Prenominal adjective order is such a fat big deal because adjectives are ordered by likely need

Chris Westbury¹

Published online: 22 July 2020
© The Psychonomic Society, Inc. 2020

Abstract

When multiple adjectives precede a noun in English, they are often ordered in a way that is implicitly understood by all fluent speakers of the language. Adjective order might be described as a *big fat deal*, but to describe it as a *fat big deal* betrays a lack of knowledge of English. Sweet (*A New English Grammar: Part II*, 1898/1955) proposed two related semantic principles to explain the phenomenon: *definiteness of denotation* (adjectives that denote a property that is most independent of the modified noun must be placed furthest from that noun) and *closeness of adjective/noun in meaning* (adjectives that denote properties essential to or inherent in the modified noun are placed closer to the noun). These observational descriptions of the phenomenon have received experimental support (Martin, *Journal of Verbal Learning and Verbal Behavior*, 8(6), 697–704, 1969). However, the issue of *why* Sweet's rules are true has not yet been solved. I propose, operationalize, test, and find strong support for a simple theory: that prenominal adjective order reflects *likely need*, the *a priori* probability that a particular adjective will be needed.

Keywords Adjective order · Semantics · Information theory · Likely need

It has long been recognized that English prenominal adjectives must usually be ordered in a specific way when they are used in sequence. For example, English speakers can speak about *a beautiful red Ferrari* or *a crazy new idea*, but (except in cases where we are trying to emphasize a particular contrast or responding in a particular context) we cannot normally speak of *a red beautiful Ferrari* or *a new crazy idea*. In this paper, I propose a new theory for adjective ordering, which is that adjectives are ordered in terms of *likely need*, the *a priori* probability that a particular adjective will be needed.

Background

Some attempts have been made to explain adjective order in terms of syntactic principles (e.g., Annear, 1964; Fries, 1952; Sussex, 1974; Vendler, 1963, 1968). These are largely deprecated now, in part because other researchers (Danks & Glucksberg, 1971; Martin, 1969) judged the proposals to be

ad hoc and in part because attempts to study the neurological underpinnings of adjective ordering have consistently implicated *semantic* rather than *syntactic* processes (Huang & Federmeier, 2012; Kemmerer, 2000; Kemmerer, Tranel, & Zdzanzyk, 2009).

There are a small number of words often coded as adjectives (notably, determiners such *several*, *many*, *single*, and *only*) that must be followed by a noun phrase due to their meaning, and therefore must go first when they are contiguous with a prenominal adjective. Numbers must precede the entire noun phrase when they modify one. We can say *two leaky boats* but never *leaky two boats*. A few adjectives have a similar quality of only making sense if they directly modify the whole noun phrase, and are therefore obligatorily placed before another adjective. For example, when we speak of the *French political system*, it is the political system that is French, so it would not make sense to speak of the *political French system*, which has an entirely different (and opaque) meaning. However, many strongly ordered adjectives have no such obvious reason for their ordering.

Both the oldest and most recent approaches to the issue explain adjective order in terms of semantic principles. In an early discussion of the issue, Sweet (1898/1955) suggested two non-mutually-exclusive semantic principles that could determine adjective order. One of his principles is *closeness*

✉ Chris Westbury
chrisw@ualberta.ca

¹ Department of Psychology, University of Alberta, P220 Biological Sciences Building, Edmonton, AB T6G 2E9, Canada

of adjective/noun meaning: “When a noun has more than one modifier, the general principle that the one most closely connected to it in meaning comes next to it” (p. 8). Gender is more essential to an animal than its body size, so by default we have to describe a large feline with two X chromosomes as *a fat female cat* and not *a female fat cat*. Sweet’s second principle is *definiteness of denotation*: adjectives that express a viewpoint must precede adjectives that report an inherent property and “there is gradation in increasing specialization from the beginning to the end of such a group” (p. 8; see also Whorf, 1945). The redness of the beautiful red Ferrari is a part of the car, but its beauty is in the eye of the beholder, so adjectives expressing beauty have to precede adjectives expressing color.

Sweet’s principles are consistent with the fact that very general modifiers (i.e., words that can modify many different nouns) such as *other*, *only*, *certain*, *particular*, and other determiners must always precede any other adjective, as in *the other/only red Ferrari*, *a certain/particular academic paper* or *three delicious green candies*. They are also consistent with the fact that increased concreteness is associated with adjectives closer to the noun (as noted by Martin, 1969), since the more definite terms are often more definite because they refer to concrete sensory properties of the world. Whether they *explain* those consistencies is open for debate. We might prefer as an explanation a principle that tells us how these semantic consistencies are computed (i.e., an explanation cast at a lower level than semantics).

After examining Sweet’s two principles experimentally, Martin (1969) concluded that definiteness of denotation was the main determinant of adjective order, though it was often confounded with closeness of adjective/noun meaning. Martin furthermore noted that more frequent adjectives were more likely to be placed further from the noun than were less frequent adjectives. This is also consistent with Sweet’s principles since less frequent words are likely to occur in fewer contexts than more frequent words and are therefore likely to be more specific in meaning. For example, the generally applicable word *pretty* is much more common (69,520 occurrences in the 560 million word Corpus of Contemporary American English [COCA; corpus.byu.edu/coca/]) than the much more specific word *blond* (6,236 COCA occurrences), so *a pretty blond lady* is strongly preferable to *a blond pretty lady*. However, Martin and Ferb (1973) pointed out that terms that have never been encountered before are immediately ordered correctly when they are learned. The fact that newly encountered adjectives are ordered appropriately as soon as their meaning is known suggests that individual word frequencies must play a relatively small role.

More recent attempts to explain prenominal adjective order have been more specific about the semantic categories to which adjectives belong. For example, Dixon (1977) proposed that adjectives were ordered according to the following order schema: *value*, *dimension*, *physical property*, *speed*,

human propensity, *age* and *color*. In the extreme, we should prefer *the wonderful large heavy slow loving old brown dog* over any other order. Dixon’s category order is, yet again, generally consistent with Sweet’s (1898/1955) more general and simpler principles. In discussing these categorical ordering attempts, Malouf (2000) noted that

unfortunately, while there are rules of thumb for ordering adjectives, none lend themselves to a computational implementation. For example, adjectives denoting size do tend to precede adjectives denoting color. However, these rules underspecify the relative order for many pairs of adjectives and are often difficult to apply in practice.” (p. 1)

Sweet’s principle of definiteness of denotation and Dixon’s (1977) order have recently been examined in a different way, as effects of subjectivity. As the examples of *the beautiful red Ferrari* and *the pretty blond lady* illustrate, less definite adjectives are sometimes indefinite because they are subjective. Scontras, Degen, and Goodman (2017) used Amazon Turk to get subjective assessments of the naturalness of the order of first 26 and then 196 concrete adjectives, as well as judgments of those words’ subjectivity. They demonstrated that the subjectivity judgments were good predictors of the prescribed order of the adjectives. In recognition of the fact that subjectivity judgments are themselves subjective, they followed this up with what they felt was a more objectively operationalized measure, which they called *faultness disagreement potential*. They asked 40 people to judge whether two speakers would both be right when they produced descriptions with the adjectives in opposite orders using the smaller adjective set. This measure was very highly correlated with the subjectivity scores.

