Evaluating a Century of Progress on the Cognitive Science of Adjective Ordering

William Dyer
AWS AI Labs
wdyer@amazon.com

Charles Torres
University of California, Irvine
charlt4@uci.edu

Gregory Scontras
University of California, Irvine
g.scontras@uci.edu

Richard Futrell
University of California, Irvine
rfutrell@uci.edu

Abstract

The literature on adjective ordering abounds with proposals meant to account for why certain adjectives appear before others in multi-adjective strings (e.g., *the small brown box*). However, these proposals have been developed and tested primarily in isolation and based on English; few researchers have looked at the combined performance of multiple factors in the determination of adjective order, and few have evaluated predictors across multiple languages. The current work approaches both of these objectives by using technologies and datasets from natural language processing to look at the combined performance of existing proposals across 32 languages. Comparing this performance with both random and idealized baselines, we show that the literature on adjective ordering has made significant meaningful progress across its many decades, but there remains quite a gap yet to be explained.

1 Introduction

Adjective ordering preferences regularly appear across the world's languages: In nominal constructions with multiple adjective modifiers (e.g., *the small brown box*), speakers often (strongly) prefer one ordering. Furthermore, these preferences are often the same across languages for translation-equivalent adjectives. This striking regularity raises the question of what aspects of language or its use in communication yield the observed preferences. After more than a century of research, linguists and cognitive scientists have proposed an array of hypotheses for predicting adjective ordering in terms of cognitive factors affecting language production and linguistic representations.

To date, most investigations of these cognitive hypotheses about adjective order have considered single predictors in isolation, or have compared their performance on a single language (i.e., English; for discussion, see Scontras, 2023). This situation is not ideal, especially considering the cognitive theories we survey below were developed in the context of predicting adjective order in English only, often leaving their cross-linguistic generality unclear. However, the recent availability of massively cross-linguistic parsed datasets and the development of NLP technologies such as word embeddings have opened up the possibility for large-scale evaluations of a wide variety of cognitive hypotheses against a wide range of data.

Our goal in this paper is to evaluate the predictive power of cognitive hypotheses for adjective order across 32 languages, and to situate their performance with respect to two baselines: (i) a lower baseline representing random chance accuracy in predicting adjective order, and (ii) a baseline that reflects the best performance that can be achieved in predicting order directly from the distributional and semantic information encoded in modern word embeddings. While this neural distributional baseline provides a strong descriptive account of adjective order, it does not provide an explanation of *why* adjectives are ordered in the way they are, as the cognitive predictors do. By situating the performance of cognitive predictors between these baselines, our goal is to determine how much progress has been made in the scientific explanation of adjective order over 125 years of research, and how much remains to be explained.

The remainder of the paper is structured as follows. Section 2 describes the data sources we draw on to operationalize and evaluate predictors of adjective order and how we extract their data.

Transactions of the Association for Computational Linguistics, vol. 11, pp. 1185–1200, 2023. https://doi.org/10.1162/tacl_a_00596
Action Editor: Marco Baroni. Submission batch: 12/2022; Revision batch: 1/2023; Published 9/2023.

Section 3 presents the cognitive predictors and how we implement them. Section 4 describes our evaluation method, including the formulation of baselines. Section 5 presents the results with some discussion, and Section 6 concludes.

2 Data

Recent years have seen massive expansions in the availability of crosslinguistic datasets. In particular, the Universal Dependencies (UD) project (Nivre et al., 2016) has gathered dependency-parsed corpora of naturalistic text in many languages. It is exactly this dependency-parsed naturalistic data that is useful for a crosslinguistic corpus study of adjective order, because it is possible to easily extract instances of multiple adjectives modifying a single noun, and then to study the ordering patterns found in these instances. Our goal is to study the ability of cognitively-motivated theories to predict the attested orders of adjectives as found in these corpora.

The primary syntactic configuration that we extract from dependency-parsed corpora is what we call a **triple**, consisting of a head noun token N with universal part-of-speech `NOUN`, modified by exactly two distinct adjective tokens A_1 and A_2, with universal part-of-speech `ADJ`, and with the syntactic relation type `amod`. Given a triple $\{A_1, A_2, N\}$ extracted based on syntactic configuration, our goal is to predict the linear order of the words: whether it is A_1A_2N, A_2A_1N, A_1NA_2, etc. We classify triples into three **templates**: noun-final (AAN), noun-medial (ANA), and noun-initial (NAA), and we predict order within each of these templates. The diversity of lexical types in the triples data is shown in Table 1, represented as type-to-token ratios for individual adjectives. The data shows reasonable diversity of types, and the type–token ratios are not significantly different across templates.[1]

Many of the cognitive predictors that we use rely on relative frequency counts for adjectives co-occurring with nouns. For these predictors, we estimate their values based on counts of **pairs**: instances of a single head noun (universal POS `NOUN`) modified by a single adjective (universal POS `ADJ`, with relation type `amod`). We use pairs extracted from the automatically parsed Wikipedia dump datasets released as part of the CoNLL 2017 Shared Task (Zeman et al., 2017). We will refer to these pairs as the **training pairs**. For our test set of triples, which will be used for the final evaluation of predictors, we use the Universal Dependencies 2.8 corpora, concatenating non-L2 corpora for each language.

	type	token	ratio
AAN			
Bulgarian	97	106	0.92
Croatian	90	114	0.79
Czech	838	1490	0.56
Danish	43	46	0.93
Dutch	122	190	0.64
English	348	546	0.64
Estonian	104	118	0.88
German	711	1258	0.57
Greek	26	32	0.81
Hindi	21	22	0.95
Polish	48	56	0.86
Russian	395	538	0.73
Slovak	25	28	0.89
Slovenian	39	42	0.93
Swedish	82	106	0.77
Turkish	81	126	0.64
Ukrainian	59	60	0.98
ANA			
Catalan	98	114	0.86
French	258	420	0.61
Italian	234	310	0.75
Polish	116	134	0.87
Portuguese	108	154	0.70
Romanian	39	44	0.89
Spanish	220	296	0.74
NAA			
Arabic	135	224	0.60
Catalan	111	136	0.82
French	288	426	0.68
Hebrew	30	30	1.00
Italian	167	270	0.62
Portuguese	97	114	0.85
Romanian	149	176	0.85
Spanish	269	330	0.82

Table 1: Type–token ratios for adjectives in held-out test triples.

[1] Unpaired t-tests comparing average type–token ratios in the different templates give $p > 0.05$ for all comparisons.

In estimating several of our predictors, we make use of word vectors. In all instances, we use the aligned word vectors provided by Facebook,[2] which were trained on data from Wikipedia (Bojanowski et al., 2017; Joulin et al., 2018).

