What Determines the Order of Adjectives in English? Comparing Efficiency-Based Theories Using Dependency Treebanks

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

William Dyer
Oracle Corporation
william.dyer@oracle.com

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

Abstract

We take up the scientific question of what determines the preferred order of adjectives in English, in phrases such as *big blue box* where multiple adjectives modify a following noun. We implement and test four quantitative theories, all of which are theoretically motivated in terms of efficiency in human language production and comprehension. The four theories we test are subjectivity (Scontras et al., 2017), information locality (Futrell, 2019), integration cost (Dyer, 2017), and information gain, which we introduce. We evaluate theories based on their ability to predict orders of unseen adjectives in hand-parsed and automatically-parsed dependency treebanks. We find that subjectivity, information locality, and information gain are all strong predictors, with some evidence for a two-factor account, where subjectivity and information gain reflect a factor involving semantics, and information locality reflects collocational preferences.

1 Introduction

Across languages, there exist strong and stable constraints on the order of adjectives when more than one is used to modify a noun (Dixon, 1982; Sproat and Shih, 1991). For example, in English, *big blue box* sounds natural and appears relatively frequently in corpora, while *blue big box* sounds less natural and occurs less frequently (Scontras et al., 2017). In this paper, we take up the scientific question of what explains these constraints in natural language. To do so, we implement quantitative models that have been proposed in previous literature as explanations for these constraints, and compare their accuracy in predicting adjective ordering data in parsed corpora of English[1].

In the last few years, adjective order has become a crucial testing ground for quantitative theories of syntax. These theories provide mathematical models that can describe the distribution of words in sentences and the way those words combine to yield the meaning of a sentence, in a way that captures the fine-grained quantitative patterns observable in large text datasets (Manning, 2003; Bresnan et al., 2007; Chen and Ferrer-i-Cancho, 2019).

Quantitative syntactic theories are often **efficiency-based**, meaning that they model word distributions as the result of a process that tries to maximize information transfer while minimizing some measure of cognitive cost; as a result, they often use the mathematical language of information theory. Such theories promise not only to describe distributions of words, but also to *explain* why they take the shape they do, by viewing human language as an efficient code subject to appropriate constraints. This work informs NLP by providing a theory of language structure that integrates with data-driven, optimization-based machine learning models.

Adjective order is a fruitful empirical target for quantitative theories of syntax because it is an area where the traditional discrete and symbolic theories become highly complex, and a quantitative approach becomes more attractive. For example, in the formal syntax literature, a standard explanation for adjective order constraints is that each adjective belongs to a certain semantic class (e.g., COLOR or SIZE) and that there exists a universal total order on these semantic classes (e.g., COLOR $<$ SIZE) shared among all languages, which determines the order of adjectives in any given instance (Cinque, 1994; Scott, 2002). Such discrete theories of adjective order become complex rapidly as the number of semantic classes to be posited becomes large (upwards of twelve in Scontras et al. 2017) and more fine-grained (see Bar-Sever et al. 2018 for discussion of the learning problem posed by such classifications).

[1] All code and data are available at https://github.com/langprocgroup/adjorder.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 2003–2012

In contrast, quantitative syntax theories typically identify a single construct that grounds out in real-valued numerical scores given to adjectives, which determine their ordering preferences. These scores can be estimated based on large-scale corpus data or based on human ratings. In what follows, we test the predictions of four such theories: the subjectivity hypothesis (Scontras et al., 2017; Simonič, 2018; Hahn et al., 2018; Franke et al., 2019; Scontras et al., 2019), the information locality hypothesis (Futrell and Levy, 2017; Futrell et al., 2017; Hahn et al., 2018; Futrell, 2019), the integration cost hypothesis (Dyer, 2017), and the information gain hypothesis, which we introduce.

We begin with a presentation of the details of each theory, then implement the theories and test their predictions against large-scale naturalistic data from English. In addition to comparing the predictors in terms of accuracy, we also perform a number of analyses to determine the important similarities and differences among their predictions. The paper concludes with a discussion of what our results tell us about adjective order and related issues, and a look towards future work.