This work was followed up in Simonič (2018) and Franke, Scontras, & Simonič (2019). They addressed the question of *why* subjectivity should play a role in adjective ordering by demonstrating in a highly constrained context that subjectivity ordering is a (likely but not necessary) consequence of the demands of successful reference resolution, using the rational speech act modeling framework (Frank & Goodman, 2012; Franke & Jäger, 2016; Goodman & Frank, 2016). The rational speech act framework builds on earlier related work of Zipf (1949) and Grice (1975) in arguing that an act of communication can be conceived of as a form of Bayesian inference in which a message recipient tries to maximize the probability of inferring the true state of the world (or, more generally, the intended referent of the speaker, or, more precisely, the receiver’s subjective certainty about the speaker’s intended referent), given some transmitted message. As discussed in Franke and Jäger (2016), this probability calculation may be influenced by many things, including the probability that the message might be incorrectly encoded (the speaker makes a

mistake) or incorrectly decoded (the listener misunderstands). Subjective adjectives such as *beautiful* or *delicious* introduce complications to the process of decoding a message. It is always possible that a message receiver does not share the subjective view of the message transmitter, in which case the subjective adjective will be unhelpful and perhaps even misleading to the receiver. Using Monte Carlo simulations, Franke et al. (2019) demonstrated that placing a subjective adjective earlier than an objective adjective was associated with a higher probability of correctly identifying a referent. Of the several reasons this is true, the main one is that “placing more subjective adjectives farther from the modified noun means that they modify a smaller context, which means that there are fewer opportunities for the listener’s subjective representation to deviate from the speaker’s” (p. 5). The subjective adjective is constrained by the local context in which it is uttered, making it easier for the listener to appropriately delimit its scope. Franke et al. (2019) demonstrated this by focusing narrowly on the phrase *the big brown bag*, in which the subjective adjective *big* can be interpreted in context to mean *big for bags*, as opposed, say, to *big for thimbles* or *big for airplanes*. One limitation of this line of reasoning is that it does not directly address how adjectives are ordered in the absence of any context. It may be that subjective ordering is carried through to uncontextualized phrases simply because language users have learned it is effective in contextualized cases. Another limitation is that this line of reasoning cannot address how two subjective adjectives are ordered. Why do we say *the big fat man*, but not *the fat big man*, when both bigness and fatness may be subjective?

Malouf (2000) considered several computational approaches to predicting adjective order. All of them involved extrapolating an observed order in a large subset of a corpus to predict adjective order in a smaller subset of that corpus. Although some of the methods performed very well (the best was accurate 89.73% of the time, and combining methods achieved a correct adjective ordering rate of 91.85%), these methods are dissatisfying for the same reason that Dixon’s (1977) ordering was: They use a *description* of the phenomena to be explained (albeit in Malouf’s case, a quantitative description) as its own explanation. Essentially, such explanations amount to saying that adjectives are ordered as they are because that is how adjectives are ordered. Although it is possible that there may be nothing more to adjective order than acquired habit, as psychologists we would prefer to address the question of *why* adjectives are ordered as they are.

Wulff (2003) undertook a multifactorial corpus study of adjective order. She used linear discriminant analysis to model the order of 3,136 adjective–adjective pairs from the British National Corpus (BNC), by trying to predict an adjective’s position (first or second). Using a large set of predictors such as word length and frequency, percentage of BNC occurrences tagged as an adjective, observed co-occurrences of adjectives,

membership in Dixon’s semantic classes, and others, she achieved a cross-validated correct adjective order classification rate of 63.2%, far above chance. Many of Wulff’s predictors were rather laborious (or, for a subset of words, impossible) to operationalize, had low (or, at least, indeterminate) psychological plausibility, and/or are theoretically opaque. Moreover, some of them have a circular relationship to the phenomenon being modeled. As noted above, Dixon’s semantic classes were presented as a *description* of how adjectives are ordered, so using them as a predictor of adjective order is problematic.

In recent years there have been several studies of adjective order that used information-theoretic principles. Hahn, Degen, Goodman, Jurafsky, and Futrell (2018) and Futrell (2019) followed up on the idea that adjectives are ordered in order of increasing specificity. They formalized specificity using a commonly used measure of collocation strength, *pointwise mutual information* (PMI):

PMI(Adjective, Noun)

$$= \log(p(\text{Adjective} \mid \text{Noun})) - \log(p(\text{Adjective})).$$

PMI measures whether a given adjective modifies a noun more often than would be expected by chance based on their individual probabilities. By computing PMI over nouns and the adjectives that modify them, it is possible to quantify the degree to which a specific adjective modifies a specific noun, whether by direct adjacency or not. A higher PMI value is associated with a higher likelihood that the adjective modifies the noun. If the adjective and the noun only occur together, then $\log(p(\text{Adjective} \mid \text{Noun})) = \log(1)$ is zero, and the PMI is $-\log(p(\text{Adjective}))$, a positive number since $p(\text{Adjective})$ is always less than 1. When the adjective occurs as often with the noun as it occurs overall (i.e., when the probabilities of the noun and adjective are independent), then $\log(p(\text{Adjective} \mid \text{Noun})) = \log(p(\text{Adjective}))$, so the PMI is zero. When the adjective is less likely to occur with the noun than it occurs without the noun, then $\log(p(\text{Adjective} \mid \text{Noun})) < \log(p(\text{Adjective}))$, so the PMI is negative.

Hahn et al. (2018) used PMI in a logistic mixed-effects model to predict the order of adjective pairs. Both PMI and subjectivity measures independently entered into the model, although the authors did not report the accuracy rate of the model. They followed this up with a more complex model that incorporated subjectivity scores and required simulation of the cost to a listener of updating her beliefs based on a communication from a speaker. This was estimated using the *surprisal* of an adjective–adjective–noun triplet: $-\log(p(\text{Adjective}_1 \mid \text{Adjective}_2 \mid \text{Noun}))$. The model was used to simulate communication between a single speaker and a single listener. It replicated the results discussed above by demonstrating that placing subjective adjectives earlier had a higher utility than placing them later. When the model was applied to a larger data set, it predicted adjective order with an accuracy of 93.7%. However, using only the subjectivity scores

performed almost as well (93.3% correct), suggesting that most of success of the model is attributable to the incorporation of subjectivity scores. A hyperparameterized version of the model achieved a cross-validated accuracy of 96.2%.

PMI is confounded with the number of followers of an adjective. Adjectives that are followed by a small number of nouns must on average have a larger PMI than adjectives that are followed by more nouns, since uncommon adjectives have fewer opportunities to occur with nouns than common adjectives. This may be a feature rather than a bug when it comes to explaining adjective order, because it captures the fact that adjectives that are followed by a small number of nouns often have a narrow meaning (and therefore must go close to the noun, as discussed above), which is precisely why they modify few nouns. Many adjectives that are followed by few nouns are fixed forms from technical contexts (e.g., *aplastic anemia* [PMI = 5.6], *after-school program* [PMI = 2.6], *chi-square tests* [PMI = 2.6], *Craftsman-style house* [PMI = 2.4], *homosporous pteridophytes* [PMI = 3.9]).

Another problem with PMI, as noted by Dyer (2018), is that nonce adjectives are ordered consistently despite the fact that they have an incomputable PMI because they have a zero probability. The same problem applies to using PMI to predict the order of real adjectives from unexperienced adjective–noun pairs. For example, the COCA does not contain a single instance of the noun *aardvark* that is preceded by an adjective, so the relevant PMIs needed to decide on the ordering of the phrase *big fat aardvark* are not directly computable from that corpus (though they might be computed by some form of statistical smoothing, as in Roberts & Chater, 2008). These considerations suggest that PMI is unlikely to be the causal factor in adjective ordering.

After rejecting PMI for because of its problem with ordering nonce adjectives, Dyer (2018) proposed a different information theoretic explanation of word order that applied more generally than just to adjectives (see also Futrell, Levy, and Gibson, 2017; Liu, Xu, and Liang, 2017). The idea depends on measuring the complexity of integrating any word with the words that it modifies. Dyer proposed that the integration complexity of an adjective could be estimated by the Shannon entropy (Shannon, 1948) of the probability distribution of the syntactic categories of that word's modifiers. Shannon entropy $H(X)$ is defined across a series of elements X_i in X as:

$$H(X) = - \sum_{i=1}^n p(X_i) \log_2 p(X_i).$$

If a word has 10 possible modifiers, of which three are adjectives, three are verbs, two are adverbs, and one is a preposition, then Dyer's proposed measure of the Shannon entropy of the probability distribution of the syntactic categories of the head word could be calculated using the formula above as

$-(0.3 \times \log_2 (0.3) + 0.3 \times \log_2 (0.3) + 0.2 \times \log_2 (0.2) + 0.1 \times \log_2 (0.1)) = 1.84$ bits. If all 10 modifiers belong to the same syntactic category, the Shannon entropy of the probability distribution of the syntactic categories would be simply $1 \times \log_2 (1) = 0$, capturing mathematically that there is no uncertainty about the syntactic category of the modifier. Dyer computed the syntactic category entropy of the adjectives *happy* (1.78 bits), *Italian* (0.76 bits), and *chemical* (0.5 bits). He argues that lower entropy words (which are easier to integrate due to their decreased uncertainty) should be placed closer to the noun, suggesting we should prefer *the happy Italian woman* to *the Italian happy woman*, as we do. He considered a few other adjective ordering examples that also fit the theory (*beautiful small black bag*; *three big houses*), but did not report an overall accuracy for adjective ordering. The idea of syntactic integration complexity is appealing because it has a theoretical motivation that suggests why adjectives are ordered as they are and because it is not limited only to English adjective ordering but applies more widely to other words and (as Dyer demonstrates) to other languages. However, it relies on having access to the full set of a word's dependency trees, a requirement that seems psychologically implausible on the face of it and that makes the measure difficult to compute (though tree annotations do exist for corpora in many languages; Nivre et al., 2016).