3 Predictors

We evaluate the performance of eight predictors from the literature on adjective ordering. Our choice of predictors is based on the criterion that predictors must have a precise operationalization that can be estimated using the data at hand.

The cognitive predictors differ in whether they predict adjectives to come *close* to the noun, or whether they predict adjectives should come generally *earlier* in the linear order of an utterance as a whole. When a predictor holds that an adjective should be close to the noun, its effect on linear order should be opposite for pre- and post-nominal adjectives, with varying and often unclear predictions for the ANA template. When the predictor holds that an adjective should be generally early, its effect on linear order should have the same sign for pre- and post-nominal adjectives. For predictors that were developed in the monolingual English context, where the only permissible template is AAN, the proper polarity of predictions is sometimes unclear for other templates such as NAA and ANA, as we discuss below.

Below, we briefly describe each predictor, its history in the linguistics and cognitive science literature, and how it was estimated for our study.

Frequency Several authors have shown adjective frequency to be a reliable predictor of adjective order, with more frequent adjectives appearing earlier (Martin, 1969; Wulff, 2003; Scontras et al., 2017; Trotzke and Wittenberg, 2019; Westbury, 2021). This effect of frequency is consistent with a broader finding that more frequent words appear earlier in sentences. The pattern has been explained in terms of a general preference for more 'accessible' words to go earlier in utterances as a result of a kind of greediness in human sentence production (Bock, 1982; Ferreira and Dell, 2000; Chang, 2009). To date, existing studies of frequency effects have focused on English ordering. Our frequency predictor consists of the log-transformed raw counts of adjectives appearing as dependents in the training pairs.

Length Another accessibility-based predictor of word order is word length: There is a general tendency for short words and phrases to go before long ones, as evidenced in production experiments and corpus studies (Behaghel, 1909; Stallings et al., 1998; Bresnan et al., 2007). Applied to adjective order, this predictor has been evaluated successfully only in English (Wulff, 2003; Scontras et al., 2017; cf. Kotowski and Härtl, 2019 for a different finding in German). The general short-before-long preference is also considered an accessibility effect (Stallings et al., 1998), since short words and phrases are easier to access and produce than long ones.

Meaning Specificity One of the oldest ideas in the literature on adjective ordering holds that adjectives more ''special in meaning'' appear nearer to the noun (Sweet, 1898, p. 8). A common way of interpreting meaning specificity concerns the range of nouns an adjective can modify; adjectives applicable to a narrower range of nouns will have a more specific meaning (Ziff, 1960; Seiler, 1978). Here we explore two different operationalizations of meaning specificity. The first is **integration complexity** (IC), which quantifies the entropy of the probability distribution of a word's heads in dependency trees; adjectives combining with a broader range of nouns as heads will have higher integration complexity and should appear farther form the modified noun (Dyer, 2017, 2018; Futrell et al., 2020a). The distribution on head nouns given adjectives is estimated from a training corpus to be described below.

The second operationalization of meaning specificity is in terms of Westbury's (2021) notion of 'likely need'. The intuition for this measure is that adjectives with a multi-purpose meaning will be used across a wider range of contexts, and so, across contexts, speakers will be more likely to need to use those more flexible, more general adjectives and will use them earlier. We adopt Westbury's (2021) operationalization of this idea: Adjectives whose semantic vector is closer to the average adjective vector—the 'category-defining vector', or **CDV**—will have a more general meaning and will therefore appear earlier. The average adjective meaning is determined according to the

[2] https://fasttext.cc/docs/en/aligned-vectors.html.

token frequency of adjectives in our training pairs. The predictions of the theory for other templates are not clear.

Meaning Closeness Another predictor with a century-long history concerns the meaning connection between the adjective and noun. Crucially, these predictors depend on the specific noun being modified; the other predictors described so far are only a function of individual adjectives. According to Sweet (1898, p. 8), the adjective ''most closely connected with [the noun] in meaning'' comes nearest to it. This idea of meaning closeness has resurfaced in various forms in the intervening years (e.g., Ziff, 1960; Hetzron, 1978; Byrne, 1979; McNally and Boleda, 2004; Bouchard, 2005; Svenonius, 2008). For our purposes, we consider two operationalizations. The first, pointwise mutual information, or **PMI**, quantifies the information that adjectives and nouns have in common on the basis of the extent to which they occur together (Fano, 1961; Church and Hanks, 1990; Futrell et al., 2020a); adjectives with higher PMI with the modified noun should appear closer to the noun. PMI has a cognitive justification in terms of minimizing processing difficulty under memory limitations (Futrell, 2019; Futrell et al., 2020b). We calculate PMI using the additively smoothed distribution on head nouns given adjectives in the training data (with smoothing constant $\alpha = .001$).

The second operationalization of meaning closeness is inspired by the distributional hypothesis (Firth, 1957), where an adjective operates on a noun by changing its distribution in vector space (Baroni and Zamparelli, 2010). We quantify that change by the vector cosine distance, or **VCosD**, between the noun and summed noun-adjective vectors, which are meant to represent the composition of the adjective and the noun (Paperno and Baroni, 2016). The intuition is that some adjectives may drastically change the distribution of the noun, while others do so only negligibly, and this change may relate to adjective order: Adjectives with larger VCosD should appear farther from the modified noun. Interestingly, VCosD has been related to PMI by Ethayarajh et al. (2019).

Information Gain Proposed by Dyer et al. (2021), information gain quantifies the amount of information about a referent provided by the occurrence of an adjective. The cognitive motivation for this predictor is the idea that speakers greedily maximize information gain, resulting in adjectives which offer a greater reduction in uncertainty appearing earlier. In the previous work, information gain was shown to be a strong predictor of adjective order across languages and templates.

Subjectivity A separate line of research has proposed that adjective subjectivity predicts ordering preferences, with less subjective adjectives appearing closer to the noun (Quirk et al., 1972; Hetzron, 1978; Scontras et al., 2017). The subjectivity hypothesis has been extensively tested in English (Scontras et al., 2017; Hahn et al., 2018; Futrell et al., 2020a), and in several other languages (Samonte and Scontras, 2019; Kachakeche and Scontras, 2020; Shi and Scontras, 2020; Scontras et al., 2020). A number of cognitive justifications for subjectivity as a predictor of adjective order have been offered. For example, Franke et al. (2019) show that orders with more subjective adjectives farther from the noun can maximize communicative success in a setting where semantic composition is noisy. Previously, adjective subjectivity has been estimated behaviorally by asking participants how ''subjective'' a given adjective is, or by having them assess its potential for faultless disagreement (i.e., whether two people could both be right while disagreeing about whether some adjective holds of an object; Kölbel, 2004; MacFarlane, 2014).

