Investigating Cross-Linguistic Adjective Ordering Tendencies with a Latent-Variable Model

Jun Yen Leung🐳 **Guy Emerson**🐳 **Ryan Cotterell**🐳,🦈
🐳University of Cambridge 🦈ETH Zürich
junyenle@gmail.com, gete2@cam.ac.uk, ryan.cotterell@inf.ethz.ch

Abstract

Across languages, multiple consecutive adjectives modifying a noun (e.g. “the big red dog”) follow certain unmarked ordering rules. While explanatory accounts have been put forward, much of the work done in this area has relied primarily on the intuitive judgment of native speakers, rather than on corpus data. We present the first purely corpus-driven model of multi-lingual adjective ordering in the form of a latent-variable model that can accurately order adjectives across 24 different languages, even when the training and testing languages are different. We utilize this novel statistical model to provide strong converging evidence for the existence of universal, cross-linguistic, hierarchical adjective ordering tendencies.

1 Introduction

Most native speakers of a language would agree that certain adjective orderings are preferable to others. For instance, in English, “the big red dog” sounds natural while “the red big dog” sounds very awkward. Similar ordering preferences have been found to apply universally across the languages in the world: for example, the adjective for “big” in most languages tends to be farther away from the noun, syntactically, than “red.” For an overview of these phenomena, see Cinque (2010).

There are many explanatory accounts of cross-linguistic adjective ordering in the linguistics literature, the most popular being hierarchical tendencies based on semantic categories of adjectives (Dixon, 1982; Sproat and Shih, 1991; Cinque, 1994, 2010). For instance, Sproat and Shih (1991) and Cinque (2010) note that adjectives describing SIZE tend to be placed further from the noun than those describing COLOR in most languages. However, most of these studies have relied primarily on the judgment of native speakers rather than on corpus data, and those corpus-based models that do exist have focused exclusively on English (Shaw and Hatzivassiloglou, 1999; Malouf, 2000; Wulff, 2003; Mitchell, 2009; Dunlop et al., 2010; Mitchell et al., 2011; Hill, 2012; Scontras et al., 2017; Hahn et al., 2018; Futrell et al., 2020). In this paper, we make use of tools and techniques from statistical modeling to provide strong converging evidence supporting a hierarchical theory of cross-linguistic adjective ordering.

Specifically, we present a novel interpretable, multi-lingual, latent-variable model of adjective ordering that directly enforces a hierarchy of semantic classes and is trained entirely using corpus data. We empirically show that our model accurately orders adjectives across 24 different languages, even when tested on languages that it has not been trained on. In doing so, we demonstrate the existence of universal, cross-linguistic, hierarchical tendencies in adjective ordering.

2 Adjective Ordering

Consider the following English phrases, taken from Teodorescu (2006):

(1) A beautiful small black purse
(2) a. # A beautiful black small purse[1]
b. # A small beautiful black purse
c. # A small black beautiful purse

None of these phrases are ungrammatical, yet most native English speakers would contend that only (1) is correct in most contexts. Further complicating the phenomenon, there are many unmarked cases where ordering rules can be broken without hurting correctness. For example, now consider:

(3) A brown Chinese bear
(4) A Chinese [brown bear][2]

[1] # denotes an infelicitous phrase

[2] [ ] denotes an adjective-noun collocate

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 4016–4028,
November 16–20, 2020. ©2020 Association for Computational Linguistics

Here, (3) presents the most natural ordering of "brown" and "Chinese" (to illustrate this, substitute "bear" with "house"), but (4) is also correct because a "brown bear" is an adjective–noun collocate. For a more detailed discussion on adjective ordering exceptions, see Teodorescu (2006).

2.1 Common Theories

All adjective ordering theories put adjectives on a scale. What differentiates them is the granularity of that scale and the metric used to rank adjectives. This section describes the most notable theories, which appeal to a hierarchy of semantic classes, inherentness, modification strength, and subjectivity. We adopt the hierarchical approach in this paper because it is more general and so allows a closer fit to the data. While the more functional explanations (i.e. inherentness, modification strength, and subjectivity) might allow us to derive a hierarchy from something more fundamental, current theories only appear to account for a portion of adjective ordering preferences.

Hierarchical theories. Hierarchical theories of adjective ordering posit that each adjective belongs to a class of semantically similar adjectives, and that these classes follow a rigid order. Several theories describing how prenominal adjective classes are ordered have been suggested, most famously Cinque (2010)'s: VALUE → SIZE → SHAPE → COLOR → PROVENANCE. Dixon (1982) observes that postnominal adjectives follow the opposite order as do prenominal ones. To illustrate, consider the following phrase in both English and Spanish:

(5) An ugly black shirt

(6) Una camisa negra fea
a shirt black ugly

Inherentness. The inherentness theory (Whorf, 1945) posits that adjectives fall into two broad categories: adjectives that describe inherent properties of nouns—such as color, material, physical state, provenance, breed, nationality, function, use, etc.—and adjectives that describe non-inherent properties, and that inherent adjectives are usually placed closer to the noun than non-inherent ones.