Dye, Milin, Futrell, and Ramscar (2018) specifically discussed the relationship between adjective order and Shannon entropy, though without directly testing how well entropy could predict adjective order.

Ideally, a theory of adjective ordering should explain *why* adjectives are ordered as they are, rather than using their observed patterns of use as an explanation. Here, I propose an information-theoretic model that does that has a strong theoretical basis and high psychological plausibility, and that uses just two theoretically grounded predictors.

A new theoretical approach to adjective order

The theoretical basis of the model is the concept of *likely need*, which was originally proposed in the context of memory research (Anderson, 1991; Anderson & Milson, 1989; Anderson & Schooler, 1991; Anderson and Tweney, 1997). Anderson and Milson (1989) argued for an evolutionary adaptive perspective on memory, based on an analogy to human-engineered (therefore, perfectly understood) information retrieval systems. In engineered systems, there is a time and energy cost for retrieving an irrelevant item, just as there may be metabolic, time, lost opportunity, and other costs when an animal retrieves an irrelevant memory. Given such costs, it is rational to retrieve items in order of their likelihood of being needed in any particular situation (following a closely related suggestion in Zipf, 1949). A retrieval cost places a

lower limit on retrieval probabilities since it is also rational that a biological or man-made retrieval system should stop trying to retrieve an item when the probability of retrieving it is low enough that the expected gain from retrieving it is less than the cost of retrieving it. Anderson and Milson (1989) considered a number of factors that contributed to the retrieval probability of an item (which they called the item's *need probability*), including *time since last retrieval*, *the average frequency with which an item is retrieved*, *the strength of the retrieval cues*, and *the number of different contexts in which the item appears*. They presented a model based on retrieving items in order of their need probability given a set of contextual cues that successfully reproduced many well-known phenomena in memory research.

The idea of need probability has been imported into the study of language processing (Adelman, Brown, & Quesada, 2006; Jones, Johns, & Recchia, 2012; Hollis, 2017), where it has been suggested that the large lexical access facilitation effects usually attributed to word frequency might be more reasonably attributed to need probability. Word frequency is just one component of this, though it may be the most important one when there are no contextual cues, as in the single word decision tasks that are widely used in psycholinguistic experiments.

Need probability is closely related to Shannon entropy (Shannon, 1948). Words with high need probability provide less reduction in uncertainty than words with low need probability. Intuitively, this is clear if we take as an example an obviously high need probability word such as *the*. Since the definite article can be followed by many words (it has 104,333 distinct followers in the COCA), it is difficult to guess that its following word will be. By contrast, if we consider the low need probability adjective *tillable*, almost all words have a zero probability of following, and very few are likely. Not many things are tillable. The word has only two distinct followers in the COCA, which are (as the reader might guess) *soil* and *land*. The word's occurrence hugely reduces uncertainty about the following word (see discussion in Hale, 2003).

Figure 1 shows the relationship between the log₂ number of words that follow an adjective, and its log₂ rank in terms of numbers of following words. There is a very strong linear fit ($r = -.97$, $p < 2e-16$) over the 51% (2,048/4,002) of adjectives with the most followers, suggesting that the distribution of adjectives is largely optimized informationally, except when there are few followers. Danks and Glucksberg (1971) argued long ago that pragmatics could explain adjective order since it is communicatively optimal to order adjectives in terms of increasingly fine discriminations, from least to most informative.

Hollis (2017) used a *predict model of language* to estimate the average need probability of a word. In his work on modeling memory, Anderson (1991) had suggested that the degree of similarity between perceived cues (which serve as memory probes) and the features of a remembered item (the memory

Fig. 1 Log₂ number of words following an adjective, as a function of Log₂ rank order. The linear fits shown are fitted to the adjectives with Log₂(rank) < 11 (indicated with a shading change), which includes the 51.1% of adjectives with the most followers

trace) is proportional to the need probability of that remembered item. For example, we are highly likely to recall a person when that person is standing right in front of us, because the perceived cues (the person's appearance) resemble our stored memories of that person's appearance. Hollis (2017) pointed out that Anderson's observation translated naturally to distributional semantic models. In predict models (Mikolov, Chen, Corrado, & Dean, 2013), individual words are represented by vectors (by convention, of length 300) of the weights on the hidden units of a simple three-layer neural network used to predict a target word's context, or to predict a target word from its context. Those vectors can be naturally conceived of as cues to linguistic context, similar to the facial features that served as cues in the example above. If I ask you to tell me what word comes to mind after I say *purr*; *pet*, *fur*, you are likely to think of the word *cat*, because the contextual cues (the first three words) are related to the context of the word you think of. If we take the average vector of the three cue words, the vector most similar to that vector that is not one of its component words is *cat*. Hollis (2017) noted that this observation can generalize because "Words with high cosine similarity to other words *on average* should be precisely the words that are needed across a broad range of contexts" (p. 1354; emphasis added). The word we are most likely to need on average is that word whose context is most like the average linguistic context. Hollis averaged together cosine distances of a word's vector from 1,000 frequency-weighted random words to estimate that word's *likely need*. Concretely, we are more likely to find ourselves needing the word *interesting*

than the word *ontological*, since the word *interesting* has wide and general application (occurs in many different contexts), while the word *ontological* is highly specific (occurs in a small number of specialized contexts). This can be captured quantitatively if we average together the vectors of the most common 500 words in the dictionary and compute the cosine distance of the words from that average vector. The word *interesting* is the 2,789th closest word to that likely word contexts vector, while the word *ontological* is the 69,476th closest word, in a dictionary of 78,278 words—that is, the context of the word *interesting* is more similar to the general context of words than the context of the word *ontological* is. Hollis showed that his measure of average need probability was linearly related to lexical access times and to logged values of both word frequency and *context diversity*, the number of different documents within which a word appears.

Following Westbury and Hollis (2018), in this paper we apply this idea of need probability to the word class of focal interest here, adjectives. Westbury and Hollis (2018) showed that distance from the averaged vector of a large number of words from the same part of speech (e.g., nouns, verbs, adverbs, or adjectives) was a good predictor of membership in that word class. They called such averaged vectors *category-defining vectors* (CDVs). For example, the 1000 words whose vectors are most similar to a CDV defined by averaging together the vectors of 100 randomly-selected verbs are themselves all verbs. Westbury and Hollis (2018) also tested human sensitivity to distance from CDVs, using a series of go/no-go experiments. Participants were asked to hit a button if a word belonged to a particular word class, or else to do nothing. Distance from a word class's CDV was predictive both of accuracy and correct-decision speed in making such decisions.

Distance from word class CDVs can be conceived of as a measure of average need for that word class, because increasing distance from the CDV is associated with systematic decreases along two of the dimensions that Anderson and Milson (1989) explicitly identified as part of need probability: *the number of different contexts in which the item appears* and *the average frequency with which an item is retrieved* (which is obviously related to, though not perfectly correlated with, another explicitly identified aspect of need probability, *time since last retrieval*). To demonstrate this in the present context, a CDV for the class of adjectives was constructed by averaging together the vectors of the 1,000 most common adjectives, using a word2vec skip-gram matrix (<https://code.google.com/archive/p/word2vec/>) defined over Google's multi-billion-word news corpus (available from that word2vec website).

Number of contexts

In language, the number of contexts in which a word appears can be estimated by two measures: *degree of synonymy* and

generality of context. Degree of synonymy is important for adjective order because many adjectives have very similar meanings. The idea of *goodness*, for example, can be expressed with a wide variety of words. The raw frequency of the word *good* is not a useful measure of the need probability of *the concept of goodness*, because English speakers have many alternate words they can use to express roughly the same concept: *cool*, *excellent*, *superb*, *great*, and many others. If adjective order depends on the need probability of an adjective's *meaning*, we want a measure of need probability that takes synonymy into account.