2 Theories of adjective order

2.1 Subjectivity

Scontras et al. (2017) show that adjective order is strongly predicted by adjectives' **subjectivity scores**: an average rating obtained by asking human participants to rate adjectives on a numerical scale for how subjective they are. Adjectives that are rated as more subjective typically appear farther from the noun than adjectives rated as less subjective, and the strength of ordering preferences tracks the subjectivity differential between two adjectives. For example, in *big blue box*, the adjective *big* has a subjectivity rating of 0.64 (out of 1), and the adjective *blue* has a subjectivity rating of 0.30. If adjectives are placed in order of decreasing subjectivity, then *big* must appear before *blue*, corresponding to the preferred order. The notion of subjectivity as a predictor of adjective order was previously introduced by Hetzron (1978).

Subsequent work has attempted to explain the role of subjectivity in adjective ordering by appealing to the communicative benefit afforded by ordering adjectives with respect to decreasing subjectivity. For example, Franke et al. (2019) use simulated reference games to demonstrate that, given a set of independently-motivated assumptions concerning the composition of meaning in multi-adjective strings, subjectivity-based orderings lead to a greater probability of successful reference resolution; the authors thus offer an evolutionary explanation for the role of subjectivity in adjective ordering (see also Simonič, 2018; Hahn et al., 2018; Scontras et al., 2019).

2.2 Information locality

The theory of **information locality** holds that words that have high mutual information are under pressure to be close to each other in linear order (Futrell and Levy, 2017; Futrell et al., 2017). Information locality is a generalization of the well-supported principle of **dependency length minimization** (Liu et al., 2017; Temperley and Gildea, 2018). In the case of adjective ordering, the prediction is simply that adjectives that have high **pointwise mutual information** (PMI) with their head noun will tend to be closer to that noun. The PMI of an adjective a and a noun n is (Fano, 1961; Church and Hanks, 1990):

$$\text{PMI}(a:n) \equiv \log \frac{p(a,n)}{p(a)p(n)}. \quad (1)$$

In this paper, we take the relevant joint distribution $p(a,n)$ to be the distribution of adjectives and nouns in a dependency relationship, with the marginals calculated as $p(a) = \sum_n p(a,n)$ and $p(n) = \sum_a p(a,n)$.

Information locality is motivated as a consequence of a more general theory of efficiency in human language. In this theory, languages should maximize information transfer while minimizing cognitive information-processing costs associated with language production and comprehension. Information locality emerges from these theories when we assume that the relevant measure of information-processing cost is the surprisal of words given lossy memory representations (Hale, 2001; Levy, 2008; Smith and Levy, 2013; Futrell and Levy, 2017; Futrell, 2019).

2.3 Integration Cost

The theory of integration cost is also based in the idea of efficiency with regard to information-processing costs. It differs from information locality in that it assumes that the correct metric of processing difficulty for a word w is the **entropy**

over the possible heads of w:

$$\begin{aligned}\text{Cost}(w) &\propto H[T|w] \\ &= \sum_t -p_T(t|w) \log p_T(t|w), \end{aligned} \quad (2)$$

where T is a random variable indicating the head t of the word w (Dyer, 2017). This notion of cost captures the amount of uncertainty that has to be resolved about the proper role of the word w with respect to the rest of the words in the sentence. Like information locality, the theory of integration cost recovers dependency length minimization as a special case. For the case of predicting adjective order, the prediction is that an adjective a will be closer to a noun when it has lower integration cost:

$$\text{IC}(a) = H[N|a], \quad (3)$$

where N is a random variable ranging over nouns.

Integration cost corresponds to an intuitive idea previously articulated in the adjective ordering literature. The idea is that adjectives that can modify a smaller set of nouns appear closer to the noun: for example, an order such as *big wooden spoon* is preferred over *wooden big spoon* because the word *big* can modify nearly any noun, while *wooden* can only plausibly modify a small set of nouns (Ziff, 1960). The connection between integration cost and set size comes from the information-theoretic notion of the **typical set** (Cover and Thomas, 2006, pp. 57–71); the entropy of a random variable can be interpreted as the (log) cardinality of the typical set of samples from that variable. When we order adjectives by integration cost, this is equivalent to ordering them such that adjectives that can modify a larger typical set of nouns appear farther from the noun. The result is that each adjective gradually reduces the entropy of the possible nouns to follow, thus avoiding information-processing costs that may be associated with entropy reduction (Hale, 2006, 2016; Dye et al., 2018).

