. (2015). Word Meanings across Languages Support Efficient Communication. *The handbook of language emergence*, *87*, 237.

Roth, E. (1993). *Forrest Gump* [screenplay].

Scontras, G., Degen, J., & Goodman, N. D. (2017). Subjectivity Predicts Adjective Ordering Preferences. *Open Mind*, *1*, 53–66.

Scott, G.-J. (2002). Stacked adjectival modification and the structure of nominal phrases. In G. Cinque (Ed.), *The cartography of syntactic structures* (pp. 91–120). Oxford.

Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G., & Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, *71*(2), 109–147.

Sproat, R., & Shih, C. (1991). The Cross-Linguistic Distribution of Adjective Ordering Restrictions. In C. Georgopoulos & R. Ishihara (Eds.), *Interdisciplinary Approaches to Language*.

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

Traugott, E. C. (2010). Revisiting subjectification and intersubjectification. In K. Davidse, L. Vandelanotte, & H. Cuyckens (Eds.), *Subjectification, intersubjectification and grammaticalization* (pp. 29–71). Berlin: De Gruyter Mouton.

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

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books. *arXiv:1506.06724 [cs]*. (arXiv: 1506.06724)

Ziff, P. (1960). *Semantic Analysis*. Ithaca, NY.