Predicting cross-linguistic adjective order with information gain

William Dyer
Oracle Corporation
william.dyer@oracle.com

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

Zoey Liu
Boston College
ying.liu.5@bc.edu

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

Abstract

Languages vary in their placement of multiple adjectives before, after, or surrounding the noun, but they typically exhibit strong intra-language tendencies on the relative order of those adjectives (e.g., the preference for 'big blue box' in English, 'grande boîte bleue' in French, and 'alṣundūq al'azraq alkabīr' in Arabic). We advance a new quantitative account of adjective order across typologically-distinct languages based on maximizing information gain. Our model addresses the left-right asymmetry of French-type ANA sequences with the same approach as AAN and NAA orderings, without appeal to other mechanisms. We find that, across 32 languages, the preferred order of adjectives mirrors an efficient algorithm of maximizing information gain.

1 Introduction

Languages that allow multiple sequential adjective modifiers tend to exhibit strong tendencies on the relative order of adjectives, as in 'big blue box' vs. 'blue big box' in English (Dixon, 1982). To date, most of the research on adjective ordering has focused on preferences in pre-nominal languages like English where adjectives precede the modified noun (Futrell et al., 2020a), or in post-nominal languages like Arabic where adjectives follow the noun (Kachakeche and Scontras, 2020). This research usually posits a metric, such as information locality (Futrell et al., 2020b) or subjectivity (Scontras et al., 2017), which governs the preferred distance between a noun and its adjectives. Because these theories predict only the relative linear *distance* between noun and adjective, they cannot be straightforwardly applied to mixed languages like French, where adjectives regularly appear both before and after the modified noun, at least not without added assumptions about hierarchical distance (Cinque, 1994). Instead, these mixed languages are often modeled with constraints on which adjective classes or functions can appear before or after a noun (Cinque, 2010; Fox and Thuilier, 2012).

Traditional accounts of adjective ordering in the linguistics literature often assume a **tree structure** in which the target measure is the hierarchical distance from noun (N) to adjective (A). According to syntactic accounts, ordering regularities are predicted by a universal hierarchy of lexical semantic classes (e.g., color adjectives are hierarchically closer to the modified noun than size adjectives; Cinque, 1994; Scott, 2002). Alternative accounts use aspects of adjective meaning to predict adjective order, making appeal to notions like 'inherentness' (Whorf, 1945) or 'definiteness of denotation' (Martin, 1969). Recently, Scontras et al. (2017) provide experimental evidence that their synthesis of semantic predictors into a continuum based on subjectivity reliably predicts ordering preference in English; followup studies have found subjectivity to be a reliable predictor in other languages as well (Tagalog: Samonte and Scontras, 2019; Mandarin: Shi and Scontras, 2020; Arabic: Kachakeche and Scontras, 2020; Spanish: Rosales Jr. and Scontras, 2019; Scontras et al., 2020). Explanations for the role of subjectivity in adjective ordering show how subjectivity-based orderings are more **efficient** than alternative orderings, thereby maximizing communicative success (Simonič, 2018; Hahn et al., 2018; Franke et al., 2019; Scontras et al., 2019).

Other efficiency-based approaches to adjective order quantify efficiency with information-theoretic measures of word distributions such as surprisal or entropy (Cover and Thomas, 2006; Levy, 2008). Models in this vein have a long conceptual history in the field, originating with the idea that semantic closeness between words is reflected in syntactic closeness in a surface realization (Sweet, 1900; Jespersen, 1922; Behaghel,

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 957–967

1932). Modern quantitative incarnations include integration cost (Dyer, 2017) and information locality (Futrell et al., 2020b), both generalizations of the widely-accepted principle of dependency distance minimization (Liu et al., 2017; Temperley and Gildea, 2018).

Crucially, while previous approaches are able to model symmetrical structures within the noun phrase, as in the mirror-image A_1A_2N orders of English and the NA_2A_1 orders of Arabic, a hierarchical approach cannot model the **left–right asymmetry** of Romance A_1NA_2 without an appeal to other, usually syntactic, mechanisms (Cinque, 2009, 2010).

We advance an information-theoretic factor that predicts adjective ordering across the three typological 'templates' of adjective order—pre (AAN), mixed (ANA), and post (NAA)—based on **information gain** (IG), a measure of the reduction in uncertainty attained by transforming a dataset. IG is used in machine learning for ordering the nodes of a decision tree (Quinlan, 1986; Norouzi et al., 2015), where nodes are most often ordered in a greedy fashion such that the information gain of each node is maximized. By analogy, we view the noun phrase as a decision tree for reducing a listener's uncertainty about a speaker's intended meaning. Each word acts as a node in the decision tree; preferred adjective orders thus reflect an efficient ordering of nodes.