Modification strength. Vecchi et al. (2013) apply a compositional distributional semantics approach to studying English adjective–adjective–noun phrases, and note that in correctly ordered phrases, the adjective closer to the noun contributes more to the meaning of the phrase than does the adjective further from the noun. For instance, "different architectural style" is more similar to "architectural style" than it is to "different style".

Subjectivity. The subjectivity theory (Hill, 2012; Scontras et al., 2017; Hahn et al., 2018) ranks adjectives by subjectivity on a continuous scale and posits that the less subjective an adjective is, the closer it should be placed to the noun.

2.2 Binomial Ordering

A closely related phenomenon to adjective ordering is binomial ordering. Binomials are pairs of words joined by a conjunction, such as "salt and pepper" or "ball and chain". Adjective ordering and binomial ordering have been studied in similar ways, and have in many cases been found to behave similarly (Benor and Levy, 2006; Copestake and Herbelot, 2011; Ivanova and Levy, 2018).

3 A Latent-Variable Model

A natural mathematical formalization of adjective ordering is as a latent-variable model. A latent-variable model relates a set of observable variables to a set of unobservable (latent) ones. Here, we observe how adjectives are ordered in corpus data and from this infer an ordered set of latent adjective classes. This allows us to determine the ordering of an arbitrary set of adjectives by referencing their class memberships and the class order.

Like other latent-variable models, such as latent semantic analysis (Dumais et al., 1988) and latent Dirichlet allocation (Blei et al., 2003), our model aims to fit the data using a lower-dimensional space. In particular, the number of adjective classes is much smaller than the size of the vocabulary or the size of the pre-trained adjective embeddings.

3.1 Ordering English Adjectives

Consider an English noun phrase where k unique adjectives $\mathbf{a} = \{a_1, \ldots, a_k\}$ modify a noun n, and $k \geq 2$. Let $\mathcal{C}$ be an ordered set of latent adjective classes labeled $[1, 2, \ldots, |\mathcal{C}|]$ and let d be the dimensionality of our pre-trained word embedding vectors $\mathbf{e}(\cdot)$. Our goal is to simultaneously learn a mapping $\mathbf{V} \in \mathbb{R}^{d \times |\mathcal{C}|}$ from adjective embeddings to latent classes and learn an interaction matrix $\mathbf{W} \in \mathbb{R}^{|\mathcal{C}| \times |\mathcal{C}|}$ which reflects the preferred ordering of those classes.

We develop a probabilistic model of each of the $k!$ possible permutations $\boldsymbol{\pi}$ of $\mathbf{a}$ as in eq. (1), which factorizes the distribution in terms of the latent classes $\mathbf{c}$ (a k-length tuple of class labels, one per adjective in the permutation). The i^{th} class, c_i, denotes the class assigned to the i^{th} adjective, a_i.

$$p(\boldsymbol{\pi} \mid \mathbf{a}) = \sum_{\mathbf{c} \in \mathcal{C}^k} p(\boldsymbol{\pi} \mid \mathbf{c}) \prod_{i=1}^{k} p(c_i \mid a_i) \tag{1}$$

Given latent classes, the distribution over permutations is given in eq. (2), using the scoring function in eq. (3), where π_i indexes the adjective in the i^{th} position of the permutation, and so c_{π_i} is the latent class in the i^{th} position. Thus, eq. (3) sums the ordering preference scores between each consecutive pair of adjective classes in the permutation, using the pairwise preferences in $\mathbf{W}$. Using these scores, eq. (2) produces a distribution, normalizing over the set of all permutations S_k:

$$p(\boldsymbol{\pi} \mid \mathbf{c}) = \frac{\exp \operatorname{score}(\boldsymbol{\pi}, \mathbf{c})}{\sum_{\boldsymbol{\pi}' \in S_k} \exp \operatorname{score}(\boldsymbol{\pi}', \mathbf{c})} \tag{2}$$

$$\operatorname{score}(\boldsymbol{\pi}, \mathbf{c}) = \sum_{i=1}^{k-1} W_{c_{\pi_i}, c_{\pi_{i+1}}} \tag{3}$$

Finally, the distribution over latent classes is obtained with $\mathbf{V}$, making use of a pre-trained embedding $\mathbf{e}(a_i)$ for each adjective:

$$p(c_i \mid a_i) = \operatorname{softmax}\left(\mathbf{V}\,\mathbf{e}(a_i)\right)_{c_i} \tag{4}$$

To summarize, we compute the probability of each permutation by considering all possible assignments of latent classes. The probability of a permutation is a weighted sum (eq. (1)) of normalized scores (eqs. (2) and (3), using $\mathbf{W}$), weighted according to the likelihood of the latent classes (eq. (4), using $\mathbf{V}$). Both $\mathbf{W}$ and $\mathbf{V}$ are learned through batch gradient descent.

To predict an ordering, we enumerate all permutations of $\mathbf{a}$, compute their probabilities as described, and pick the highest scoring one.