This behavioral measure is logistically challenging to collect for a large set of languages and adjectives. Therefore, we adopt the method of **semantic norm extrapolation** (Tang et al., 2014; Tsvetkov et al., 2014; Ljubešić et al., 2018): We use new and existing experimental datasets to train a neural network to predict subjectivity ratings from word embeddings, and then use this network to deliver estimated subjectivity ratings for adjectives. We use these estimated subjectivity scores in all cases. We use aligned word vectors (Joulin et al., 2018) so that we can transfer this network cross-linguistically to yield extrapolated subjectivity scores across languages.

In order to train networks for subjectivity prediction, we adapted existing datasets of experimentally elicited subjectivity ratings for adjectives. For non-English languages, subjectivity ratings were elicited in previous work using the

	train		dev		test		
	type	token	type	token	type	token	source
Arabic[a]	20	640	2	64	3	96	Kachakeche and Scontras (2020)
English	533	11147	65	1349	68	1389	Futrell et al. (2020a), new data
German	20	803	2	81	3	121	Scontras et al. (2021)
Greek[a]	20	2030	2	210	3	315	Scontras et al. (2021)
Hebrew	20	420	2	42	3	63	Scontras et al. (2021)
Mandarin	21	735	2	70	3	105	Shi and Scontras (2020)
Spanish[a]	21	882	2	84	3	126	Rosales Jr. and Scontras (2019)
Tagalog	20	220	2	22	3	33	Samonte and Scontras (2019)
Vietnamese	14	238	1	17	2	34	Scontras et al. (2021)

[a] Language data generated after duplication by gender to normalize the ratings.

Table 2: Number of distinct predicates (types) and the collected responses (tokens) used to train subjectivity model.

faultless disagreement task. For English, we also collected additional subjectivity ratings for 343 adjectives in the English Universal Dependencies corpus that appeared in multi-adjective strings. Using the ''subjectivity'' method from Scontras et al. (2017), participants ($n = 235$) rated the subjectivity of 30 unique adjectives, with an average of 21 ratings collected per adjective. The characteristics of these datasets are shown in Table 2.

For the subjectivity prediction network's architecture, we used a feedforward network with a single hidden layer of 128 neurons and ReLU activations, trained using Adam (Kingma and Ba, 2014). We split the above dataset in three ways in an 80/10/10 train/dev/test split within languages. Training was performed on the training set until there was no learning on the development set for 10 continuous epochs. Evaluations on the resulting test set showed a strong correlation with the empirical data (Spearman's $\rho = 0.86$, Pearson's $r = 0.87$).

4 Evaluation Method

4.1 Goals

To date, nearly all of the quantitative investigations of cognitive theories of adjective ordering have evaluated the performance of a single predictor in a single language, with a few exceptions (cf. Wulff, 2003; Scontras et al., 2017; Futrell et al., 2020a; Dyer et al., 2021). The result is that it is not clear how robustly cognitive predictors generalize across languages, nor how well they perform in aggregate in predicting adjective order. Our goal is to evaluate this aggregate cross-linguistic performance, and to situate that performance with respect to a lower baseline of random chance guessing and a baseline representing the best that can be attained using the full semantic and distributional information contained in modern word embeddings of adjectives. The results give a picture of (1) how robust and consistent the different cognitive predictors are across languages and templates, (2) how much variance in adjective order is explained by cognitive theories, and (3) how much remains to be explained, in terms of the discrepancy between the performance of an ensemble of cognitive predictors vs. the distributional baseline.

Our main goal is not to directly compare the predictors of adjective order on their accuracy. The reason for this choice of goal is twofold. First is a practical consideration: Given the sizes of the existing datasets, the accuracy values for the different predictors have overlapping confidence intervals, and so it is not possible to confidently state that one predictor is more accurate than another robustly. This limitation is not only due to the sizes of the test sets: There is also considerable uncertainty in the values of the predictors as estimated from the training pairs and word embeddings. The second reason to forego head-to-head comparisons between predictors is the emerging consensus within the literature on adjective ordering that a full account necessarily involves multiple predictors, some of them exerting competing pressures (Wulff, 2003; Futrell

et al., 2020a; Scontras, 2023). Nevertheless, our results will reveal that some predictors are more robust than others in terms of being consistently informative across languages.

Our study differs in its goal from descriptive studies such as Malouf (2000) and Leung et al. (2020), which study how well adjective ordering preferences can be learned from examples of ordered adjectives in text corpora. Such descriptive studies correspond to our distributional baseline: indeed, our distributional baseline implementation is closely related to the descriptive model of Leung et al. (2020), differing primarily in that we do not impose a total ordering constraint.

In contrast to such studies, our goal is to examine explanatory accounts of adjective ordering, in which adjective order is predicted *a priori* based on cognitive theories. To the extent that the values of our cognitive predictors depend on corpus counts, these counts themselves do not depend on the order of the adjectives in those corpora: They are based on training pairs which are extracted solely based on syntactic dependency configuration and not on word order. In practical terms, we are evaluating the ability of cognitive theories to provide a zero-shot feature set that is informative about adjective order.

Below, we describe our evaluation procedure in terms of our distributional baseline, lower baseline, and ensemble of cognitive predictors, and how these models are evaluated against test sets of triples.

4.2 Distributional Baseline

The distributional baseline for adjective order prediction represents how well the order of adjectives in a triple can be predicted based on full distributional information about the adjectives and noun, as present in aligned word embeddings. In theory, this baseline, which does not operate under the constraint of being cognitively motivated, should always outperform the cognitive predictors to some extent, simply because it is less constrained.

To calculate the distributional baseline, we trained batches of deep neural networks on a designated training set before evaluating their performance on a designated test set.[3] The fastText vectors we used as input were always submitted to the network with the adjectives' vectors concatenated in alphabetical order.[4] The network's target was to predict if this ordering was the attested linear order for the triple. For each template and language, thirty such networks were trained until performance on a designated development set failed to improve further from training. We trained networks both with and without hidden layers.

Word embeddings based on distributional information are widely accepted to contain (or at least correlate with) semantic information about words; however, they may be missing some information that cannot be easily recovered from words' distribution in text (e.g., information that would allow for the disambiguation of word senses based on context; Lenci et al., 2022). These are limitations that prevent our distributional baseline from achieving the full accuracy that might be possible from predicting adjective order from adjective semantics. At the same time, depending on the specific training method, word embeddings may contain some indirectly-encoded information about relative word orders, which would make the distributional baseline higher than it would be if it were purely semantic. For example, fastText vectors are trained by predicting words based on context words within a window of varying size 1–5 (Bojanowski et al., 2017). If an adjective consistently appears with many other adjectives modifying the same noun, and consistently appears far from that noun, then the noun may drop out of its context window during training. Because of these limitations, we refer to this baseline as a 'distributional baseline' rather than a semantic baseline.