2 Empirical background

Empirical investigations of adjective ordering have focused on the cross-linguistic stability of these preferences across a host of unrelated languages (e.g., Dixon, 1982; Hetzron, 1978; Sproat and Shih, 1991). For example, where English speakers prefer 'big blue box' to 'blue big box', Mandarin speakers similarly prefer *dà-de lán-de xiāng-zi* 'big blue box' to *lán-de dà-de xiāng-zi* 'blue big box' (Shi and Scontras, 2020). In post-nominal languages, we find the mirror-image of the English pattern, such that adjectives that are preferred closer to the noun in pre-nominal languages are also preferred closer to the noun in post-nominal languages.[1] For example, speakers of Arabic prefer *alṣundūq al'azraq alkabīr* 'the box blue big' to *alṣundūq alkabīr al'azraq* 'the box big blue'.

[1]Celtic languages have been claimed to be an exception to this trend (Sproat and Shih, 1991), though our own investigations into Irish suggest that it behaves like other post-nominal languages, at least with respect to information gain.

In support of the cross-linguistic stability of adjective ordering preferences, Leung et al. (2020) present a latent-variable model capable of accurately predicting adjective order in 24 languages from seven different language families, achieving a mean accuracy of 78.9% on an average of 1335 sequences per language. Importantly, the model succeeds even when the training and testing languages are different, thus demonstrating that different languages rely on similar preferences. However, Leung et al.'s study was limited to AAN and NAA templates. There has been very little corpus-based empirical work on ordering preferences in the mixed ANA template, where adjectives both precede and follow the modified noun.[2]

While Leung et al. (2020) learn adjective order by training on observed adjective pairs, an alternate strategy is to posit one or more a priori metrics as an underlying motivation for adjective order (e.g., Malouf, 2000, in part). This approach allows for the study of why adjective orders might have come about. To that end, Futrell et al. (2020a) report an accuracy of 72.3% for English triples based on a combination of subjectivity and information-theoretic measures derived from the distribution of adjectives and nouns.

One of the information-theoretic measures analyzed by Futrell et al. (2020a) is an implementation of information gain based on the partitioning an adjective performs on the space of possible noun referents. However, it is unclear how this formulation of information gain could be implemented for post-nominal adjectives, in which the noun has presumably already been identified. Instead, the current study implements information gain based on feature vectors, as outlined in §3.

To our knowledge, the current study is the first attempt at predicting adjective order across all three templates, with an eye not only to raw accuracy, but in hopes of illuminating the functional pressures which might contribute to word ordering preferences in general. While we acknowledge that multiple factors are likely involved in adjective order preferences, our contribution here is a single quantitative factor capable of predicting adjective order across typologically distinct languages.

[2]We note three empirical studies that have examined the placement of a single adjective or adjective phrase before or after the noun in Romance languages: Thuilier (2014), Gulordava et al. (2015) and Gulordava and Merlo (2015). However, these studies do not tackle the question of order preferences among ANA triples.

3 Information gain

3.1 Picture of communication

We assume that a speaker is trying to communicate a **meaning** to a listener, with a meaning represented as a binary vector, where each dimension of the vector corresponds to a feature. Multiple features can be true simultaneously. For example, a speaker might have in mind a vector like $m_1 = [111\ldots0]$ in Figure 1, where the vector has value 1 in the dimensions for 'is-big' (f_0), 'is-grey' (f_1), and 'is-elephant' (f_2), and 0 for all other features. A meaning of this sort would be conveyed by the noun phrase 'big grey elephant'. We call m a **feature vector** and the set of feature vectors M.

The listener does not know which meaning m the speaker has in mind; the listener's state of uncertainty can be represented as a probability distribution over all possible feature vectors, $P(m)$, corresponding to the prior probability of encountering a given feature vector. We call this distribution the **listener distribution** L.

By conveying information, each word in a sequence causes a change in the listener's prior distribution. Suppose as in Figure 1 that a listener starts with probability distribution L, then hears a word w conveying a feature (f_2), resulting in the new distribution L'. The amount of change from L to L' is properly measured using the Kullback–Leibler (KL) divergence $D_{\text{KL}}[L'||L]$ (Cover and Thomas, 2006). Therefore, the divergence $D_{\text{KL}}[L'||L]$ measures the amount of information about meaning conveyed by the word.

Another measure of the change induced by a word is the information gain, an extension of KL divergence to include the notion of negative evidence. Let $\bar{L}'$ represent the listener's probability distribution over feature vectors conditional on the negation of w. By taking a weighted sum of the positive and negative KL divergence, we recover **information gain** (Quinlan, 1986):

$$\text{IG} = \frac{|L'|}{|L|} D_{\text{KL}}[L'||L] + \frac{|\bar{L}'|}{|L|} D_{\text{KL}}[\bar{L}'||L], \quad (1)$$

where $|L|$ indicates the number of elements in the support of L with non-zero probability. Information gain represents the information conveyed by a word and also the information conveyed by its negation.