2.4 Information gain

We propose a new efficiency-based predictor of adjective order: information gain. The idea is to view the noun phrase, consisting of prenominal adjectives followed by the noun, as a **decision tree** for identifying a referent, where each word partitions the space of possible referents. Each partitioning is associated with some information gain, indicating how much the set of possible referents shrinks. In line with the logic for integration cost, we propose that the word with *smaller* information gain will be placed earlier, so that the set of referents is gradually narrowed by each word.

As generally implemented in decision trees, information gain refers to the reduction of entropy obtained from partitioning a set on a feature (Quinlan, 1986). In our case, the distribution of nouns N is partitioned on a given adjective a, creating two partitions: N_a and its complement ${N_a}^c$. The difference between the starting entropy $H[N]$ and the sum of the entropy of each partition, conditioned on the size of that partition, is the information gain of a:

$$\begin{aligned}\text{IG}(a) &= H[N] \\ &- \left[\frac{|N_a|}{|N|} H[N_a] + \frac{|{N_a}^c|}{|N|} H[{N_a}^c]\right]. \end{aligned} \quad (4)$$

Information gain is therefore comprised of both positive and negative evidence. That is, specifying an adjective such as *big* partitions the probability distribution of nouns into N_{big}, the subset of N which takes *big* as a dependent, and ${N_{big}}^C$, the subset of N which does not.

Crucially, $H[N_a]$ is not $H[N|a]$ in general. $H[N|a]$ is the conditional entropy of nouns given a specific adjective, while $H[N_a]$ is the entropy of a distribution over nouns whose support is limited to noun types that have been observed to occur with an adjective a. Combined with $H[{N_a}^c]$, information gain tells us how much the entropy of N is reduced by partitioning on a. This means that information gain and integration cost, while conceptually similar, are not mathematically equivalent.

To our knowledge, information gain has not been previously suggested as a predictor of adjective ordering, although Danks and Glucksberg (1971) expressed a similar intuition in proposing that adjectives are ordered according to their 'discriminative potential'. Although decision-tree algorithms such as ID3 choose the highest-IG feature first, we predict that the lower-information-gain adjective will precede the higher one.

3 Related Work

Previous corpus studies of adjective order include Malouf (2000), who examined methods for ordering adjectives in a natural language generation context, and Wulff (2003), who examined effects of phonological length, syntactic category ambiguity, semantic closeness, adjective frequency, and

a measure similar to PMI called noun specificity. Our work differs from this previous work by focusing on recently-introduced predictors that have theoretical motivations grounded in efficiency and information theory.

The theories we test here (except information gain) have been tested in previous corpus studies, but never compared against each other. Scontras et al. (2017) validate that subjectivity is a good predictor of adjective order in corpora, and Hahn et al. (2018) and Futrell et al. (2019) evaluate both information locality and subjectivity. Dyer (2018) uses integration cost to model the order of same-side sibling dependents cross-linguistically and across all syntactic categories.

4 Methods

Our task is to find predictors of adjective order based solely on data about individual adjectives and nouns. More formally, the goal is to find a scoring function $S(A, N)$ applying to an adjective A and a noun N, such that the order of *two* adjectives modifying a noun A_1A_2N can be predicted accurately by comparing $S(A_1, N)$ and $S(A_2, N)$. Furthermore, the scoring function S should not include information about relative order in observed sequences of the form A_1A_2N—the scoring function should be based only on corpus data about co-occurrences of A and N, or on human ratings about A and/or N. We apply this restriction because our goal is to evaluate scientific theories of *why* adjectives are ordered the way they are, rather than to achieve maximal raw accuracy.

4.1 Data sources

Corpus-based predictors We estimate information-theoretic quantities for adjectives using a large automatically-parsed subsection of the English Common Crawl corpus (Buck et al., 2014; Futrell et al., 2019). The use of a parsed corpus is necessary to identify adjectives that are dependents of nouns in order to calculate PMI and IC. As described in Futrell et al. (2019), this corpus was produced by heuristically filtering Common Crawl to contain only full sentences and to remove web boilerplate text, and then parsing the resulting text using SyntaxNet (Andor et al., 2016), obtaining a total of ∼1 billion tokens of automatically parsed web text. In this work, we use a subset of this corpus, described below.