4.3 Lower Baseline

While the lower accuracy baseline for a binary classification task is naïvely $1/2$—for example, in the NAA template, a choice between NA_1A_2 and NA_2A_1 orders—when classifying across a set of adjective–noun triples, the lower baseline may be different due to an uneven distribution of adjectives. Therefore, for our lower baseline we

[3] Distributional baseline DNNs had three hidden layers (300, 150, 75) and were trained with Adam (Kingma and Ba, 2014) at a learning rate of .0001, with both hyperparameters determined by a gridsearch over learning rates and neural network depth, choosing the highest mean performance over our dataset's development sets.

[4] Alphabetical ordering is used as a canonical ordering for triples that creates an approximately even split between triples whose true order is the canonical one and those whose true order is the opposite.

simply created a random predictor: Each adjective wordform was assigned a random uniform value in $[0, 1]$, then adjective order for a triple was predicted in a logistic regression based on the difference in random predictors for the two adjectives. Averaging 100 runs of this process yields a lower baseline which takes into account the distribution of adjectives across our triples.

4.4 Cognitive Predictors

To evaluate cognitive predictors, we train logistic regressions to predict the order of adjectives in a triple based on features consisting of values of our cognitive predictors for the two adjectives. The cognitive features are presented as the difference between values for the alphabetically first adjective minus the second. The logistic regression classifier predicts whether the alphabetical order of adjectives is the attested order or whether it was flipped. We use logistic regression rather than other classifier methods because, in addition to maximizing the accuracy of predictions on the training set, logistic regression provides easily interpretable coefficients for the predictors. A positive coefficient in the regression indicates that an adjective with a larger value of a predictor should go earlier.

We evaluate cognitive predictors in isolation—with only one cognitive feature used as a predictor in the logistic regression—as well as in an ensemble model, which includes only those predictors found to have a significant slope in the individual regression (at $p < .05$), and also only those predictors that receive the same sign in the ensemble regression as in the original regression. These exclusions of predictors are made to ensure that we are making principled predictions that accord with the cognitive theories underlying these predictors.

Given the goal of evaluating the performance of cognitive predictors across languages, we report the accuracy of a model trained on data from all our languages save one, with that one held-out language's data serving as the test set. Our distributional baselines are similarly calculated with this approach of holding out a single language, made possible by the use of aligned fastText vectors. Finally, we report aggregate result for each template based on a 80:20 train–test split of all the triples within that template, across languages. We save the question of choice between different templates for future study—especially pertinent to Romance languages in which a choice between A_1NA_2 and NA_2A_1 is often possible.

4.5 Data Handling

In an effort to provide as close to an apples-to-apples comparison as possible, we implemented a number of constraints around our data prior to analysis. We limit our set of languages to those from which at least 100 triples can be extracted from Universal Dependencies corpora, and further specify that at least 10% of a language's triples must be of a given template in order for that template to be included in our results—the motivation being that we want to analyze productive templates for a language, not spurious triples derived from incorrect parsing or foreign sequences. Finally, in assembling our ensembles of cognitive predictors, we only measure those triples on which all predictors can operate. That is, due to sparsity, typos, or other noise in the data, some predictors may not give a prediction for a triple while other predictors can; these triples are not reported in our results for these predictors. The distributional baseline is evaluated using the full available training and test data.

5 Results

Tables 3 and 4 present the results of our cognitive predictors, both in terms of the best-performing single predictor (in blue) and the best-performing ensemble of predictors (in red); we also include the lower and distributional baselines. It should be noted up front that, although we present 'best' single predictors and ensembles, the confidence intervals around the predictions are large enough to include nearly all of the alternatives. Still, the picture that emerges is clear: Ensemble models outperform single predictors, suggesting that no single predictor yet considered will explain all of the ordering regularities; and the distributional baseline exceeds the performance of the ensemble models, suggesting that cognitive science has yet to exhaustively characterize the factors that enter into determining adjective order. However, the progress that has been made over the past century of research is non-negligible: Single predictors and ensemble models outperform the lower baseline significantly, at least at the template level, as evidenced by the non-overlapping confidence intervals.

		lower	single		cognitive	distributional	
	n	baseline	predictor	accuracy	ensemble	baseline	
AAN	1916	0.51 ± 0.02	subj	0.62 ± 0.02	0.71 ± 0.02	0.88 ± 0.01	
ANA	649	0.55 ± 0.04	length	0.75 ± 0.03	0.87 ± 0.03	0.89 ± 0.02	
NAA	676	0.52 ± 0.04	pmi	0.63 ± 0.04	0.66 ± 0.04	0.83 ± 0.02	

Table 3: Lower baseline, best-performing single cognitive predictor, best-performing ensemble of cognitive predictors, and distributional baseline test accuracy derived from 80:20 train/test split. n is the number of test triples per template for which cognitive predictors can be evaluated.