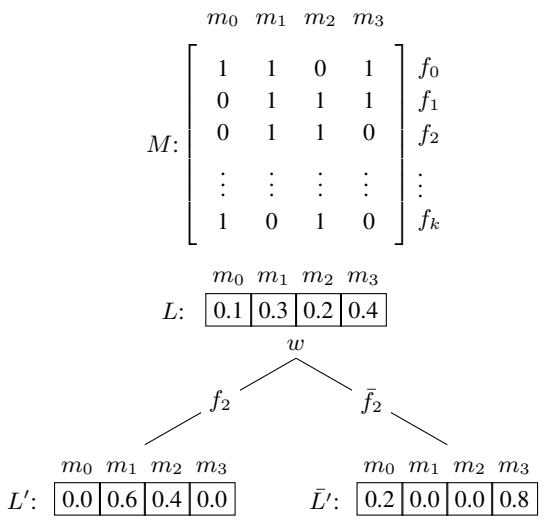

Figure 1: A toy universe composed of four feature vectors m defined by k binary features f and an associated probability distribution L. Partitioning L on f_2 yields L', the probability distribution of the feature vectors containing a 1 for f_2, viz. m_1 and m_2, as well as $\bar{L}'$, the distribution of feature vectors containing a 0 for f_2, or $\bar{f}_2$.

3.2 Relationship to other quantities

Our IG quantity in Eq. 1 is drawn from the ID3 algorithm for generating decision trees (Quinlan, 1986). The goal of ID3 is to produce a classifier for some random variable (call it L) which works by successively evaluating some set of binary features in some order. The optimal order of these features is given by greedily maximizing information gain, where information gain for a feature f is a measure of how much the entropy of L is decreased by partitioning the dataset into positive and negative subsets based on whether f is present or absent. Our application of information gain to word order comes from treating each word as a binary indicator for the presence or absence of the associated feature, and then applying the ID3 algorithm to determine the optimal order of these features.

The first term of Eq. 1, the divergence $D_{\text{KL}}[L'||L]$, measures the amount of information about L conveyed by the word w and has been the subject of a great deal of study in psycholinguistics. In particular, Levy (2008) shows that if the word w and the context c can be reconstructed perfectly from the updated belief state L', then the amount of information conveyed by w reduces to the **surprisal** of word w in context c:

$$D_{\text{KL}}[L'||L] = -\log p(w|c). \quad (2)$$

Importantly for our purposes, the positive evidence term $D_{\mathrm{KL}}[L'||L]$ in Eq. 1 alone is unlikely to make useful predictions about cross-linguistic ordering preferences, because surprisal is invariant to reversal of word order across a language as a whole (Levy, 2005; Futrell, 2019): the same surprisal values would be measured for any given language and a language with all the same sentences in reverse order. As such, these metrics are unable to predict any a priori asymmetries in word-order preferences between pre- and post-nominal positions.

3.3 Negative evidence

The new feature of information gain, which has not been presented in previous information-theoretic models of language, is the negative evidence term in $D_{\mathrm{KL}}[\bar{L}'||L]$, indicating the change in the listener's belief about L given the negation of the features indicated by word w, a quantity related to extropy (Lad et al., 2015). For example, consider *académie*/NOUN *militaire*/ADJ 'military/ADJ academy/NOUN' in French. Let L represent a listener's belief state after having heard the noun *académie* 'academy'. Upon hearing the adjective *militaire* 'military', L is partitioned into L'—the portion of L in which *militaire* is a feature—and $\bar{L}'$, the portion of L in which *militaire* is not a feature. Put another way, $\bar{L}'$ is the probability distribution over non-military academies.

The negative evidence portion of information gain is of primary interest to us because it breaks the symmetry to word-order reversal that we would have if we used the positive evidence term alone. That is, because the sum of surprisals of words w_1 and w_2 in the context of w_1 is the log joint probability of a sequence:

$$-\log p(w_1) - \log p(w_2|w_1) = -\log p(w_1, w_2), \tag{3}$$

the sum of w_2 and w_1 in the context of w_2 necessarily yields the same quantity. Conversely, IG's negative-evidence value is related to the log probability of w_2 conditional on the event of *not* observing w_1, and as such the sum of negative evidence values is not equivalent to the joint surprisal.

Information gain can therefore predict left–right asymmetrical word-order preferences such as the order of adjectives in ANA templates. Further, it maps onto a well-known decision rule for the ordering of trees.

3.4 An efficient algorithm

The goal of algorithms such as ID3 is to produce a decision tree which divides a dataset into equal-sized and mutually-exclusive partitions, thereby creating a shallow tree (Quinlan, 1986). While finding the smallest possible binary decision tree is NP-complete (Hyafil and Rivest, 1976), ID3's locally-optimal approach has proven quite effective at producing shallow trees capable of accurate classification (Dobkin et al., 1996).

By analogy, the ordering of adjectives in a noun phrase by maximizing information gain likewise produces a tree with balanced positive and negative partitions at each node. Specifically, adjectives that minimize the entropy of both the positive and negative evidence are placed before adjectives which are less 'decisive' at partitioning feature vectors.

4 Methodology

4.1 Data

Our study relies on two types of source data, both extracted from the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies (Ginter et al., 2017; Zeman et al., 2017): a set of Common Crawl and Wikipedia text data from a variety of languages, automatically parsed according to the Universal Dependencies scheme with UDPipe (Straka and Straková, 2017). First, we extract **noun phrases** (NPs) containing at least one adjective to populate feature vectors (§4.3). Second, we extract **triples**, instances of a noun and two dependent adjectives, where the three words are sequential in the surface order and neither the noun nor the adjectives have other dependents.