The relationships between the 282 most strongly vectors correlated ($r \geq 0.61$ with at least one other word) of the 500 closest words to this adjective CDV are shown in Fig. 2. Along with many smaller clusters, there are several clearly discernible large clusters of synonymous or near-synonymous words, including clusters (marked on Fig. 2) relating to *preposterousness* (e.g., *ridiculous*, *ludicrous*, *laughable*, *farfetched*, *nonsensical*), *shamefulness* (e.g., *inexcusable*, *unforgivable*, *deplorable*, *scandalous*, *reprehensible*), *terribleness* (e.g., *atrocious*, *dreadful*, *horrendous*, *horrible*, *bad*), and *excellence* (e.g., *terrific*, *awesome*, *phenomenal*, *exceptional*, *great*).

To quantitatively demonstrate the decreasing synonymy with increasing distance from the CDV, I computed the average pairwise correlation of the vectors of all words in each block of 100 words, from closest to most distant to the CDV. As shown in Fig. 3, the average pairwise correlation shows a reliable decrease in each block, from the first until the 14th (i.e., across the vectors of the closest 1,400 words to the CDV). In particular, the block consisting of the closest 100 words has much higher average correlation (average [SD] $r = 0.25$ [0.12]) between the vectors of its component words than the average correlation between the vectors of the next 100 words (average [SD] $r = 0.20$ [0.12]; $t(19175) = 32.7$, $p < 2.2e-16$).

It is clear from inspection that the main clusters of words in Fig. 2 (as well as many of the smaller clusters) are also composed of words that have extremely general applicability. Many different kinds of things can be *ludicrous*, *deplorable*, *horrible*, or *awesome*. To quantify the decreasing applicability (increasing specificity) of adjectives as they are further from the CDV, I computed the number of distinct words that follow each of the first 1,600 words in the CDV in the 560 million word Corpus of Contemporary American English (COCA; corpus.byu.edu/coca/). To normalize the measure, I took the ratio of average $\log_2(\# \text{ of followers})/\text{average } \log_2(\text{COCA frequency of occurrence})$. This measure is also negatively correlated with distance from the CDV (see Fig. 4; $r = -.16$, 95% CI $[-0.21, -0.11]$, $p = 7.3e-09$). Adjectives closer to the CDV tend to be followed by a wider range of different words than adjectives further from the CDV, after controlling for word frequency.

Fig. 2 The correlation relationship among the 282 words with vectors that correlated at $r \geq .61$ with the vector of at least one other word among the 500 closest words to the adjective category-defining vector (CDV).

The threshold was selected by human judgment to maximize interpretability of the image. Large gray texts are human annotations labeling some clusters. Distance between disconnected clusters is arbitrary

Average frequency

Another aspect of need probability explicitly specified by Anderson and Milson (1989) was *the average frequency with*

Fig. 3 Average correlation of all word vectors within each block of 100 words, as a function of distance from the adjective category-defining vector (CDV). All differences are reliable by t test ($p < .05$), except the final shown difference, between Blocks 14 and 15

which an item is retrieved. In language, this can be estimated (independently of any context) by a word's frequency. As shown in Fig. 5, there is a significant negative correlation between average $\log_2$ (COCA frequency) and distance of that word from the CDV ($r = -.21$, 95% CI $[-0.25, -0.16]$, $p = 4.77 \times 10^{-16}$). More frequent words are closer to the CDV.

Based on these observations that words closer to the adjective CDV have more synonyms, more general meanings, and higher frequency than words further from that CDV (i.e., that distance from the adjective CDV is correlated with *the number of different contexts in which the item appears* and with *the average frequency with which an item is retrieved*), and given that the CDV is literally defined as the average vector of the word category, it is reasonable to consider that distance from a word category CDV is a proxy for likely need within that category. Having such a measure allows us to test the likely need theory of adjective ordering.

Computational validation

Stimuli

I began by extracting all 34,376 of the adjective–adjective–noun triplets that appeared at least three times in the part-of-speech-

Fig. 4 Logged number of distinct followers per logged occurrence, as a function of distance from the adjective category-defining vector (CDV), for words that occurred at least 100 times in the Corpus of Contemporary American English.

tagged COCA, which includes text from spoken English, fiction, magazines, newspapers, and academic texts. I discarded the nouns and collapsed the adjective pairs, to end up with 18,025 distinct adjective pairs that had appeared before a noun, containing only words in our co-occurrence matrix (described below). Of these, 10 were discarded because they consisted of the same adjective repeated twice (e.g., *new new* or *great great*), rendering issues of adjective order irrelevant.

A problem with assessing the performance of any proposed prenominal adjective ordering mechanism is that it is not

obvious what the optimal hit rate should be (an issue I return to in the second study below). Although 18,015 distinct pairs occur in the COCA, that does not necessarily mean that all those pairs are in their canonical order (see discussion in Vecchi, Zamparelli, & Baroni, 2013). Some adjective pairs may have no canonical order, being equally acceptable in either order. Other pairs may be ordered as they are for reasons that are largely irrelevant to our focal interest. For example, as noted above, some adjective pairs include an adjectival modifier in the first position (e.g., *other*, *only*, *actual*), which

Fig. 5 Logged Corpus of Contemporary American English frequency, as a function of distance from the adjective category-defining vector (CDV)

almost always has to go first for syntactic reasons. Some adjective pairs are fixed forms or part of a larger fixed form (i.e., *saturated fat*, *gross national*, *bitter cold*). In particular, it has been previously suggested that the words *young*, *little*, and *old* should not be considered in studies of adjective order because “these three adjectives tend to form a sort of petrified compound with certain nouns” (Vendler, 1968, p. 132; see also Goyvaerts, 1968; Hetzron 1978; Martin & Ferb, 1973; Sweet, 1898/1955; Wulff, 2003). I will discuss these particular words further in the conclusion of this paper. Finally, it is acceptable to reverse the canonical adjective order in situations where there is a deliberate intention of emphasis. We would normally say *the old brown dog*, but if we were trying to distinguish between two old dogs of different colors, we might specify that we were referring to *the brown old dog* without raising any eyebrows about the noncanonical adjective order. Although some of these complexities might be dealt with by hand-choosing pairs to reject, others of them are difficult or impossible to adjudicate. Rather than trying to hand-select the “right” pairs to eliminate, I chose to model all adjective pairs in the COCA. I discuss below a data-based method of estimating human reliability in judging adjective order. Stronger empirical evidence for that reliability is presented in Study 2.

Method

To put the model developed below into a wider perspective, let us first consider how some related individual predictors are related to adjective order. The number of words following the second adjective was less than the number of words following the first adjective 55.8% of the time, suggesting this measure is itself a weak but statistically significant cue to word order (exact binomial $p < 2e-16$). The frequency of the first word was greater than the frequency of the second 52.1% of the time (exact binomial $p = 6.25e-09$). Given the great variance in word frequency, a more reasonable measure than the raw number of followers is one that controls of word frequency. I divided the number of a word’s followers by the word’s frequency, to get a measure of followers per occurrence. The number of followers per occurrence for the second adjective was less than the number of followers per occurrence for the first adjective 57.9% of the time (exact binomial $p < 2e-16$).

These systematic differences in the expected directions lend some support to the hypothesis that adjective order reflects *likely need*. To test the hypothesis adjective order more rigorously, I modeled it using just two predictors, distance from the adjective CDV and logged word frequency. Logged frequency was included as a separate predictor both because it is only weakly correlated with distance from the CDV, because it is related to likely need by definition, and because including it does improve the model substantially. To build and validate our model, I randomly split the adjective pairs

into three sets of 6,005 pairs each. In one set, I flipped the predictors so that the predictors for the second adjective were treated as predictors of the first adjective and vice versa. I put this flipped set together with a second set, so that binomial regression could be used to predict whether an adjective pair was in its observed order (1) or its flipped order (0). The third set of 6,005 pairs was retained to cross-validate the model.