		lower	single		cognitive	distributional	
	n	baseline	predictor	accuracy	ensemble	baseline	
AAN							
Bulgarian	166	0.55 ± 0.08	length	0.56 ± 0.08	0.66 ± 0.07	0.84 ± 0.05	
Croatian	198	0.54 ± 0.07	subj	0.54 ± 0.07	0.68 ± 0.06	0.80 ± 0.06	
Czech	2813	0.49 ± 0.02	subj	0.61 ± 0.02	0.73 ± 0.02	0.86 ± 0.10	
Danish	99	0.46 ± 0.10	subj	0.60 ± 0.10	0.78 ± 0.08	0.76 ± 0.11	
Dutch	329	0.48 ± 0.05	subj	0.67 ± 0.05	0.70 ± 0.05	0.81 ± 0.03	
English	1214	0.45 ± 0.03	length	0.59 ± 0.03	0.71 ± 0.03	0.85 ± 0.20	
Estonian	198	0.39 ± 0.07	subj	0.65 ± 0.07	0.67 ± 0.07	0.78 ± 0.19	
German	2341	0.48 ± 0.02	ic	0.54 ± 0.02	0.69 ± 0.02	0.91 ± 0.19	
Greek	57	0.60 ± 0.13	subj	0.53 ± 0.13	0.65 ± 0.12	0.92 ± 0.11	
Hindi	60	0.50 ± 0.13	subj	0.58 ± 0.12	0.75 ± 0.11	0.77 ± 0.17	
Polish	123	0.61 ± 0.09	subj	0.59 ± 0.09	0.70 ± 0.08	0.82 ± 0.18	
Russian	1056	0.45 ± 0.03	subj	0.63 ± 0.03	0.76 ± 0.03	0.87 ± 0.14	
Slovak	64	0.55 ± 0.12	subj	0.61 ± 0.12	0.64 ± 0.12	0.98 ± 0.22	
Slovenian	51	0.49 ± 0.14	subj	0.51 ± 0.14	0.65 ± 0.13	0.97 ± 0.20	
Swedish	252	0.45 ± 0.06	subj	0.62 ± 0.06	0.74 ± 0.05	0.76 ± 0.19	
Turkish	296	0.40 ± 0.06	length	0.40 ± 0.06	0.47 ± 0.06	0.70 ± 0.45	
Ukrainian	126	0.53 ± 0.09	subj	0.63 ± 0.08	0.64 ± 0.08	0.79 ± 0.11	
ANA							
Catalan	286	0.54 ± 0.06	length	0.81 ± 0.04	0.86 ± 0.04	0.92 ± 0.06	
French	907	0.57 ± 0.03	length	0.80 ± 0.03	0.91 ± 0.02	0.89 ± 0.07	
Italian	749	0.49 ± 0.04	length	0.81 ± 0.03	0.86 ± 0.02	0.88 ± 0.01	
Polish	201	0.57 ± 0.07	length	0.61 ± 0.07	0.71 ± 0.06	0.85 ± 0.17	
Portuguese	255	0.60 ± 0.06	length	0.75 ± 0.05	0.85 ± 0.04	0.90 ± 0.01	
Romanian	73	0.70 ± 0.11	length	0.78 ± 0.09	0.84 ± 0.08	0.91 ± 0.07	
Spanish	790	0.55 ± 0.03	length	0.73 ± 0.03	0.89 ± 0.02	0.88 ± 0.01	
NAA							
Arabic	134	0.38 ± 0.08	pmi	0.46 ± 0.08	0.51 ± 0.08	0.69 ± 0.23	
Catalan	313	0.56 ± 0.06	pmi	0.70 ± 0.05	0.73 ± 0.05	0.78 ± 0.08	
French	925	0.53 ± 0.03	pmi	0.71 ± 0.03	0.73 ± 0.03	0.75 ± 0.06	
Hebrew	88	0.51 ± 0.10	pmi	0.57 ± 0.10	0.66 ± 0.10	0.85 ± 0.14	
Italian	572	0.53 ± 0.04	pmi	0.60 ± 0.04	0.60 ± 0.04	0.77 ± 0.11	
Portuguese	202	0.59 ± 0.07	pmi	0.70 ± 0.06	0.68 ± 0.06	0.76 ± 0.08	
Romanian	342	0.55 ± 0.05	pmi	0.59 ± 0.05	0.63 ± 0.05	0.79 ± 0.18	
Spanish	698	0.51 ± 0.04	pmi	0.66 ± 0.04	0.66 ± 0.03	0.75 ± 0.11	

Table 4: Lower baseline, best-performing single cognitive predictor, best-performing ensemble of cognitive predictors, and distributional baseline test accuracy derived by training on all languages in a given template except one held-out test language. n is the number of test triples per language and template for which cognitive predictors can be evaluated.

In Table 5, we present the cognitive predictors used by the best-performing ensemble models, both by template and by language. As mentioned above, these best-performing models have confidence intervals that overlap with several other ensemble models, which means one should be careful to not over-interpret the presence of a predictor in a best-performing ensemble (or, conversely, to over-interpret the absence of a predictor).

Still, the results reveal some striking regularities in terms of which predictors are informative

	CDV	IC	IG	LENGTH	FREQ	VCosD	PMI	SUBJ
AAN			+	−			−	+
Bulgarian		+	+	−		+	−	+
Croatian		+	+	−			−	+
Czech			+	−			−	+
Danish		+	+	−			−	+
Dutch		+	+	−			−	+
English	−		+	−		+	−	+
Estonian		+	+	−		+	−	+
German	−	+	+	−			−	+
Greek		+	+	−		+	−	+
Hindi		+	+	−			−	+
Polish		+	+	−		+	−	+
Russian			+	−			−	+
Slovak		+	+	−			−	+
Slovenian		+	+	−			−	+
Swedish		+	+	−			−	+
Turkish	−		+	−		+	−	+
Ukrainian		+	+	−		+	−	+
ANA	−		+	−			−	+
Catalan	−		+	−			−	+
French	−		+	−			−	+
Italian	−	+	+	−		−	−	+
Polish	−		+	−			−	+
Portuguese	−		+	−			−	+
Romanian	−		+	−			−	+
Spanish	−		+	−			−	+
NAA	+	−		+	+		+	−
Arabic	+	−		+	+		+	
Catalan	+	−		+	+		+	−
French	+	−		+	+		+	
Hebrew	+	−		+	+		+	
Italian		−			+	−	+	
Portuguese	+	−			+		+	−
Romanian	+	−		+	+		+	
Spanish	+	−		+	+		+	−

Table 5: Matrix showing which cognitive predictors are used by the best-performing ensemble per template and per language. Results by template are for the within-language evaluation described in the text.

across languages and templates. For the best single predictors, we see similarity within templates: For AAN, subjectivity is most often the best single predictor; for ANA, the best single predictor is length; and for NAA, PMI is the best single predictor. In ensembles, as shown in Table 5, the predictors which most consistently emerge as informative are PMI, subjectivity, and length. PMI is a significant predictor across all languages, with signs that accord with its cognitive justification. The next most robust predictors are subjectivity (again with consistent signs) and length, although for length the signs in the NAA template are contrary to what would be expected from the cognitive motivation—under accessibility-based accounts, short words should generally go before

long words, which would predict a negative sign across all languages. In fact, no predictor shows the same sign across all templates.

5.1 Discussion

The results show that cognitive predictors of adjective order have broad crosslinguistic validity, and furthermore reveal intriguingly consistent patterns in terms of which predictors appear to be informative across languages and templates. For example, for the best-single-predictor results, we saw that subjectivity performs consistently best in AAN languages, length performs best in ANA, and PMI performs best in NAA. This regularity is unlikely to hold by chance.[5]

The literature on predictors provides some clues about why these patterns may arise. For subjectivity, several accounts attribute its role in adjective ordering to successful referential communication: Ordering adjectives with respect to subjectivity maximizes the chances that a listener will arrive at the intended referent (e.g., identifying the correct box when hearing *the small brown box*; for details, see Scontras et al., 2019; Franke et al., 2019; Scontras et al., 2020). We see that subjectivity consistently performs as the best single predictor in AAN—but not ANA or NAA—languages. There is independent evidence to support the idea that AAN languages are more likely to use adjectives for the purpose of establishing reference (e.g., singling out a specific box among a set of boxes)—as opposed to, say, commenting on speaker judgments on objects in common ground (e.g., commenting on the size or color of the unique box in a communicative context; Hahn et al., 2018). Rubio-Fernández (2016) argues that pre-nominal adjectives are more useful for incrementally establishing nominal reference than post-nominal adjectives: Hearing *small* and *brown* before *box* helps a listener narrow in on the potential referents before they reach the noun; encountering the adjectives after the noun is less useful for this purpose (see also Kachakeche et al., 2021). Perhaps AAN languages are more likely to use adjectives for the purpose of establishing reference, which is why subjectivity plays such a prominent role in predicting adjective order in these languages.