We restrict triples in this way to minimize the effect that other dependents might have on order preferences. For example, while single-word adjectives tend to precede the noun in English, as in 'the *nice* people', adjectives in larger right-branching phrases often follow: 'the people *nice to us*' (Matthews, 2014), a trend also seen in Romance (Gulordava et al., 2015; Gulordava and Merlo, 2015). Similarly, conjunctions have been shown to weaken or neutralize preferences (Fox and Thuilier, 2012; Rosales Jr. and Scontras, 2019; Scontras et al., 2020).

NPs and triples extracted from the Wikipedia dumps are used to generate feature vectors and to train our regression (§4.4). We use triples from the Common Crawl dumps to perform hold-out accuracy testing.

4.2 Normalization

Because our source data are extracted from dumps of automatically-parsed text, they contain a large amount of noise, such as incorrectly assigned syntactic categories, HTML, nonstandard orthography, and so on. To combat this noise, we extract all lemmas with UPOS marked as ADJ and NOUN in all Universal Dependencies (UD) v2.7 corpora (Zeman et al., 2020) for a given language—the idea being that the UD corpora are of higher quality—and include only NPs and triples in which the adjectives and nouns are in the UD lists. All characters are case-normalized, where applicable.

4.3 Feature vectors

Each NP attested in the Wikipedia corpus for a given language corresponds to a feature vector with value 1 in the dimension associated with each adjective or noun lemma. For example, an NP such as "the best room available" generates a vector containing 1 for 'is-available', 'is-best', and 'is-room'.

The relative count of each NP in the Wikipedia corpus yields a probability distribution on feature vectors. It is this distribution which is transformed by partitioning on each lemma in a triple.

4.4 Evaluation

For a given typological template (AAN, ANA, or NAA) there are two competing variants; our tasks are to (i) predict which of the variants will be attested in a corpus and (ii) show a cross-linguistic consistency in how that prediction comes about.

Because we are limiting our study to the two competing variants within each template, the position of the noun is invariant, leaving only the relative order of the two adjectives to determine the order of a triple. Our problem thus reduces to whether the information gain of the first linear adjective is greater than that of the second.

In the case of AAN and ANA triples, the IG of each adjective is calculated by partitioning the entire set of feature vectors L on each of the two adjectives. In the case of NAA triples, however, IG is calculated by partitioning only those feature vectors which 'survive' the initial partition by the noun, and are therefore part of L'. Thus we calculate $\text{IG}(L, a)$ before the noun and $\text{IG}(L', a)$ after.

Rather than simply implement the ID3 algorithm and choose adjectives based on their raw information gain, we train a logistic regression to predict surface orders based on the *difference* of IG between the attested first and second adjective, a method previously used by Morgan and Levy (2016) and Futrell et al. (2020a). The benefits of this approach are two-fold: we are able to account for bias in the distribution of adjectival IGs, and we can more easily deconstruct how strong information gain is as a predictor of adjective order.

Within each template, for each attested triple τ, let π_1 be the lexicographically-sorted first permutation of τ and π_2 be the second, with α_1 being the first linear adjective in π_1 and α_2 being the first linear adjective in π_2. Our independent variable p is whether π_1 is attested in the corpus, and our dependent variable is the difference between the information gain of α_1 and α_2. We train the coefficients β_0 and β_1 in a logistic regression of the form

$$p = \begin{cases} 1, & \text{if } \pi_1 \text{ is attested} \\ 0, & \text{if } \pi_2 \text{ is attested} \end{cases}$$
$$\log \frac{p}{1-p} \sim \beta_0 + \beta_1[\text{IG}(\alpha_1) - \text{IG}(\alpha_2)]. \quad (4)$$

A positive value for β_1 tells us that permutations in which the larger-IG adjective is placed first tend to be attested. The value of β_0 tells us whether there is a generalized bias towards a positive or negative $\text{IG}(\pi_1) - \text{IG}(\pi_2)$. The accuracy we achieve by running the logistic regression on held-out testing data tells us the effectiveness of an IG-based algorithm at predicting adjective order.

4.5 Reporting results

We report results for languages from which at least 5k triples could be analyzed, and for templates representing at least 10% of a language's triples in UD corpora. The count of analyzable triples for each language is a product of those available in the 2017 CoNLL Shared Task, those with sufficiently large UD v2.7 corpora, and those that meet our extraction requirements (§4.1).

Because we are interested in exploring a cross-linguistic predictor of adjective order, we report macro-average accuracies and β_1 coefficients. That is, each language's accuracy and coefficient are calculated independently and are then averaged. We report both type- and token-accuracy, using the latter in our analysis based on the intuition that the preference for the order of a commonly-occurring triple is stronger than a more rare one.