Results

The binomial regression model developed on two-thirds of the data is presented in Table 1. Among the 12,010 pairs in the development set, the model was able to correctly classify whether an adjective pair was flipped or not with an accuracy of 72.1% (unflipped: 72.1%; flipped: 72.1%; Exact $p < 2e-16$). It correctly accepted the unflipped order of the 6,005 adjective pairs in the validation set 72.4% of the time (exact $p < 2e-16$).

A better cross-validation set is composed of prenominal adjective pairs specifically attested to be in the correct order. Since almost everyone who has written about the topic has provided such examples, I was able to collect 229 published examples (from Byrne, 1979; Danks & Glucksberg, 1971; Huang & Federmeier, 2012; Kemmerer, 2000; Kemmerer et al., 2009; Kemmerer, Weber-Fox, Price, Zdanczyk, & Way, 2007; Martin, 1969; Scheffelin, 1971; Sichelschmidt, 1986; Sproat & Shih, 1991; Sussex, 1974; Vandekerckhove, Sandra, & Daelemans, 2013; Vendler, 1963; Wulff, 2003). I eliminated the 17 pairs that included a determiner as the first adjective, leaving 212 attested pairs. The logistic regression equation correctly accepted 77.7% of these. Among the 146 pairs that occurred at least once in the COCA, the correct acceptance rate was 82.5%.

It is possible to obtain a better estimate of adjective pair frequency from the Google Ngram corpus (<https://books.google.com/ngrams>), which contains 465.5 billion words published in books between 1800 and 2012. I used the R package *ngramr* (<https://github.com/seancarmody/ngramr>) to compute the frequencies in that corpus of all the attested adjective pairs, as well as the frequencies of the same words in the reverse order. The results underscore the difficulty of achieving a perfect classification of adjective order by providing us with a plausible upper bound for human classification. Among the 212 prenominal pairs specifically listed by scholars as examples of the correct order, only 83.9% appeared more often in the attested order than in the opposite order, providing suggestive evidence that this level of accuracy may be the closest to a “gold standard score” that could be achieved by a model (see Discussion section of Study 2). The logistic regression equation performed equally well on both sets, correctly accepting 79.2% of the 178 pairs that appeared more often in the attested order, compared with 71.4% of 34 pairs that appeared more often in the opposite order ($\chi^2 = 1.23$, $p > .05$).

Table 1 Model development (top) and best model (m2, bottom) for predicting whether two adjectives are in their observed order or not, with chance at 0.5

Model ID	Model specification	AIC
m0	LogFreq1 + LogFreq2	16581
m1	m0 + CDVDistance1 + CDVDistance2	13433
m2	LogFreq1 * CDVDistance1 + LogFreq2 * CDVDistance2	13349

Predictor	Estimate	SE	z	p
(Intercept)	0.13	0.69	0.18	0.86
CDVDistance1	−4.96	0.56	−8.93	<2E-16
CDVDistance2	4.75	0.54	8.88	<2E-16
LogFreq2	−0.25	0.05	−4.74	2.16E-06
LogFreq1	0.24	0.05	4.37	1.25E-05
LogFreq2:CDVDistance2	0.42	0.06	6.80	1.05E-11
LogFreq1:CDVDistance1	−0.40	0.06	−6.22	5.08E-10

Discussion

The cross-validated classification rate of 72.4% on 6,005 randomly selected pairs is much better than the cross-validated accuracy of 63.2% achieved by Wulff (2003) on 3,136 selected pairs modeling a closely related decision with a larger set of sometimes circular predictors ($\chi^2 = 81.96$, $p < 2e-16$). More importantly than the improved cross-validated accuracy is the fact that this accuracy was achieved using a very simple linear model that has a priori plausibility and is tied to the measure of likely need that has previously been shown to have psychological relevance. Westbury and Hollis (2018) showed that access time to a word was strongly predicted by the distance of that word to its CDV, and it is well known that more frequent words are accessed more quickly.

Experimental validation

To further validate the likely need theory of adjective order, an experiment was conducted to measure how well human judgments of adjective order were predicted by likely need, as defined above. The experiment was reviewed ethically by the University of Alberta research ethics board.

Method

Participants

Participants were 52 self-reported native English speakers (32 female, 20 male; 45 self-reported right-handers, seven self-reported left-handers) who participated after giving informed consent in return for partial course credit. They had an average [*SD*] age of 19.6 [3.2] years and an average [*SD*] of 13.6 [1.9] years of education. Though all attested to having learned

English as a native speaker before the age of 5, half [26] of the subjects reported speaking another language as well as English.

Stimuli

The stimuli were drawn from the adjective–adjective–noun pairs that appeared in the COCA. I computed the likely need measure described above for all the adjective pairs. After ordering them from least to most likely to be correctly ordered according to that measure, I took every 67th phrase, or the next nearest phrase if the chosen one included a determiner or an obvious fixed form. This provided 500 phrases that spanned the range of order probability, from a low estimated probability of being correctly ordered (e.g., *chopped bittersweet chocolate* [2.6%], *minced fresh basil* [7.6%], *old bad thing* [8.1%]) to a high estimated probability of being correctly ordered (e.g., *strong red wine* [96.2%], *small nonstick saucepan* [96.4%], *serious gray eyes* [97.4%]).

Procedure

Each participant was shown the following instructions:

“We will show you two phrases, one above and one below. They are the same except for the order of the adjectives. We ask you to let us know which order sounds better to you. If you think the one on the top is better, please press the ‘k’ key. If you think the one below is better, press the ‘m’ key. Note that the ‘k’ key is above the ‘m’ key on the keyboard, so you will press the top key when the phrase on top is better, and the bottom key when the phrase on the bottom is better.

Sometimes you may feel that both orders are equally acceptable. In that case you can press either key randomly.”

Every participant made decisions about 250 phrases in two blocks, with a self-timed break after the first 125 decisions. Phrases were randomly selected from the 500 for each participant and presented in random order. The position of the phrases was controlled so that the COCA-attested order was seen equally often in the top and bottom position.

After seeing the instructions, every participant saw an example pair of phrases, *big fat dog* above *fat big dog*, while a research assistant was in the room. Neither of these phrases was among the 500 phrases selected as stimuli. Participants were asked if they had any questions before continuing. When they were ready to continue, the research assistant left the room, and participants completed the rest of the trials themselves.

Data were collected in three small testing rooms using custom-written software running on three identical Apple G4 Macintosh Minis running Apple OS 10.15, attached to identical 17.1-in. monitors. The screens' resolutions were set to $1,280 \times 1,024$ pixels. The phrases were shown in black 75-point Times New Roman font, centred on the screen 30% and 50% down from the top, against a white background.

Results

The reaction times were highly variable (average [*SD*]: 3,695 [3,022] ms). Logged reaction times more than 3 standard deviations from the mean logged RT were removed. This resulted in the removal of 32 trials (0.25% of all data) with an average [*SD*] RT of 364.7 [112.00] ms and a maximum time of 557 ms, which is probably too short for participants to have had time to read the two phrases and reach a decision about them. The remaining data were analyzed using linear mixed-effects models, using the Akaike information criterion (AIC; Akaike, 1974) to adjudicate between models.

The most important result is the relationship between participant decisions about adjective order and the likely need estimate of adjective order described above. A base model to predict whether humans accepted the most probable order was defined using only participants with random intercepts (AIC: 16,863). Adding stimulus order as a random effect hardly changed the AIC value (AIC: 16,861), so was not included. Adding the estimated probability of the most likely order (i.e., using the estimates derived from the model in Table 1) decreased the AIC value by 933 to 15,930, confirming that these estimates are strong predictors of adjective order. The final simple model is shown in Table 2. The effect is illustrated graphically in Fig. 6, which shows that there is a strong linear relationship between the estimated probability of the most likely adjective order and the likelihood that a human judge would select that order as the best order. Across

Table 2 Fixed effects from best model for predicting human judgments of adjective order. See Fig. 6

Predictor	Estimate	SE	<i>t</i>	<i>p</i>
(Intercept)	0.74	0.025	29.87	<2e-16
Standardized likely need estimate	0.59	0.020	29.24	<2e-16

all 500 pairs, the model's binarized estimate of order probability (1 if > 0.5 ; 0 otherwise) agreed with the binarized majority human decision (1 if accepted by $> 50\%$ of humans; 0 otherwise) 70.2% of the time (exact binomial $p < 2e-16$).