In NAA languages where PMI outperforms the other predictors, pressures from successful referential communication may be less strong, given the communicative role of adjectives post-nominally. In other words, it may be the case that adjectives in NAA languages are less likely to be used for the purpose of establishing reference. As a result, meaning-based predictors like subjectivity (and also information gain, as seen in Table 5) play less of a role in adjective ordering post-nominally. With meaning-based pressures less relevant, production pressures like PMI stand out; supporting this idea that production pressures play a larger role post-nominally in the absence of meaning-based pressures, we also see an increased role for adjective frequency in the ensemble models for NAA languages (Table 5).

For ANA languages, particularly Romance languages like Spanish or French, the set of adjectives that occur in pre-nominal position are often reduced versions of post-nominal adjectives (e.g., *gran* vs. *grande* in Spanish; Butt et al., 2018). If it is the case that ANA languages allow only a restricted set in pre-nominal position, and pre-nominal adjectives are often shortened, it should come as no surprise that length should perform well as a predictor. Indeed, the cognitive ensembles perform nearly as well as the distributional baseline for ANA templates in many languages.

The results also raise some questions: For example, although length is fairly consistent as a predictor across languages, its sign is not consistent with its cognitive motivation based on accessibility. One alternative explanation for a length effect, which would predict the sign pattern found in Table 5, is that speakers may prefer to put shorter adjectives farther from the noun in order to minimize dependency length between adjectives and nouns (Dyer, 2017; Temperley and Gildea, 2018; Liu et al., 2017; Futrell et al., 2020c), with dependency length crucially measured in terms of the phonetic lengths of intervening words, rather than in terms of the number of intervening words.

One predictor which is surprisingly non-robust across languages is frequency, which only participates in the best-performing ensemble in the NAA template. The reason for this non-robustness

[5]To test this claim statistically, we performed a permutation test with 100,000 samples for the hypothesis that subjectivity appears as the best single predictor for $> 75\%$ of AAN languages, length for 100% of ANA languages, and PMI for 100% of NAA languages. The test was performed by permuting the predictors in the single-predictor column of Table 4. We find $p < 0.001$.

	PMI	IC	IG	FREQ	LENGTH	CDV	VCosD	SUBJ
AAN								
PMI		−0.33	−0.18	−0.26	0.14	0.07	0.01	−0.06
IC	−0.33		**0.63**	**0.79**	−0.27	−0.17	−0.02	0.10
IG	−0.18	**0.63**		**0.70**	−0.29	−0.09	−0.03	0.03
FREQ	−0.26	**0.79**	**0.70**		−0.24	−0.02	−0.04	−0.01
LENGTH	0.14	−0.27	−0.29	−0.24		0.06	−0.01	−0.16
CDV	0.07	−0.17	−0.09	−0.02	0.06		0.01	−0.21
VCosD	0.01	−0.02	−0.03	−0.04	−0.01	0.01		−0.01
SUBJ	−0.06	0.10	0.03	−0.01	−0.16	−0.21	−0.01	
ANA								
PMI		−0.24	−0.12	−0.14	0.19	0.12	−0.02	−0.12
IC	−0.24		**0.71**	**0.79**	−0.32	−0.33	−0.04	0.21
IG	−0.12	**0.71**		**0.79**	−0.34	−0.29	−0.05	0.15
FREQ	−0.14	**0.79**	**0.79**		−0.23	−0.18	−0.07	0.08
LENGTH	0.19	−0.32	−0.34	−0.23		0.24	0.04	−0.31
CDV	0.12	−0.33	−0.29	−0.18	0.24		0.03	−0.34
VCosD	−0.02	−0.04	−0.05	−0.07	0.04	0.03		−0.01
SUBJ	−0.12	0.21	0.15	0.08	−0.31	−0.34	−0.01	
NAA								
PMI		−0.20	0.02	−0.09	0.10	0.05	−0.03	0.08
IC	−0.20		0.02	**0.77**	−0.23	−0.18	0.00	0.02
IG	0.02	0.02		0.01	−0.02	0.02	−0.04	0.01
FREQ	−0.09	**0.77**	0.01		−0.16	−0.07	−0.00	0.00
LENGTH	0.10	−0.23	−0.02	−0.16		−0.01	0.00	−0.08
CDV	0.05	−0.18	0.02	−0.07	−0.01		−0.03	−0.16
VCosD	−0.03	0.00	−0.04	−0.00	0.00	−0.03		−0.00
SUBJ	0.08	0.02	0.01	0.00	−0.08	−0.16	−0.00	

Table 6: Correlation matrices for individual predictors grouped by template. Correlations above 0.5 and below −0.5 are shown in bold.

could be that frequency is correlated with other factors, such as length (Zipf, 1935). Table 6 shows correlations among predictors in what predictions they make about triple order. Given the conceptual relatedness of many of the predictors, they are not as correlated as might be expected; nonetheless, IC, IG, and Frequency are strongly correlated and so may reflect a single underlying factor.

5.2 Closing the Gap

Here we discuss what it would take to improve cognitive theories so that they capture more about adjective order, closing the gap between the cognitive predictors and the distributional baseline, especially for AAN and NAA templates. There are (at least) three possible explanations for this gap: (i) cognitive predictors do not provide an informative enough feature set for prediction of adjective order; (ii) the estimates of the cognitive predictors based on our training pairs could be improved; or (iii) the cognitive predictors have enough information, but they need to be combined together in a different way (other than logistic regression).

To evaluate the last possibility, we trained a feedforward neural network with one hidden layer to predict adjective order given the cognitive features as input. The neural network allows for nonlinear interactions among cognitive features. We also trained a classifier based on linear word embedding features, creating a linear form of the distributional baseline. The resulting set of models lets us determine whether the discrepancy between the cognitive predictors and the distributional baseline is due to nonlinear interactions among features or the information represented by the features themselves. That is, if the neural network models perform better, the discrepancy is due to feature interactions; if the embedding-based models perform better even using a linear classifier, the discrepancy is due to the features.