AAN	language	n	β_1	P	token acc.	type acc.
mean β_1	Bulgarian	13018	20.058	0.000	0.650	0.649
18.591 [15.740, 21.443]	Chinese	5909	18.604	0.000	0.724	0.766
	Croatian	15555	21.246	0.000	0.666	0.634
mean token accuracy	Czech	27899	28.207	0.000	0.671	0.665
0.656 [0.630, 0.683]	Danish	11226	17.506	0.000	0.786	0.770
	Dutch	11279	12.201	0.000	0.609	0.605
mean type accuracy	English	23311	22.076	0.000	0.643	0.647
0.645 [0.616, 0.674]	Finnish	12605	15.342	0.000	0.655	0.644
	German	16391	16.210	0.000	0.601	0.606
	Greek	5506	18.383	0.000	0.631	0.643
	Latvian	5290	15.826	0.000	0.594	0.551
	Russian	25397	25.697	0.000	0.658	0.651
	Slovak	11933	25.935	0.000	0.700	0.651
	Slovenian	18859	28.192	0.000	0.670	0.661
	Swedish	10937	11.462	0.000	0.717	0.711
	Turkish	12115	12.579	0.000	0.576	0.577
	Ukrainian	11474	15.949	0.000	0.593	0.592
	Urdu	6432	9.170	0.000	0.673	0.593
ANA	language	n	β_1	P	token acc.	type acc.
mean β_1	Basque	3322	-9.623	0.000	0.703	0.678
31.313 [16.786, 45.841]	Catalan	3117	45.135	0.000	0.818	0.814
	Croatian	4912	-3.411	0.106	0.608	0.604
mean token accuracy	French	5673	43.349	0.000	0.771	0.756
0.737 [0.674, 0.799]	Galician	5020	68.290	0.000	0.805	0.806
	Indonesian	1521	-2.462	0.138	0.543	0.524
mean type accuracy	Italian	9484	36.658	0.000	0.681	0.698
0.726 [0.665, 0.787]	Persian	2598	43.242	0.000	0.794	0.766
	Polish	13481	24.873	0.000	0.684	0.655
	Portuguese	7580	32.374	0.000	0.734	0.725
	Romanian	2426	46.823	0.000	0.730	0.739
	Spanish	9212	57.813	0.000	0.744	0.738
	Vietnamese	2636	24.013	0.000	0.962	0.931
NAA	language	n	β_1	P	token acc.	type acc.
mean β_1	Arabic	11595	4.595	0.000	0.693	0.660
4.140 [3.128, 5.152]	Basque	4899	1.957	0.000	0.626	0.635
	Catalan	2878	5.024	0.000	0.710	0.722
mean token accuracy	French	8368	5.143	0.000	0.737	0.749
0.680 [0.639, 0.721]	Galician	1334	5.776	0.000	0.716	0.694
	Hebrew	6751	1.115	0.000	0.558	0.560
mean type accuracy	Indonesian	5724	4.631	0.000	0.740	0.734
0.687 [0.647, 0.726]	Italian	4523	4.057	0.000	0.713	0.739
	Persian	12683	1.583	0.000	0.605	0.606
	Portuguese	5139	5.329	0.000	0.726	0.730
	Romanian	8492	5.333	0.000	0.742	0.746
	Spanish	6245	6.214	0.000	0.713	0.745
	Vietnamese	3354	3.068	0.000	0.561	0.606
	comprehensive mean		18.08		0.687	0.681

Table 1: Results by template and language: n triples analyzed, regression coefficient β_1 and P-value, and test accuracies. Means with 95% confidence intervals shown for each template.

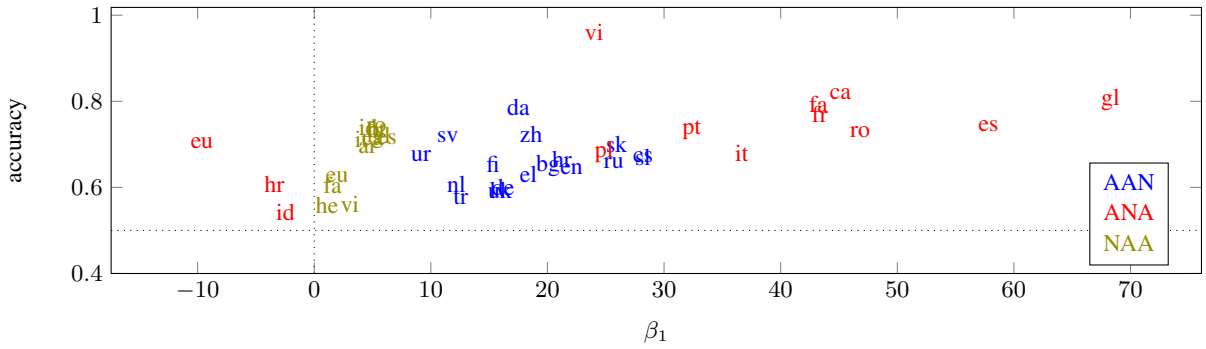

Figure 2: Plot of accuracy and β_1 coefficient, categorized by template type.