From this corpus, we extract two forms of data. First, we extract **adjective–noun (AN) pairs**: a set of pairs $\langle A, N\rangle$ where A is an adjective and N is a noun and N is the head of A with dependency type *amod*. As in Futrell (2019), we define A as an adjective iff its part-of-speech is `JJ` and its wordform is listed as an adjective in the English CELEX database (Baayen et al., 1995). We define N as a noun iff its part-of-speech is `NN` or `NNS` and its wordform is listed as a noun in CELEX. These AN pairs are used to estimate the information-theoretic predictors that we are interested in. We extracted 33,210,207 adjective–noun pairs from the parsed Common Crawl corpus.

Second, we extract **adjective–adjective–noun (AAN) triples**: a set of triples $\langle A_1, A_2, N\rangle$ where A_1 and A_2 are adjectives as defined above, and A_1 and A_2 are both adjective dependents with relation type *amod* of a single noun head N. Furthermore, A_1 and A_2 must not have any further dependents, and they must appear in the order A_1A_2N in the corpus with no intervening words. We extracted a total of 842,714 AAN triples from the parsed Common Crawl corpus.

The values of all corpus-based predictors are estimated using the AN pairs. The AAN triples are used only for fitting regressions from the predictors to adjective orders, and for evaluation.

Ratings-based predictors We gathered subjectivity ratings for all 398 adjectives present in AAN triples in the English UD corpus. These subjectivity ratings were collected over Amazon.com's Mechanical Turk, using the methodology of Scontras et al. (2017). 264 English-speaking participants indicated the subjectivity of 30 random adjectives by adjusting a slider between endpoints labeled "completely objective" (coded as 0) and "completely subjective" (coded as 1). Each adjective received an average of 20 ratings.

Test set As a held-out test set for our predictors, we use the English Web Treebank (EWT), a hand-parsed corpus, as contained in Universal Dependencies (UD) v2.4 (Silveira et al., 2014; Nivre, 2015). Following our criteria, we extract 155 AAN triples having scores for all our predictors. Because this test set is very small, we also evaluate against a held-out portion of the parsed Common Crawl data. In the Common Crawl test set, after including only AAN triples that have scores for all of our predictors, we have 41,822 AAN triples.

4.2 Estimation of predictors

Our information-theoretic predictors require estimates of probability distributions over adjectives and nouns. To estimate these probability distributions, we first use maximum likelihood estimation as applied to counts of wordforms in AN pairs. We call these estimates **wordform estimates**.

Although maximum likelihood estimation is sufficient to give an estimate of the general entropy of words (Bentz et al., 2017), it is not yet clear that it gives a good measure for conditional entropy or mutual information, due to data sparsity, even with millions of tokens of text (Futrell et al., 2019).

Therefore, as a second method that alleviates the data sparsity issue, we also calculate our probability distributions not over raw wordforms but over clusterings of words in an embedding space, a method which showed promise in Futrell et al. (2019). To derive word clusters, we use `sklearn.cluster.KMeans` applied to a pretrained set of 1.9 million 300-dimension GloVe vectors[2] generated from the Common Crawl corpus (Pennington et al., 2014). We classify adjectives into $k_A = 300$ clusters and nouns into $k_N = 1000$ clusters. These numbers k were found by choosing the largest k multiple of 100 that did not result in any singleton clusters. We then estimated probabilities $p(a, n)$ by maximum likelihood estimation after replacing wordforms a and n with their cluster indices.

This clustering method alleviates data sparsity by reducing the size of the support of the distributions over adjectives and nouns, to k_A and k_N respectively, and by effectively spreading probability mass among words with similar semantics. The clusters might also end up recapitulating the semantic categories that have played a role in more traditional syntactic theories of adjective order (Dixon, 1982; Cinque, 1994; Scott, 2002). We call these estimates **cluster estimates**.

4.3 Evaluation

Fitting predictors to data Most of our individual predictors come along with theories that say what their effect on adjective order should be. Adjectives with low PMI should be farther from the noun, adjectives with high IC should be farther from the noun, and adjectives with high subjectivity should be farther from the noun. Therefore, strictly speaking, it is not necessary to fit these predictors to any training data: we can evaluate our theories based on their *a priori* predictions simply by asking how accurately we can predict the order of adjectives in AAN triples based on the rules above.

However, we can get a deeper picture of the performance of our predictors by using them in classifiers for adjective order. By fitting classifiers using our predictors, we can easily extend our models to ones with multiple predictors, in order to determine if a combined set of the predictors gives increased accuracy over any one.