These classifier accuracies are shown in Figure 1. We find that model architecture (DNN

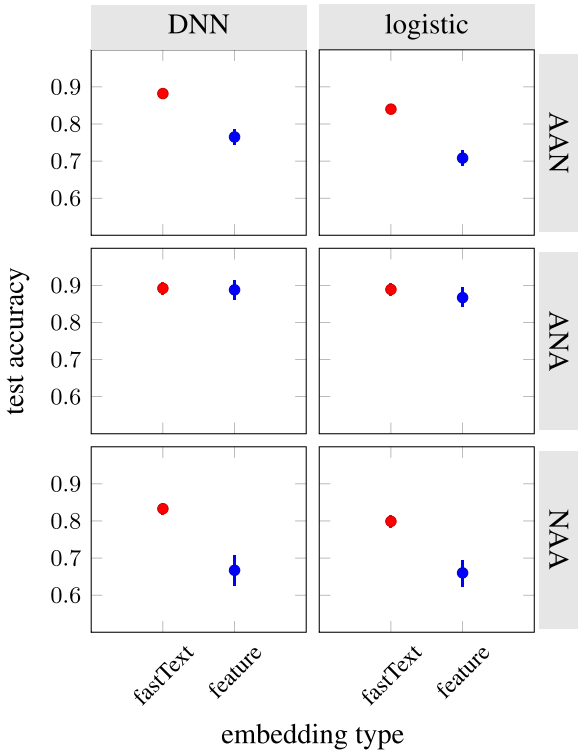

Figure 1: A comparison of accuracy between fastText and cognitive feature-based embeddings.

vs. logistic) has little effect on accuracy, suggesting that the distributional baseline's performance is more due to information in its embeddings than it is to nonlinear interactions among features.

This result has consequences for cognitive theories of adjective order: it suggests that progress will not be made by combining existing features in new ways, but rather by coming up with new features that better reflect the kind of information contained in word embeddings.

6 Conclusion

Adjective order is an important object of study because it appears to be in many ways universal across languages, and thus offers a test bed for understanding how universal properties of human cognition have shaped language. Our results reveal that cognitive theories have made real progress in explaining adjective order across languages: Despite these theories being formulated primarily based on the analysis of English, they do yield predictors with fairly consistent crosslinguistic validity. Nevertheless, considerable variance remains unexplained.

Our approach also shows that a massively cross-linguistic approach to comparing and combining cognitive theories is now possible, and we believe this style of approach offers a meaningful way forward for the development and evaluation of future theories of how cognition shapes language.

References

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1193, Cambridge, MA. Association for Computational Linguistics.

Otto Behaghel. 1909. Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern. *Indogermanische Forschungen*, 25:110–142.

J. Kathryn Bock. 1982. Toward a cognitive psychology of syntax: Information processing contributions to sentence formulation. *Psychological Review*, 89:1–47. https://doi.org/10.1037/0033-295X.89.1.1

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146. https://doi.org/10.1162/tacl_a_00051

Denis Bouchard. 2005. Sériation des adjectifs dans le SN et formation de concepts. *Recherches linguistiques de Vicennes*, 34:125–142. https://doi.org/10.4000/rlv.1383

Joan Bresnan, Anna Cueni, Tatiana Nikitina, and Harald Baayen. 2007. Predicting the dative alternation. In *Cognitive Foundations of Interpration*, pages 69–94. Royal Netherlands Academy of Science, Amsterdam.

John B. Butt, Carmen Benjamin, and Moreira-Rodriguez Antonia. 2018. *A New Reference Grammar of Modern Spanish*. Routledge. https://doi.org/10.4324/9781315648446

B. Byrne. 1979. Rules of prenominal adjective order and the interpretation of ''incompatible'' adjective pairs. *Journal of Verbal*

Learning and Verbal Behavior, 18(1):73–78. https://doi.org/10.1016/S0022-5371(79)90574-7

Franklin Chang. 2009. Learning to order words: A connectionist model of Heavy NP Shift and accessibility effects in Japanese and English. *Journal of Memory and Language*, 61:374–397. https://doi.org/10.1016/j.jml.2009.07.006

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

William Dyer. 2018. Integration complexity and the order of constituents. In *Proceedings of the Second Workshop on Universal Dependencies*, pages 55–65. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6007

William Dyer, Richard Futrell, Zoey Liu, and Gregory Scontras. 2021. Predicting cross-linguistic adjective order with information gain. In *Findings of the Association for Computational Linguistics*, pages 957–967. https://doi.org/10.18653/v1/2021.findings-acl.83

William Edward Dyer. 2017. *Minimizing Integration Cost: A General Theory of Constituent Order*. Ph.D. thesis, University of California, Davis.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Towards understanding linear word analogies. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3253–3262, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1315

Robert M. Fano. 1961. *Transmission of Information: A Statistical Theory of Communication*. MIT Press, Cambridge, MA.

Victor S. Ferreira and Gary S. Dell. 2000. Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive Psychology*, 40:296–340. https://doi.org/10.1006/cogp.1999.0730, PubMed: 10888342

John Rupert Firth. 1957. A synopsis of linguistic theory 1930–3195. In *Studies in Linguistic Analysis*, pages 1–32. Philological Society, Oxford.

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

Richard Futrell. 2019. Information-theoretic locality properties of natural language. In *Proceedings of the First Workshop on Quantitative Syntax (Quasy, SyntaxFest 2019)*, pages 2–15, Paris, France. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-7902

Richard Futrell, William Dyer, and Gregory Scontras. 2020a. What determines the order of adjectives in English? Comparing efficiency-based theories using dependency treebanks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2003–2012. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.181

Richard Futrell, Edward Gibson, and Roger P. Levy. 2020b. Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, 44:e12814. https://doi.org/10.1111/cogs.12814, PubMed: 32100918

Richard Futrell, Roger P. Levy, and Edward Gibson. 2020c. Dependency locality as an explanatory principle for word order. *Language*, 96(2):371–413. https://doi.org/10.1353/lan.2020.0024

Michael Hahn, Judith Degen, Noah D. Goodman, Dan Jurafsky, and Richard Futrell. 2018. An information-theoretic explanation of adjective ordering preferences. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society*, pages 1766–1771.

R. Hetzron. 1978. On the relative order of adjectives. In H. Seiler, editor, *Language Universals*, pages 165–184. Narr, Tübingen.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2979–2984, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1330

Zeinab Kachakeche, Richard Futrell, and Gregory Scontras. 2021. Word order affects the frequency of adjective use across languages. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 43:3006–3012.