5 Results

We extracted and analyzed at least 5k triples from 32 languages across a variety of families.[3] Because some languages contain triples in two typological templates, we report results for 44 sets of triples. Table 1 reports language-specific results and means for each template, including n triples analyzed, regression coefficient β_1 and P-value, token and type accuracy, and 95% confidence intervals. Figure 2 shows a plot of accuracy and β_1 coefficient for each language, categorized by template.

As reported in Table 1, we find above-chance ($> 50\%$) accuracy for all languages tested. We accurately predict 65.6% of AAN triples, 73.7% of ANA triples, and 68.0% of NAA triples, for a comprehensive accuracy across all languages of 68.7%. Overlapping 95% confidence intervals across template suggest that IG-based prediction performs equally well across templates.

The high performance on Vietnamese ANA triples (96.2%) is largely due to the algorithm correctly predicting that the highly-frequent adjective *nhiều* 'many' should be placed before the noun, while most other adjectives are placed after.[4]

Though we cannot make a direct comparison to other studies due to a lack of shared data, Table 2 shows that our cross-linguistic accuracy of 68.7% bests any single predictor applied to a similar set of English AAN triples by Futrell et al. (2020a).

The learned β_1 coefficient is not significantly different between AAN (18.591) and ANA (31.313) triples, though that of NAA (4.140) triples is significantly smaller than the other two. More generally, of the 44 datasets tested, β_1 is positive in 41 (93.2%), suggesting that there is a strong preference to maximize information gain. Further, of the three instances of a negative β_1, two (Croatian and Indonesian ANA) do not reach significance, perhaps due to a paucity of data. The sole significant negative β_1 is from Basque ANA triples.

	n	accuracy	confidence interval
IG-FV	44	**0.687**	[0.686, 0.688]
Subj.	1	0.661	[0.657, 0.666]
PMI	1	0.659	[0.654, 0.664]
IG-NR	1	0.650	[0.645, 0.654]
IC	1	0.642	[0.634, 0.646]

Table 2: Comparison across n languages of the current metric, IG of feature vectors (IG-FV), and subjectivity, PMI, IG of noun referents (IG-NR), and integration cost (IC) (Futrell et al., 2020a)

6 Discussion

6.1 β_1 coefficient

Our results show a strong tendency across typological templates and across languages for the adjective which yields a larger information gain to be placed before the other, as evidenced by a positive β_1. However, the absolute value of β_1 is difficult to interpret without understanding the relative magnitudes of the underlying IG scores, magnitudes that vary across datasets and word distributions.

In general, we observe that a larger value of β_1 indicates that IG is a more reliable predictor within a dataset. More specifically, a value of $\beta_1 = 1$ indicates that if the IG difference between orders is equal to one bit, then the log odds of the order with larger IG increases by one.

[3] https://github.com/wmdyer/infogain

[4] One might worry about the classification of 'many' as an adjective. While widely extant across languages, the class of adjectives is not entirely homogeneous. As such, the equivalent of a word like 'many' in some languages might be marked as an adjective, determiner, or other syntactic category. For the current study, we simply follow the UD annotation scheme.

	n	rate	confidence interval
AAN	18	0.017	[0.012, 0.022]
ANA	13	**0.007**	[0.002, 0.011]
NAA	13	0.022	[0.013, 0.032]
all	44	0.016	[0.012, 0.020]

Table 3: Macro-average rate of adjectives attested in both possible orders within each template, showing n languages, rate of attestation, and 95% confidence intervals.

6.2 Asymmetries

The preference for one variant of an ANA triple over the other is an asymmetry without a straightforward explanation in a distance-based model; there is no clear mapping from ANA onto the other templates, which means that an adjective's relative distance to the noun is not informative. Our algorithm is novel in that the placement of the adjectives is governed by greedy IG, not distance to the noun—an innovation that allows us to break the symmetry between the adjectives in ANA triples. Similarly, IG makes no a priori prediction as to whether a mirror- or same-order will emerge between AAN and NAA triples: both pre- and post-nominal behavior is a product of ordering adjectives such that information gain is maximized, and IG itself is fundamentally derived from the distribution of adjectives and nouns that populate a language's possible feature vectors for conveying meaning.

Another left–right asymmetry that has been posited in the linguistics literature holds that dependents placed before the head in a surface realization (e.g., the adjectives in an AAN triple) follow a more rigid ordering than those placed after (e.g., the adjectives in a NAA triple; Hawkins, 1983). Both noun modifiers in general and adjectives specifically have been reported to follow this pattern, with a largely-universal pre-nominal ordering and a mirror, same, or 'free' post-nominal order (Hetzron, 1978). There is as yet no large-scale empirical evidence for this claim, though Trainin and Shetreet (2021) suggest that Hebrew NAA order preferences may be weaker than English AAN for a restricted set of adjective classes.

In an effort to empirically assess the claim that post-nominal orderings are more flexible compared to orderings pre-nominally across languages, Table 3 reports the average prevalence of adjective pairs attested in both possible orders (e.g., A_1A_2N and A_2A_1N, where N can be any noun) within each template in our dataset. At 95% confidence the difference between AAN and NAA does not reach significance, though the rate for ANA is significantly lower than the other two. More generally, the mean rate of just 1.6% across templates reinforces the notion that ordering preferences are quite robust regardless of template, at least for our normalized triples from the languages analyzed here.