To compare this to a PMI-based estimate, the PMI of each of the two adjectives with their noun was computed, over all prenominal adjectives up to three. The adjective that is more likely to go close to the noun should have a higher PMI. This model achieved an average agreement with human ratings of 64.6% (exact binomial $p < 2e-16$). By chi-square test, this was no worse than the likely need model ($\chi^2 = 2.3$, $p = 0.092$).

The PMI and the likely need models are compared in Fig. 7, which shows the relation between the human acceptance rate and the models' acceptance rate, collapsed into 20 bins of 25 adjectives across the range of human acceptance. The PMI model was better than the 17/25 chance level ($p = .022$) for 6/25 bins (24%), all of which had human agreement $> 80\%$. It agreed 60% of the time with the only bin that had unanimous acceptance of one order from humans. The likely need model was better than the 17/25 chance level for 9/25 bins (36%), all of which had human agreement $> 80\%$. It agreed 92% of the time with the only bin that had unanimous acceptance of one order from humans.

Although the instructions in the experiment did not mention decision speed, analysis of the RTs may add further strength to any conclusions about the validity of the likely need model. It can be hypothesized that participants should find decisions with a strongly predicted order easier to make than decisions with a weakly predicted order. To analyze the trial RTs, I divided them by the number of characters in the two phrases seen in that trial, to get a measure of RT per character. Both participant and stimulus order contributed as random effects, as measured by decreasing the AIC value by at least 5 points. Adding the estimated probability of order decreased the AIC value by 80 from 152,300 to 152,220, confirming that these estimates are strong predictors of decision time. The final fixed effect model is shown in Table 3. The effect is shown graphically in Fig. 8. As hypothesized, RTs are shorter for adjective orders that have a higher estimated probability from the likely need model.

Discussion

These results lend additional support to the computational model developed in Study 1, and the psychological reality of

Fig. 6 Estimated percentage of adjectives chosen by experimental participants in their most probable order, as a function of the computed likely need probability, fitted with GAMs. 95% confidence intervals are present, but nearly invisible. The model is presented in Table 2

the measure it is based on. The estimates derived from likely need are strong predictors of both adjective order decisions and adjective decision time.

Fig. 7 Point-wise mutual information (PMI) model and likely need model compared with human acceptance rates, for the 500 adjective pairs in the adjective order decision experiment, collapsed into 20 bins containing 25 adjective pairs each. The light dotted lines indicate chance levels ($p \geq .02$)

These judgment data allow us to provide a more definitive answer to the question raised above: What is the “gold standard” hit rate for adjective order? The experimental data confirm the existence of the phenomenon under study: There are indeed adjective pairs that have very strong order constraints (see Fig. 9). On average, the human judges agreed about the optimal adjective order for 81.5% of the 500 adjective pairs. However, only 35 (7.0%) of the 500 attested adjective pairs were judged unanimously in a single direction (e.g., *lovely little face*, *tall green grass*, *small urban college*). Only 181 (36.2%) were judged in one direction at least 90% of the time. Forty-nine adjective pairs (9.8%) were indeterminate, defined as having a judged probability no higher than 60% of going in either order (e.g., *thin crispy crust*, *white sandy beach*, *creamed chipped beef*). These data suggest that the implication sometimes found in the literature that all English prenominal adjectives are strictly ordered is not true.

Table 3 Best model for predicting RTs per character for human judgments of adjective order. See Fig. 8

	Estimate	SE	<i>t</i>	<i>p</i>
(Intercept)	98.03	4.28	22.92	<2E-16
Standardized likely need estimate	-5.68	0.66	-8.60	<2E-16

Fig. 8 Adjective order decision time per character as a function of the likely need estimated probability of that order, fitted with GAMs, with shaded 95% confidence intervals. The model is presented in Table 3

We can use these data to better estimate the success of the likely need model. As noted above, overall, the likely need binary estimate of the order agreed with the majority human decision 70.2% of the time. Limiting the data to only the 177 adjective pairs with a strong human ordering preference

($\geq 90\%$ in favor of one order), the likely need model agreed with the judged order 95.1% of the time. The largest errors among those 177 were for the phrases *thick white paint* (accepted by human judges in that order 96.2% of the time but predicted to have only a 50.08% chance of going in that order); *lovely little face* (accepted by human judges in that order 100% of the time, but predicted to have only a 54.35% chance of going in that order); *raw judicial power* (accepted: 95.65%; predicted: 50.30%); *silky blond hair* (accepted: 94.59%; predicted: 51.97%) and *tall young man* (accepted: 96.15%; predicted: 55.64%).

Fig. 9 Histogram of observed human agreement about adjective order, for 500 adjective pairs

General discussion

As discussed above, several researchers have previously suggested that adjective order may reflect information-theoretic constraints. Information theoretic explanations and the explanation offered here in terms of likely need are not opposition to each other as explanatory mechanisms for prenominal adjective order. The two explanations are closely related and can be made arbitrarily closer by changing how we define a word's context in computing its entropy. Except in the very simplest (and most unnatural) cases, measures of a complex

system's entropy are never pure measures of *the system itself*, but rather measures of *a model of that system*. The simplest case makes this clear. Shannon (1948) famously proved that the entropy of a messaging system consisting of N elements was $\log_2(N)$. However, that measure is based on the assumptions that the N elements are equiprobable and uncorrelated with each other. To the extent that either of these assumptions is wrong (as Shannon himself discussed), it is possible to take advantage of that departure from perfect randomness to define a representation of the system of N elements that has an entropy less than $\log_2(N)$. In complex systems like the linguistic system, which has many correlated elements at different levels of description, it is not obvious what is the “true” information-theoretic representation. To know that we would need a full understanding of every correlated linguistic element to which humans are sensitive.

One apparent practical problem with a simple information theoretic analysis of adjective order is that Shannon entropy is usually computed over specific strings. Martin and Ferb's (1973) observation that newly encountered adjectives are ordered correctly with other adjectives according to their *meaning* might seem to suggest that it is not the simple information value of *strings* that matters in adjective ordering, but the information value of their *meanings*—that is, it is not a question of how often a particular *string* is followed (or otherwise contextualized) by other strings, but rather a question of how often *words with similar meanings* are followed by different strings. There are semantic correlations. Simple local-context information-theoretic representations cannot explain how newly encountered adjectives can be ordered since a string encountered for the first time has no computable entropy from local co-occurrence.

It may not be obvious how to compute the Shannon entropy of word *meanings* (though, see the Discussion section). Consider, for example, the words *difficult*, *dubious*, *terrible*, and *appalling*. These all have very similar meanings whose wide applicability as modifiers guarantees they will have low entropy. Their role in prenominal adjective order should therefore be similar (i.e., they should all be far from the noun). However, their information theoretic value based on their immediate follower context is wildly divergent. The word *difficult* occurs 54,751 times in the COCA, and is followed by 766 different words. Guessing the following word given the occurrence of the word *difficult* therefore has a 0.13% ($1/766$) chance of being correct (if we make the simplification of ignoring frequency information to treat the followers as if they were randomly drawn from a bag of words). The word *appalling* occurs just 471 times in the COCA, followed by 67 different words, making it over ten times easier ($1/67$ or 1.49%) to guess the following word given its occurrence. The word *dubious* has 141 COCA followers and the word *terrible* has 832. Depending on how we define their local context, these words might have quite different information

theoretic (PMI or surprisal) values. However, all four words are treated almost identically in our analysis in terms of likely need, since they all occur in the closest 30 words (out of 76,278 words) to the adjective CDV. If a person learned for the first time that the word *horrendous* has a similar *meaning* to the word *terrible*, that person could contextualize that word and immediately know that the phrase *horrendous dyed hair* is more acceptable than *dyed horrendous hair*.

Ultimately, predict models of word meaning *are* information theoretic models, since they are only measuring similarity between string contexts (Levy & Goldberg, 2014). Others (Meyer, 2016; Rong, 2014) have pointed out that the loss function in word2vec models (the measure that is minimized while training the models) is a measure of *cross-entropy*, a measure of the number of bits needed to identify (contextualize) an event from one probability distribution (the true distribution of words) using a sample estimate of that probability distribution (the observed distribution of words). Likely need as defined here is precisely an information theoretic measure, albeit one that relies upon a complex representation of context (that which is encoded in a word's vector, or in the averaged contexts of the multiple words that make up the CDV) in computing a word's entropy.