Logistic regression method We fit logistic regressions to predict adjective order in AAN triples using our predictors. Our goal is to predict the order of the triple from the *unordered set* of the two adjectives $\{A_1, A_2\}$ and the noun N. To do so, we consider the adjectives in lexicographic order: Given an AAN triple, let A^1 denote the lexicographically-first adjective, and A^2 the second. Then any given AAN triple is either of the form $\langle A^1, A^2, N\rangle$ or $\langle A^2, A^1, N\rangle$. We fit a logistic regression to predict this order given the *difference* in the values of the predictors for the two adjectives. That is, we fit a logistic regression of the form in Figure 1. This method of fitting a classifier to predict order data was used previously in Morgan and Levy (2016). Based on theoretical considerations and previous empirical results, we expect that the fitted values of β_1 will be negative for PMI and positive for IC and subjectivity. The regression in Figure 1 can easily be extended to include multiple predictors, with a separate β for each.

Evaluation metrics We evaluate our models using raw accuracy in predicting the order of held-out AAN triples. We also calculate 95% confidence intervals on these accuracies, indicating our uncertainty about how the accuracy would change in repeated experiments. Following standard experimental practice, if we find that two predictors achieve different accuracies, but their confidence intervals overlap, then we conclude that we do not have evidence that their accuracies are reliably different. We say a difference in accuracy between predictors is **significant** if the 95% confidence intervals do not overlap.

Evaluation on held-out hand-parsed data It is crucial that we not evaluate solely on automatically-parsed data. The reason is that both

[2] `http://nlp.stanford.edu/data/glove.42B.300d.zip`

$$\log \frac{p(\langle A^1, A^2, N\rangle)}{p(\langle A^2, A^1, N\rangle)} = \beta_0 + \beta_1(S(A^1, N) - S(A^2, N)) + \epsilon$$

Figure 1: Logistic regression for adjective order. The function $S(A, N)$ is the predictor to be evaluated, β_0 and β_1 are the free parameters to be fit, and ϵ is an error term to be minimized.

PMI and IC, as measures of the strength of statistical association between nouns and adjectives, could conceivably double as predictors of parsing accuracy for automatic dependency parsers. If that is the case, then we might observe that AAN triples with low PMI or high IC are rare in automatically parsed data. However, this would not be a consequence of any interesting theory of cognitive cost, but rather simply an artifact of the automatic parser used. To avoid this confound, we include an evaluation based on held-out hand-parsed data in the form of the English Web Treebank.

5 Results

Table 1a shows the accuracies of our predictors in predicting held-out adjective orders in the Common Crawl test set, visualized in Figure 2a. We find that the pattern of results depends on whether predictors are estimated based on wordforms or based on distributional clusters. When estimating based on wordforms, we find that subjectivity and PMI have the best accuracy. When estimating based on clusters, the accuracy of PMI drops, and the best predictor is subjectivity, with IG close behind. We find a negative logistic regression weight for information gain, indicating that the adjective with lower information gain is placed first.

This basic pattern of results is confirmed in the hand-parsed EWT data. Accuracies of predictors on the EWT test set are shown in Table 1b and visualized in Figure 2b. When estimating based on wordforms, the best predictors are subjectivity and PMI, although the confidence intervals of all predictors are overlapping. When estimating based on clusters, IG has the best performance, and PMI again drops in accuracy. For this case, IG, IC, and subjectivity all have overlapping confidence intervals, so we conclude that there is no evidence that one is better than the other. However, we do have evidence that IG and IC are more accurate than PMI when estimated based on clusters.

5.1 Multivariate analysis

Adjective order may be determined by multiple separate factors operating in parallel. In order to investigate whether our predictors might be making independent contributions to explaining adjective order, we fit logistic regressions containing multiple predictors. If the best accuracy comes from a model with two or more predictors, then this would be evidence that these two predictors are picking up on separate sources of information relevant for predicting adjective order.

We conducted logistic regressions using all sets of two of our predictors. The top 5 such models, in terms of Common Crawl test set accuracy, are shown in Table 2. The best two are cluster/wordform subjectivity and wordform PMI, followed by cluster subjectivity and cluster information gain. No set of three predictors achieves significantly higher accuracy than the best predictors shown in Table 2.