Zeinab Kachakeche and Gregory Scontras. 2020. Adjective ordering in Arabic: Post-nominal structure and subjectivity-based preferences. *Proceedings of the Linguistic Society of America*, 1:419–430. https://doi.org/10.3765/plsa.v5i1.4726

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Max Kölbel. 2004. Faultless disagreement. *Proceedings of the Aristotelian Society*, 104:53–73. https://doi.org/10.1111/j.0066-7373.2004.00081.x

Sven Kotowski and Holden Härtl. 2019. How real are adjective ordering constraints? Multiple prenominal adjectives at the grammatical interfaces. *Linguistics*, 57(2):395–427. https://doi.org/10.1515/ling-2019-0005

Alessandro Lenci, Magnus Sahlgren, Patrick Jeuniaux, Amaru Cuba Gyllensten, and Martina Miliani. 2022. A comparative evaluation and analysis of three generations of distributional semantic models. *Language resources and evaluation*, 56(4):1269–1313. https://doi.org/10.1007/s10579-021-09575-z

Jun Yen Leung, Guy Emerson, and Ryan Cotterell. 2020. Investigating cross-linguistic adjective ordering tendencies with a latent-variable model. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/2020.emnlp-main.329

Haitao Liu, Chunshan Xu, and Junying Liang. 2017. Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*, 21:171–193. https://doi.org/10.1016/j.plrev.2017.03.002, PubMed: 28624589

Nikola Ljubešić, Darja Fišer, and Anita Peti-Stantić. 2018. Predicting concreteness and imageability of words within and across languages via word embeddings. In *Proceedings of the Third Workshop on Representation Learning for NLP*, pages 217–222, Melbourne, Australia. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-3028

J. MacFarlane. 2014. *Assessment Sensitivity*. Clarendon Press, Oxford. https://doi.org/10.1093/acprof:oso/9780199682751.001.0001

Robert Malouf. 2000. The order of prenominal adjectives in natural language generation. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 85–92. https://doi.org/10.3115/1075218.1075230

J. E. Martin. 1969. Semantic determinants of preferred adjective order. *Journal of Verbal Learning and Verbal Behavior*, 8:697–704. https://doi.org/10.1016/S0022-5371(69)80032-0

Louise McNally and Gemma Boleda. 2004. Relational adjectives as properties of kinds. *Empirical Issues in Formal Syntax and Semantics*, 5:179–196.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, and others. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1659–1666.

Denis Paperno and Marco Baroni. 2016. Squibs: When the whole is less than the sum of its parts: How composition affects PMI values in distributional semantic vectors. *Computational Linguistics*, 42(2):345–350. https://doi.org/10.1162/COLI_a_00250

Ranfolph Quirk, Signey Greenbaum, Geoffrey Leech, and Jan Svartik. 1972. *A Grammar of Contemporary English*. Longman, London.

Cesar Manuel Rosales Jr. and Gregory Scontras. 2019. On the role of conjunction in adjective

ordering preferences. *Proceedings of the Linguistic Society of America*, 4(32):1–12. https://doi.org/10.3765/plsa.v4i1.4524

Paula Rubio-Fernández. 2016. How redundant are redundant color adjectives? An efficiency-based analysis of color overspecification. *Frontiers in Psychology*, 7:153. https://doi.org/10.3389/fpsyg.2016.00153, PubMed: 26924999

Suttera Samonte and Gregory Scontras. 2019. Adjective ordering in Tagalog: A cross-linguistic comparison of subjectivity-based preferences. *Proceedings of the Linguistic Society of America*, 4(33):1–13. https://doi.org/10.3765/plsa.v4i1.4511

Gregory Scontras. 2023. Adjective ordering across languages. *Annual Review of Linguistics*, 9. https://doi.org/10.1146/annurev-linguistics-030521-041835

Gregory Scontras, Galia Bar-Sever, Zeinab Kachakeche, Cesar Manuel Rosales Jr., and Suttera Samonte. 2020. Incremental semantic restriction and subjectivity–based adjective ordering. *Proceedings of Sinn und Bedeutung 24*, 24(2):253–270.

Gregory Scontras, Judith Degen, and Noah D. Goodman. 2017. Subjectivity predicts adjective ordering preferences. *Open Mind: Discoveries in Cognitive Science*, 1(1):53–65. https://doi.org/10.1162/OPMI_a_00005

Gregory Scontras, Judith Degen, and Noah D. Goodman. 2019. On the grammatical source of adjective ordering preferences. *Semantics and Pragmatics*. https://doi.org/10.3765/sp.12.7

Gregory Scontras, Zeinab Kachakeche, Austin Nguyen, Cesar Rosales, Suttera Samonte, Einat Shetreet, Yuxin Shi, Elli Tourtouri, and Nitzan Trainin. 2021. Cross-linguistic evidence for subjectivity-based adjective ordering preferences. Talk presented at Theoretical and Experimental Approaches to Modification.

Hansjakob Seiler. 1978. Determination: A functional dimension for interlanguage comparison. In Hansjakob Seiler, editor, *Language Universals, Papers from the Conference held at Gummersbach/Cologne, Germany, October 3–8, 1976*, pages 301–328. Narr, Tübingen.

Yuxin Shi and Gregory Scontras. 2020. Mandarin has subjectivity-based adjective ordering preferences in the presence of *de*. *Proceedings of the Linguistic Society of America*, 5(1):410–418. https://doi.org/10.3765/plsa.v5i1.4711

Lynne M. Stallings, Maryellen C. MacDonald, and Padraig G. O'Seaghdha. 1998. Phrasal ordering constraints in sentence production: Phrase length and verb disposition in heavy-NP shift. *Journal of Memory and Language*, 39(3):392–417. https://doi.org/10.1006/jmla.1998.2586

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

Henry Sweet. 1898. *A New English Grammar, Logical and Historical*, volume 2: Syntax, Clarendon Press, London. Republised by Cambridge University Press in 2014.

Duyu Tang, Furu Wei, Bing Qin, Ming Zhou, and Ting Liu. 2014. Building large-scale Twitter-specific sentiment lexicon: A representation learning approach. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 172–182, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

David Temperley and Daniel Gildea. 2018. Minimizing syntactic dependency lengths: Typological/cognitive universal? *Annual Review of Linguistics*, 4:1–15. https://doi.org/10.1146/annurev-linguistics-011817-045617

Andreas Trotzke and Eva Wittenberg. 2019. Long-standing issues in adjective order and corpus evidence for a multifactorial approach. *Linguistics*, 57(2):273–282. https://doi.org/10.1515/ling-2019-0001

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*,

pages 248–258, Baltimore, Maryland. Association for Computational Linguistics. https://doi.org/10.3115/v1/P14-1024

Chris Westbury. 2021. Prenominal adjective order is such a fat big deal because adjectives are ordered by likely need. *Psychonomic Bulletin & Review*, 28:122–138. https://doi.org/10.3758/s13423-020-01769-w, PubMed: 32700119

Stefanie Wulff. 2003. A multifactorial corpus analysis of adjective order in English. *International Journal of Corpus Linguistics*, 8(2):245–282. https://doi.org/10.1075/ijcl.8.2.04wul

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič Jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19. Association for Computational Linguistics. https://doi.org/10.18653/v1/K17-3001

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

George Kingsley Zipf. 1935. *The Psycho-biology of Language*. Houghton, Mifflin.