6.3 Ablation

Equation 1 defines information gain as the conditioned sum of two elements, the positive evidence $D_{\mathrm{KL}}[L'||L]$ and the negative evidence $D_{\mathrm{KL}}[\bar{L}'||L]$. The positive evidence alone is akin to surprisal, a well-studied quantity in psycholinguistics (§3.2), while the negative evidence is related to extropy (§3.3). By ablating the IG formulation into the two terms discretely, we can show empirically that the proportionally-combined positive and negative evidence yield more accurate and consistent results than either of the two constituent terms alone.

Table 4 shows the mean accuracy and polarity proportion of the β_1 coefficient across languages and templates. The polarity of β_1 tells us whether maximizing IG (positive) or minimizing IG (negative) is the better strategy. Thus a polarity percentage close to 0 or 1 indicates more consistent behavior across templates.

For example, while the accuracy of using only positive evidence, $D_{\mathrm{KL}}[L'||L]$, for AAN triples is 0.565, that accuracy is realized due to a 0.000 rate of positive β_1 coefficient—that is, the 56.5% accuracy is achieved by minimizing IG, placing the adjective with the lower IG first. On the other hand, while using only positive evidence to predict NAA triples yields the same accuracy, 0.565, the coefficient polarity proportion of 0.769 means that, in most NAA cases, IG should be maximized. The three templates together reflect a modest accuracy (0.566) and an inconsistent coefficient polarity proportion (0.273).

Using only negative evidence, $D_{\mathrm{KL}}[\bar{L}'||L]$, yields even worse accuracies and similarly inconsistent coefficients as positive only. The accuracy across templates is little better than chance at 0.535, and the average coefficient polarity proportion of 0.273 likewise demonstrates that using negative evidence alone does not produce consistent behavior across templates.

	accuracy				proportion of positive β_1			
	AAN	ANA	NAA	all	AAN	ANA	NAA	all
$D_{\mathrm{KL}}[L'\|\|L]$	0.565	0.567	0.565	0.566	0.000	0.154	0.769	0.273
$D_{\mathrm{KL}}[\bar{L}'\|\|L]$	0.533	0.548	0.526	0.535	0.167	0.231	0.462	0.273
IG	**0.657**	**0.737**	**0.680**	**0.687**	**1.000**	**0.769**	**1.000**	**0.932**

Table 4: Ablation on accuracy and the proportion of positive coefficients for positive evidence ($D_{\mathrm{KL}}[L'||L]$) alone, negative evidence ($D_{\mathrm{KL}}[\bar{L}'||L]$) alone, and proportionally combined terms (IG). Boldfaced values indicate the highest accuracy or coefficient polarity proportion in each column.

The full IG calculation, including both positive and negative evidence, yields the highest accuracy across templates (0.687), as well as the highest for each template—AAN (0.657), ANA (0.737) and NAA (0.680). IG also demonstrates the most consistent behavior across languages and templates: at a rate of 0.932, maximizing IG yields the highest accuracy, regardless of whether adjectives precede or follow the noun.

7 Summary

We have taken a novel approach to the problem of predicting the surface order of adjectives across languages, casting it as a decision tree operating on a probability distribution over binary feature vectors. As each adjective is uttered, probability mass is partitioned into positive and negative subsets: those vectors that contain the feature and those that do not. The information gained by this partition can be used to order adjectives in a greedy manner, similarly to well-known algorithms for ordering nodes in a decision tree.

An IG-based approach allows us to provide the first quantitative information-theoretic account predicting the order of ANA triples. Further, with this approach we need not stipulate mirror- or same-orders for AAN and NAA triples. Because IG is not a distance metric between adjective and noun, and because IG incorporates negative evidence, both ANA and pre- or post-nominal asymmetries are able to emerge within an IG framework, without appeal to other mechanisms.

Our results show that information gain is a good predictor of adjective order across languages. Importantly, IG-based prediction follows a consistent pattern across the three typological templates, namely that adjectives that maximize information gain tend to be placed first.

References

Otto Behaghel. 1932. *Deutsche Syntax eine geschichtliche Darstellung*, volume IV. Carl Winters Unversitätsbuchhandlung, Heidelberg.

Guglielmo Cinque. 1994. On the Evidence for Partial N-Movement in the Romance DP. In Guglielmo Cinque, Jan Koster, Jean-Yves Pollack, Luigi Rizzi, and Raffaella Zanuttini, editors, *Paths Towards Universal Grammar. Studies in Honor of Richard S. Kayne,*, pages 85–110. Georgetown University Press, Washington, DC.

Guglielmo Cinque. 2009. *The Fundamental Left-Right Asymmetry of Natural Languages*, pages 165–184. Springer Netherlands, Dordrecht.

Guglielmo Cinque. 2010. *The Syntax of Adjectives: A Comparative Study*. The MIT Press, Camb., Mass.