5.2 Qualitative analysis

We manually examined cases where each model made correct and incorrect predictions in the hand-parsed EWT data. Table 3a shows example AAN triples that were ordered correctly by PMI, but not by subjectivity. These are typically cases where a certain adjective–noun pair forms a common collocation whose meaning is in some cases even noncompositional; for example, “bad behaviors” is a common collocation when describing training animals, and “ulterior motives” and “logical fallacy” are likewise common English collocations. In contrast, when subjectivity makes the right prediction and PMI makes the wrong prediction, these are often cases where a word pair which normally would form a collocation is broken up by another adjective, such as “dear sick friend”, where “dear friend” is a common collocation.

We also performed a manual qualitative analysis to determine the contribution of information gain beyond subjectivity and PMI. Table 3b shows examples of such cases from the EWT. Many of these seem to be cases with weak preferences, where both the attested order and the the flipped order are acceptable (e.g., “tiny little kitten”).

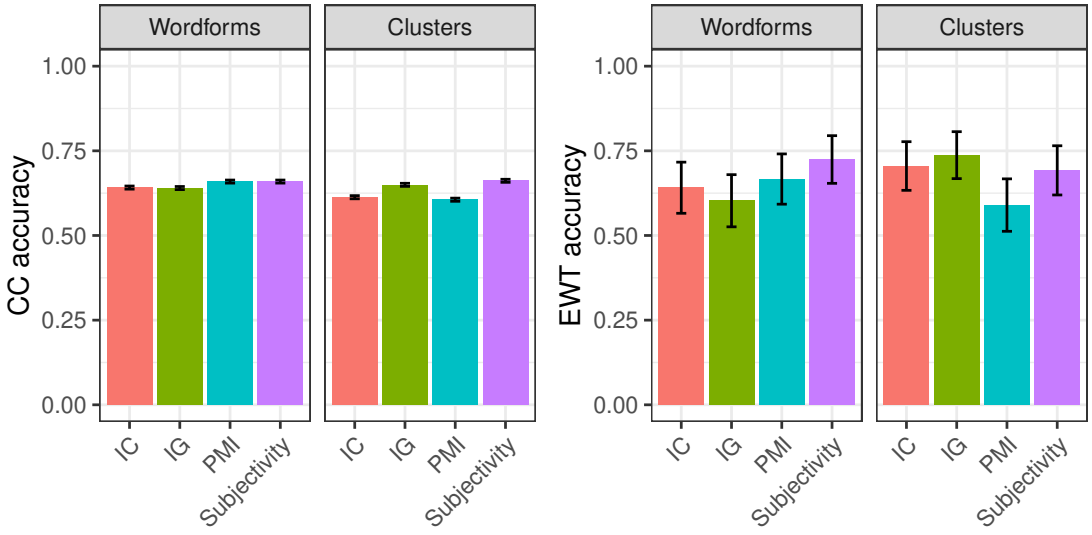
Predictor	Accuracy	Conf. Interval
Subj. (cluster)	**.661**	[.657, .666]
PMI (wordform)	**.659**	[.654, .664]
Subj. (wordform)	**.659**	[.654, .664]
IG (cluster)	.650	[.645, .654]
IC (wordform)	.642	[.634, .646]
IG (wordform)	.640	[.635, .645]
IC (cluster)	.613	[.608, .618]
PMI (cluster)	.606	[.601, .610]

(a) Common Crawl ($N = 41822$).

Predictor	Accuracy	Conf. Interval
IG (cluster)	.737	[.668, .806]
Subj. (wordform)	.724	[.654, .795]
IC (cluster)	.705	[.633, .777]
Subj. (cluster)	.692	[.620, .765]
PMI (wordform)	.667	[.592, .741]
IC (wordform)	.641	[.566, .717]
IG (wordform)	.603	[.526, .680]
PMI (cluster)	.590	[.512, .667]

(b) Hand-parsed EWT ($N = 155$). All confidence intervals overlap, other than cluster-based PMI and IG.

Table 1: Accuracies of the predictors on AAN triples in the held-out test data.



(a) Common Crawl ($N = 41822$).

(b) Hand-parsed EWT ($N = 155$)

Figure 2: Accuracies of predictors on AAN triples in the held-out test data, with 95% confidence intervals shown.