An apparent complication to the claim that adjective order is determined by likely need is that some languages with postnominal adjectives do not exhibit the same ordering as in English, but rather show the mirror image order, with more particular adjectives coming immediately after the adjective followed by more general adjectives (i.e., Romanian; Marchis & Alexiadou, 2009). A possible explanation for this observation is that linguistic likely need is conditioned on the immediately prior word or words. One piece of evidence in support of this possibility is that a CDV for concrete nouns (defined by the average vector of 1,000 vectors of concrete nouns) has a very strong negative correlation with the adjective CDV ($r = -.72$, $p < 2e-16$ across all 78.2k words in the co-occurrence matrix)—that is, words close to the noun CDV are closer to the words that are distant from adjective CDV (the semantically specific adjectives that tend to occur close to the adjective in English) than to words that are close to the adjective CDV (the less specific adjectives that tend to occur further from the adjective in English). The effect is illustrated in Fig. 10, which shows the average distance of the adjectives and nouns from the CDVs of both parts of speech for all triplets that were always preferred in one order by the participants in the experiment. As discussed above, adjectives occurring in the second position are further from the adjective CDV (average [*SD*] cosine distance: 0.89 [0.09]) than adjectives occurring in the first position (CDV average [*SD*] cosine distance: 0.69 [0.11]), $t(64.1) = 8.6$, $p = 3.2e-12$. The opposite is true with respect to the noun CDV: Adjectives occurring in the second position are closer to the noun CDV (average [*SD*] cosine distance:

Fig. 10 Average cosine distance of the adjectives and nouns from the adjective and noun category-defining vectors (CDVs) of both parts of speech, for all triplets that were always preferred in one order by the participants in the experiment. Adjectives coming first in prenominal adjective pairs are much closer to the adjective CDV than to the noun CDV

0.96 [0.16]) than adjectives occurring in the first position (average [*SD*] cosine distance: 1.03 [0.11]), $t(58.7) = 2.3$, $p = .02$.

Another apparent complication to a likely need explanation of adjective order is that two adjectives separated by *and* (which I will call *conjoined adjectives*) are not always ordered in the same way as two prenominal adjectives. Conjoined adjectives are rare in English compared with multiple prenominal adjectives. The relationship between prenominal and conjoined adjectives is shown graphically in Fig. 11, for the only 45 adjective pairs (9%) that occurred at least 10 times in at least one order in the COCA with *and* between them, from the 500 pairs used in the experiment. The probability of conjoined adjectives was computed as the number of times it occurred in the most probable prenominal order (as judged by humans), divided by the total number of occurrences in either order. In keeping with the claims of Rosales and Scontras (2019) that prenominal adjective order preferences “neutralize in cases where multi-adjective strings are formed via conjunction” (p. 1), the data suggest that the relationship between human-preferred prenominal adjective order and the probability of conjoined adjective order is weak, though the correlation is marginally significant ($r = .30$, $p = .04$; 95% CI [0.0079, 0.55]). The relationship between likely need estimate and probable conjoined adjective order is not significant ($r = .13$, $p = 0.40$; 95% CI [−0.17, 0.40]). Several adjective pairs occur only in a single conjoined order. Thirteen pairs (28.9%) are perfectly consistent with both human preference and the likely need estimate, occurring conjoined in the COCA only in the most likely prenominal order (e.g., *long and slender*; *young and single*, *big and fat*). Almost as many (12 pairs; 26.7%)

Fig. 11 Relationship between conjoined adjective order probability (Y-axis) and human judgments and model estimates of prenominal adjective order (X-axis)

are conjoined only in the less likely prenominal order (e.g., *young and energetic*, *front and left*, *white and pale*). It is notable (and discussed further below) that half of those always-reversed conjoined pairs include either the adjective *young* or *old*. The failure of likely need to account for conjoined adjective order probably occurs because many other relevant linguistic forces can be brought to the forefront by conjoining two adjectives. For example, the term *blond and beautiful* (which is more common than the prenominally ordered *beautiful and blond*) seems to deliberately emphasize beauty over blondness, especially when it occurs at the end of a phrase. Contrast *They were all tall, blond, and beautiful* with *They were all tall, beautiful, and blond*. The latter sentence seems to place emphasis on hair color, the former on beauty. Some conjoined adjectives express temporal order, as in *four strips of bacon, fully cooked and crumbled*. The conjoined order goes against the most likely prenominal order here to convey the idea that bacon strips must be cooked before they can be crumbled. Common prenominally reversed terms like *elderly and disabled* and *young and pretty* may be expressing a similar temporal or causal order (i.e., some disabled people are disabled as a result of being elderly and some youth are pretty precisely because they are young). This may explain the apparent special status of temporal adjectives such as *old* and *young* that has been noted above. Some conjoined adjectives occur because they are embedded in a sentence where they have a different than usual meaning, as the prenominally reversed phrase *front and left* was found by Google Ngram Viewer (<http://books.google.com/ngrams>) inside the sentence “Spreading instruments between left-front and left-surround, and right-front and right-surround, gives a greater sense of envelopment” (Bartlett & Bartlett, 2009, p. 468). There are

probably many other factors that impinge on how conjoined adjectives are ordered, rendering likely need less central. These factors may have implications for understanding adjective order in languages such as Spanish, in which adjectives must obligatorily be conjoined (Rosales & Scontras, 2019).

Adjective order has often been treated as a “special case,” which needs to be explained with a tailor-made solution specific only to the phenomenon (as in Dixon, 1977; Wulff, 2003, among others). The likely need explanation of prenominal adjective ordering that is offered here is appealing because it offers an explanation that builds on a broad range of previous theoretical and empirical work in psychology. The original theory in memory work by Anderson and colleagues (Anderson, 1991; Anderson & Milson, 1989; Anderson & Schooler, 1991) was specifically conceived and explicitly noted to be consistent with both evolutionary theory and rational engineering principles. That idea was imported into psycholinguistic studies as a more plausible explanation of word frequency effects than the implicit idea that we keep a running tally of how often we encounter a word (Adelman et al. 2006; Jones et al., 2012; Hollis, 2017). Distance from word category CDVs has previously been shown to be a powerful predictor of the speed and accuracy in go/no-go part-of-speech decisions (Westbury & Hollis, 2018), proving that humans are sensitive to this measure. The likely need model of adjective order is built on this earlier work. Since it places adjectives with high frequencies and low CDV distances (both of which are known to be accessed more quickly) further from the noun (earlier in the phrase) than other words, prenominal adjectives are ordered in terms of the speed with which they can be accessed, which is the order in which they are most likely to be needed.

Acknowledgements This work was made possible by a grant from the Natural Sciences and Engineering Research Council (NSERC) of Canada.