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

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

David Dobkin, Truxton Fulton, Dimitrios Gunopulos, Simon Kasif, and Steven Salzberg. 1996. Induction of shallow decision trees. *submitted to IEEE PAMI*.

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

Gwendoline Fox and Juliette Thuilier. 2012. Predicting the position of attributive adjectives in the french np. In *New Directions in Logic, Language and Computation*, pages 1–15. Springer.

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.

Richard Futrell, William Dyer, and Greg Scontras. 2020a. What determines the order of adjectives in English? comparing efficiency-based theories using dependency treebanks. In *Proc. of the 58th Annual Meeting of ACL*, pages 2003–2012, Online. ACL.

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(3):e12814.

Filip Ginter, Jan Hajič, Juhani Luotolahti, Milan Straka, and Daniel Zeman. 2017. CoNLL 2017 shared task - automatically annotated raw texts and word embeddings. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Kristina Gulordava and Paola Merlo. 2015. Structural and lexical factors in adjective placement in complex noun phrases across Romance languages. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 247–257, Beijing, China. Association for Computational Linguistics.

Kristina Gulordava, Paola Merlo, and Benoit Crabbé. 2015. Dependency length minimisation effects in short spans: a large-scale analysis of adjective placement in complex noun phrases. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 477–482, Beijing, China. Association for Computational Linguistics.

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*, pages 1766–1772, Madison, WI. Cognitive Science Society.

John A. Hawkins. 1983. *Word Order Universals: Quantitative analyses of linguistic structure*. New York: Academic Press.

Robert Hetzron. 1978. On the relative order of adjectives. In Hansjakob Seiler, editor, *Language Universals*, pages 165–184. Gunter Narr Verlag, Tübingen.

Laurent Hyafil and R Rivest. 1976. Constructing optimal binary search trees is np complete. *Information Processing Letters*.

Otto Jespersen. 1922. *Language: its nature and development*. George Allen & Unwin Ltd., London.

Zeinab Kachakeche and Gregory Scontras. 2020. Adjective ordering in Arabic: Post-nominal structure and subjectivity-based preferences. In *Proc. of the LSA*, volume 5, pages 419–430.

Frank Lad, Giuseppe Sanfilippo, Gianna Agro, et al. 2015. Extropy: complementary dual of entropy. *Statistical Science*, 30(1):40–58.

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*.

Roger Levy. 2005. *Probabilistic Models of Word Order and Syntactic Discontinuity*. Ph.D. thesis, Stanford University, Stanford, CA.

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*, 21:171–93.

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.

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

Peter Matthews. 2014. *The Positions of Adjectives in English*. Oxford University Press, New York.

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

Mohammad Norouzi, Maxwell D. Collins, Matthew Johnson, David J. Fleet, and Pushmeet Kohli. 2015. Efficient non-greedy optimization of decision trees. *arXiv:1511.04056 [cs]*.

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

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.

Suttera Samonte and Gregory Scontras. 2019. Adjective ordering in Tagalog: A cross-linguistic comparison of subjectivity-based preferences. In *Proceedings of the Linguistic Society of America*, volume 4, pages 1–13.

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*, pages 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.

Gregory Scontras, Judith Degen, and Noah D. Goodman. 2019. On the grammatical source of adjective ordering preferences. *Semantics and Pragmatics*, 12(7).

Gary-John Scott. 2002. Stacked adjectival modification and the structure of nominal phrases. In *Functional Structure in DP and IP: The Cartography of Syntactic Structures*, volume 1, pages 91–210. Oxford University Press, New York.

Yuxin Shi and Gregory Scontras. 2020. Mandarin has subjectivity-based adjective ordering preferences in the presence of 'de'. In *Proceedings of the Linguistic Society of America*, volume 5, pages 410–418.

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

Richard Sproat and Chilin Shih. 1991. The Cross-Linguistic Distribution of Adjective Ordering Restrictions. In Carol Georgopoulos and Roberta Ishihara, editors, *Interdisciplinary Approaches to Language*, pages 565 – 93. Kluwer Academic Publishers, Boston.

Milan Straka and Jana Straková. 2017. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *Proc. of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver. ACL.

Henry Sweet. 1900. *A new English grammar, logical and historical*, volume 1. Clarendon Press, Oxford.

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

Juliette Thuilier. 2014. An Experimental Approach to French Attributive Adjective Syntax. In Christopher Piñón, editor, *Empirical Issues in Syntax and Semantics*, volume 10 of *Experimental Syntax and Semantics*, pages 287–304.

Nitzan Trainin and Einat Shetreet. 2021. It's a dotted blue big star: on adjective ordering in a post-nominal language. *Language, Cognition and Neuroscience*, 36(3):320–341.

Benjamin Lee Whorf. 1945. Grammatical Categories. *Language*, 21(1):1–11.

Daniel Zeman, Joakim Nivre, ... Abrams, and Anna Zhuravleva. 2020. Universal dependencies 2.7. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Daniel Zeman, Martin Popel, ..., and Josie Li. 2017. Multilingual parsing from raw text to universal dependencies. In *Proc. of CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19. ACL.