Predictor	Accuracy	Conf. Interval
Subj. (cluster) + PMI (wordform)	**.723**	[.719, .727]
Subj. (wordform) + PMI (wordform)	.713	[.708, .717]
Subj. (cluster) + IG (cluster)	.699	[.695, .703]
Subj. (cluster) + IC (cluster)	.690	[.686, .695]
IG (cluster) + IC (cluster)	.684	[.680, .689]

Table 2: Common Crawl test set accuracy of the top 5 models combining two predictors.

A_1	A_2	N
major	bad	behaviors
large	outstanding	debts
classical	logical	fallacy
dark	ulterior	motives
minor	fine	tuning

(a) Ordered correctly by wordform PMI, but not by wordform subjectivity.

A_1	A_2	N
tiny	little	kitten
correct	legal	name
chronic	intractable	pain
radical	religious	politics
lonely	eerie	place

(b) Ordered correctly by cluster-based information gain, but not by cluster-based subjectivity nor PMI.

Table 3: Selected examples of AAN triples ordered incorrectly by our models, from the EWT test set.

5.3 Interpretation

Our results broadly support the following interpretation. Adjective ordering preferences are largely determined by a semantic factor that can be quantified variously using wordform subjectivity or distributional-cluster-based estimates of information gain. In addition to this factor, another factor is in play: when an adjective–noun pair forms a collocation with a possibly non-compositional meaning, then the adjective in this pair will tend to be placed next to the noun. This latter factor is measured by PMI. This interpretation matches that of Hahn et al. (2018), who found separate contributions from PMI and a model-based operationalization of subjectivity.

Our interpretation is supported by the following points from the analysis above. First, among predictors based solely on wordforms, the best accuracy is obtained by a combination of subjectivity and PMI. Second, when we turn to estimates based on clusters, two things happen: the accuracy of PMI drops, and the accuracy of information gain increases while the accuracy of subjectivity stays about the same. This pattern of results suggests that PMI is measuring a factor that has more to do with specific wordforms, while IG and subjectivity are measuring a factor that has more to do with semantic uncertainty about the noun or about the relationship between the adjective and the noun.

6 Conclusion

We examined a number of theoretically-motivated predictors of adjective order in dependency treebank corpora of English. We found that the predictors have comparable accuracy, but that it is possible to identify two broad factors: a semantic factor variously captured by subjectivity scores and information gain based on word clusters, and a wordform-based factor captured by PMI.

This study provides a framework for evaluating further theories of adjective order, and for evaluating the theories given here against new data from dependency treebanks. Generalizing to larger datasets of English is straightforward. More excitingly, we now have the opportunity to bring new languages into the fold. The vast majority of research on adjective ordering, and all the corpus work to our knowledge, has been done on English, where adjectives almost always come before the noun. Studying other typologically-distinct languages provides an opportunity to disentangle the theories that we studied here in a way that cannot be done in English.

The available behavioral evidence suggests that mirror-image preferences (e.g., "box blue big") may be the norm in post-nominal adjective languages (Martin, 1969; Scontras et al., 2020). Information locality, subjectivity, and integration cost make precisely that prediction, though none addresses mixed-type languages in which adjectives can precede or follow nouns. It is an open question how to implement IG for these post- or mixed-placement adjectives; one possibility is to measure the information gained when the set of adjectives associated to a noun A_n is partitioned by an adjective a. In that case, the predictions about post-nominal order could differ substantially from the predictions about pre-nominal order.

Our dependency-treebank-based methods can be applied to any other corpus of any language, provided it has enough data in the form of adjective–noun pairs to get reliable estimates of the information-theoretic predictors. Such studies will be crucial to achieve a complete computational understanding of natural language syntax.

References

Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins. 2016. Globally normalized transition-based neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2442–2452, Berlin, Germany. Association for Computational Linguistics.

R. Harald Baayen, Richard Piepenbrock, and Leon Gulikers. 1995. *The CELEX Lexical Database. Release 2 (CD-ROM)*. Linguistic Data Consortium, University of Pennsylvania.

Galia Bar-Sever, Rachael Lee, Gregory Scontras, and Lisa S. Pearl. 2018. Little lexical learners: Quantitatively assessing the development of adjective ordering preferences. In *42nd annual Boston University Conference on Language Development*, pages 58–71.

Christian Bentz, Dimitrios Alikaniotis, Michael Cysouw, and Ramon Ferrer-i-Cancho. 2017. The entropy of words—Learnability and expressivity across more than 1000 languages. *Entropy*, 19:275–307.

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.