References

- Adelman, J. S., Brown, G. D., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science*, 17(9), 814–823.
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723. doi:<https://doi.org/10.1109/TAC.1974.1100705>
- Anderson, J. R. (1991). Is human cognition adaptive? *Behavioral and Brain Sciences*, 14(03), 471–485.
- Anderson, J. R., & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review*, 96(4), 703.
- Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science*, 2(6), 396–408.
- Anderson, R. B., & Tweney, R. D. (1997). Artifactual power curves in forgetting. *Memory & Cognition*, 25(5), 724–730.
- Annear, S. A. (1964). *The ordering of pre-nominal modifiers in English* (Project on Linguistic Analysis, Report No. 8). Columbus, OH: State University Research Foundation.
- Bartlett, B., & Bartlett, J. (2009). *Practical recording techniques: The step-by-step approach to professional audio recording* (5th). Oxford, England: Focal Press.
- Byrne, B. (1979). Rules of prenominal adjective order and interpretation of “incompatible” adjective pairs. *Journal of Memory and Language*, 18(1), 73–78. doi:[https://doi.org/10.1016/S0022-5371\(79\)90574-7](https://doi.org/10.1016/S0022-5371(79)90574-7)
- Danks, J. H., & Glucksberg, S. (1971). Psychological scaling of adjective orders. *Journal of Memory and Language*, 10(1), 63–67. doi:[https://doi.org/10.1016/S0022-5371\(71\)80095-6](https://doi.org/10.1016/S0022-5371(71)80095-6)
- Dixon, R. M. W. (1977). Where have all the adjectives gone? *Studies in Language*, 1, 19–80.
- Dye, M., Milin, P., Futrell, R., & Ramscar, M. (2018). Alternative solutions to a language design problem: The role of adjectives and gender marking in efficient communication. *Topics in Cognitive Science*, 10(1), 209–224.
- Dyer, W. (2018). Integration complexity and the order of cosisters. In: *Proceedings of the Second Workshop on Universal Dependencies* (pp. 55–65). Brussels: Association for Computational Linguistics.
- Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, 336(6084), 998–998.
- Franke, M., & Jäger, G. (2016). Probabilistic pragmatics, or why Bayes’ rule is probably important for pragmatics. *Zeitschrift für sprachwissenschaft*, 35(1), 3–44.
- Franke, M., Scontras, G., & Simonić, M. (2019). Subjectivity-based adjective ordering maximizes communicative success. In Ashok Goel, Colleen Seifert & Christian Freksa (eds.), *Proceedings of the 41st Annual Conference of the Cognitive Science Society* (pp. 344–350). Montreal, QB: Cognitive Science Society.
- Fries, C. C. (1952). *The structure of English*. New York, NY: Harcourt, Brace, and World.
- Futrell, R. (2019). Information-theoretic locality properties of natural language. In: *Proceedings of the first workshop on quantitative syntax* (pp. 2–15). Paris: Association for Computational Linguistics.
- Futrell, R., Levy, R., & Gibson, E. (2017). Generalizing dependency distance. *Physics of Life Reviews*, 21, 197–199.
- Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20(11), 818–829.
- Goyvaerts, D. L. (1968). An introductory study on the ordering of a string of adjectives in present-day English. *Philologica Pragensia*, 11, 12–28.
- Grice, P. H. (1975). *Logic and conversation*. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics: Vol. 3, Speech acts* (pp. 41–58). New York, NY: Academic Press.
- Hale, J. (2003). The information conveyed by words in sentences. *Journal of Psycholinguistic Research*, 32(2), 101–123.
- Hahn, M., Degen, J., Goodman, N. D., Jurafsky, D., & Futrell, R. (2018). An Information-Theoretic Explanation of Adjective Ordering Preferences. In: *Proceedings of the 40th Annual Meeting of the Cognitive Science Society*. London: Cognitive Science Society.
- Hetzron, R. (1978). On the relative order of adjectives. In H. Seiler (Ed.), *Language universals* (pp. 165–184). Tübingen, Germany: Gunter Narr Verlag.
- Hollis, G. (2017). Estimating the average need of semantic knowledge from distributional semantic models. *Memory & Cognition*, 45(8), 1350–1370.
- Huang, H. W., & Federmeier, K. D. (2012). Dispreferred adjective orders elicit brain responses associated with lexico-semantic rather than syntactic processing. *Brain Research*, 1475, 62–70.
- Jones, M. N., Johns, B. T., & Recchia, G. (2012). The role of semantic diversity in lexical organization. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 66(2), 115–124. doi:<https://doi.org/10.1037/a0026727>

- Kemmerer, D. (2000). Selective impairment of knowledge underlying prenominal adjective order: Evidence for the autonomy of grammatical semantics. *Journal of Neurolinguistics*, 13(1), 57–82.
- Kemmerer, D., Weber-Fox, C., Price, K., Zdanczyk, C., & Way, H. (2007). Big brown dog or brown big dog? An electrophysiological study of semantic constraints on prenominal adjective order. *Brain and Language*, 100(3), 238–256.
- Kemmerer, D., Tranel, D., & Zdanczyk, C. (2009). Knowledge of the semantic constraints on adjective order can be selectively impaired. *Journal of Neurolinguistics*, 22(1), 91–108.
- Levy, O., & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. *Advances in Neural Information Processing Systems*, 27, 2177–2185.
- Liu, H., Xu, C., & Liang, J. (2017). Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*, 21, 171–193.
- Malouf, R. (2000). The order of prenominal adjectives in natural language generation. *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics* (pp. 85–92). doi:<https://doi.org/10.3115/1075218.1075230>
- Marchis, M., & Alexiadou, A. (2009). On the distribution of adjectives in Romanian. In E. O. Aboh, E. van der Linden, J. Quer, & P. Sleeman (Eds.), *Romance languages and linguistic theory: Selected papers from 'Going Romance' Amsterdam 2007*, 1, 161–178. doi:<https://doi.org/10.1075/rllt.1.09mar>
- Martin, J. E. (1969). Semantic determinants of preferred adjective order. *Journal of Verbal Learning and Verbal Behavior*, 8(6), 697–704.
- Martin, J. E., & Ferb, T. E. (1973). Contextual factors in preferred adjective ordering. *Lingua*, 32, 75–81.
- Meyer, D. (2016). *How exactly does Word2vec work?* Retrieved from http://www.1-4-5.net/~dmm/ml/how_does_word2vec_work.pdf
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space*. ArXiv preprint, 1301.3781.
- Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., ... Zeman, D. (2016). Universal dependencies v1: A multilingual treebank collection. *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)* (pp. 1659–1666). Retrieved from <https://www.aclweb.org/anthology/L16-1262.pdf>
- Roberts, M. A., & Chater, N. (2008). Using statistical smoothing to estimate the psycholinguistic acceptability of novel phrases. *Behavior Research Methods*, 40(1), 84–93.
- Rong, X. (2014). *Word2vec parameter learning explained*. ArXiv preprint: 1411.2738
- Rosales, C. M., Jr., & Scontras, G. (2019). On the role of conjunction in adjective ordering preferences. *Proceedings of the Linguistic Society of America*, 4(1), 32. doi:<https://doi.org/10.3765/PLSA.V4I1.4524>
- Scheffelin, M. (1971). Childrens' understanding of constraints upon adjective order. *Journal of Learning Disabilities*, 4(5), 264–272.
- Scontras, G., Degen, J., & Goodman, N. D. (2017). Subjectivity predicts adjective ordering preferences. *Open Mind*, 1(1), 53
- Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3), 79–423.
- Sichelschmidt, L. (1986). Optionality and choice in the analysis of adjective order: Comments on Ney. *Studia Linguistica*, 40(2), 135–148.
- Simonič, M. (2018). *Functional explanation of adjective ordering preferences using probabilistic programming* (Unpublished master's thesis). University of Tübingen.
- Sproat, R., & Shih, C. (1991). The cross-linguistic distribution of adjective ordering restrictions. In C. Georgopoulos & R. Ishihara (Eds.), *Interdisciplinary approaches to language* (pp. 565–593). Dordrecht, Netherlands: Springer. doi:<https://doi.org/10.1007/978-94-011-3818-5>
- Sussex, R. (1974). The deep structure of adjectives in noun phrases. *Journal of Linguistics*, 10(1), 111–131.
- Sweet, H. (1955). *A new English grammar: Part II*. Oxford, England: Clarendon Press. (Original publication 1898)
- Vandekerckhove, B., Sandra, D., & Daelemans, W. (2013). Selective impairment of adjective order constraints as overeager abstraction: An elaboration on Kemmerer et al. (2009). *Journal of Neurolinguistics*, 26(1), 46–72.
- Vecchi, E. M., Zamparelli, R., & Baroni, M. (2013). Studying the recursive behaviour of adjectival modification with compositional distributional semantics. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 141–151). Retrieved from <https://www.aclweb.org/anthology/D13-1015>
- Vendler, Z. (1963). The transformation grammar of English adjectives. *Transformations and discourse analysis papers*, University of Pennsylvania.
- Vendler, Z. (1968). *Adjectives and nominalizations*. Mouton, Netherlands: The Hague.
- Westbury, C., & Hollis, G. (2018). Conceptualizing syntactic categories as semantic categories: Unifying part-of-speech identification and semantics using co-occurrence vector averaging. *Behavior Research Methods*, 51, 1371–1398. doi:<https://doi.org/10.3758/s13428-018-1118-4>
- 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.
- Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Oxford, England: Addison-Wesley.

The data and materials for all experiments are available at the Open Science Framework (<https://osf.io/5x4tw/>; doi:10.17605/OSF.IO/5X4TW).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.