Christian Buck, Kenneth Heafield, and Bas Van Ooyen. 2014. N-gram counts and language models from the common crawl. In *LREC*, volume 2, page 4. Citeseer.

Xinying Chen and Ramon Ferrer-i-Cancho, editors. 2019. *Proceedings of the First Workshop on Quantitative Syntax (Quasy, SyntaxFest 2019)*. Association for Computational Linguistics, Paris, France.

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

Guglielmo Cinque. 1994. On the evidence for partial N-movement in the Romance DP. In R S Kayne, G Cinque, J Koster, J.-Y. Pollock, Luigi Rizzi, and R Zanuttini, editors, *Paths Towards Universal Grammar. Studies in Honor of Richard S. Kayne*, pages 85–110. Georgetown University Press, Washington DC.

Thomas M. Cover and J. A. Thomas. 2006. *Elements of Information Theory*. John Wiley & Sons, Hoboken, NJ.

J. H. Danks and S. Glucksberg. 1971. Psychological scaling of adjective orders. *Journal of Verbal Learning and Verbal Behavior*, 10(1):63–67.

Robert M. W. Dixon. 1982. *Where have all the adjectives gone? And other essays in semantics and syntax*. Mouton, Berlin, Germany.

Melody Dye, Petar Milin, Richard Futrell, and Michael Ramscar. 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.

William Dyer. 2018. Integration complexity and the order of cosisters. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 55–65, Brussels, Belgium. Association for Computational Linguistics.

William E. Dyer. 2017. *Minimizing integration cost: A general theory of constituent order*. Ph.D. thesis, University of California, Davis, Davis, CA.

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

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 International Conference on Quantitative Syntax*, pages 2–15, Paris.

Richard Futrell and Roger Levy. 2017. Noisy-context surprisal as a human sentence processing cost model. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 688–698, Valencia, Spain.

Richard Futrell, Roger Levy, and Edward Gibson. 2017. Generalizing dependency distance: Comment on "dependency distance: A new perspective on syntactic patterns in natural languages" by haitao liu et al. *Physics of Life Reviews*, 21:197–199.

Richard Futrell, Peng Qian, Edward Gibson, Evelina Fedorenko, and Idan Blank. 2019. Syntactic dependencies correspond to word pairs with high mutual information. In *Proceedings of the Fifth International Conference on Dependency Linguistics (DepLing 2019)*, Paris.

Michael Hahn, Judith Degen, Noah Goodman, Daniel 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 (CogSci)*.

John Hale. 2006. Uncertainty about the rest of the sentence. *Cognitive science*, 30(4):643–672.

John T. Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics and Language Technologies*, pages 1–8.

John T. Hale. 2016. Information-theoretical complexity metrics. *Language and Linguistics Compass*, 10(9):397–412.

R. Hetzron. 1978. On the relative order of adjectives. In H. Seller, editor, *Language Universals*. Narr, Tübingen, Germany.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

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.

Robert Malouf. 2000. The order of prenominal adjectives in natural language generation. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, ACL '00, pages 85–92, Stroudsburg, PA, USA. Association for Computational Linguistics.

Christopher D. Manning. 2003. Probabilistic syntax. In *Probabilistic Linguistics*, pages 289–341. MIT Press.

J E Martin. 1969. Some competence-process relationships in noun phrases with prenominal and postnominal adjectives. *Journal of Verbal Learning and Verbal Behavior*, 8:471–480.

Emily Morgan and Roger Levy. 2016. Abstract knowledge versus direct experience in processing of binomial expressions. *Cognition*, 157:382–402.

Joakim Nivre. 2015. Towards a universal grammar for natural language processing. In *Computational Linguistics and Intelligent Text Processing*, pages 3–16. Springer.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

J. R. Quinlan. 1986. Induction of decision trees. *Machine Learning*, 1(1):81–106.

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

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

Gregory Scontras, Judith Degen, and Noah D. Goodman. 2019. On the grammatical source of adjective ordering preferences. *Semantics and Pragmatics*.

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

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Christopher D. Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.

Mihael Simonič. 2018. Functional explanation of adjective ordering preferences using probabilistic programming. Master's thesis, University of Tübingen.

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

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

David Temperley and Daniel Gildea. 2018. Minimizing syntactic dependency lengths: Typological/cognitive universal? *Annual Review of Linguistics*, 4:1–15.

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

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