3.2 Enforcing a Total Ordering

Hierarchical theories imply a total ordering of adjective classes. This means that the class order is antisymmetric, transitive, and a connex relation. While it is likely that our model learns a (predominantly) total ordering, we cannot be absolutely sure that it does. To remedy this, we enforce a total ordering of categories by modifying our model such that $\mathbf{W}$ is no longer learned, but is instead fixed as a matrix with ones above the diagonal and zeroes elsewhere. We will refer to this as an off-upper-triangular matrix. To illustrate how this enforces a total ordering, recall that each element W_{ij} of $\mathbf{W}$ represents a preference for ordering class i before class j. Then, given a $|\mathcal{C}| \times |\mathcal{C}|$ off-upper-triangular matrix of ones and zeroes:

$$\begin{bmatrix} 0 & 1 & 1 & \dots & 1 \\ 0 & 0 & 1 & \dots & 1 \\ 0 & 0 & 0 & \dots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix}$$

Class 1 precedes classes $2, 3, \dots, |\mathcal{C}|$; class 2 precedes classes $3, 4, \dots, |\mathcal{C}|$; etc. To distinguish between the previously described variant where $\mathbf{W}$ is learned and this one, we will refer to the former as the **English Learned-W model (EL)** and the latter as the **English Fixed-W model (EF)**.

3.3 Handling Postnominal Adjectives

In English, noun phrases consisting of a noun and one or more adjectives always place the adjectives *before* the noun. However, this is not the case in other languages, where the adjectives can be placed before, after, or both before and after the noun. As such, we need to modify our model to accommodate such structures.

With the EL and EF models, we use a single interaction matrix $\mathbf{W}$ to score a permutation $\boldsymbol{\pi}$ of the adjectives $\mathbf{a} = \{a_1, \dots, a_k\}$ that modifies n. But if we must now support adjectives both before and after the noun, we must decompose $\mathbf{a}$ into two sets: $\mathbf{a}^{(\text{left})} = \{a_1^{(\text{left})}, \dots, a_j^{(\text{left})}\}$ and $\mathbf{a}^{(\text{right})} = \{a_1^{(\text{right})}, \dots, a_\ell^{(\text{right})}\}, j \geq 2$ or $\ell \geq 2$. Then, we can use two separate $\mathbf{W}$ matrices, $\mathbf{W}^{(\text{left})}$ and $\mathbf{W}^{(\text{right})}$, to score the adjectives that appear directly to the left and right of n, respectively:

$$\begin{aligned} \operatorname{score}(\boldsymbol{\pi}, \mathbf{c}) &= \operatorname{score}(\boldsymbol{\pi}^{(\text{left})}, \mathbf{c}^{(\text{left})}) \\ &\quad + \operatorname{score}(\boldsymbol{\pi}^{(\text{right})}, \mathbf{c}^{(\text{right})}) \end{aligned} \tag{5}$$

Conveniently, maximizing $\operatorname{score}(\boldsymbol{\pi}, \mathbf{c})$ is equivalent to maximizing $\operatorname{score}(\boldsymbol{\pi}^{(\text{left})}, \mathbf{c}^{(\text{left})})$ and $\operatorname{score}(\boldsymbol{\pi}^{(\text{right})}, \mathbf{c}^{(\text{right})})$ independently.

As with English, we present two variants of the multi-lingual model, one where $\mathbf{W}^{(\text{left})}$ and $\mathbf{W}^{(\text{right})}$ are learned and one where they are fixed. We will refer to the former as the **Multi-lingual Learned-W model (ML)** and the latter as the **Multi-lingual Fixed-W model (MF)**. The primary challenge in implementing MF is deciding what $\mathbf{W}^{(\text{right})}$ should be. While $\mathbf{W}^{(\text{left})}$ can simply be an off-upper-triangular matrix, as $\mathbf{W}$ is in EF, we need an appropriate matching $\mathbf{W}^{(\text{right})}$ that captures the different treatment given to prenominal and postnominal adjectives. Ultimately, we adopt Dixon (1982)'s observation that postnominal adjectives follow the opposite order as do prenominal ones, and fix $\mathbf{W}^{(\text{right})}$ as a matrix with ones below the diagonal and zeroes elsewhere, i.e. an off-lower-triangular matrix.

3.4 Multi-lingual Word Embeddings

In order to predict adjective order across languages, we need a joint model for word representations. We use multi-lingual fastText (Bojanowski et al., 2017) Wikipedia supervized word embeddings of dimensionality $d = 300$ aligned in a single vector space (MUSE), provided by Conneau et al. (2018).

4 Data

This section describes our English, multi-lingual, and additional languages datasets.

4.1 English Dataset

Multi-adjective noun phrases are surprisingly rare; analyzing 54,478 English noun phrases from the Universal Dependencies (UD) project (Nivre et al., 2016; Zeman et al., 2019), we find that only 745 of them (1.37%) contain two or more adjectives. As such, we require a large corpus to train our model. The data comprising the English dataset comes from ukWaC (Baroni et al., 2009), an enormous (>2 billion words) corpus of automatically tagged and dependency-parsed online text from the .uk domain. Unfortunately, ukWaC contains a lot of low-quality data, including non-English characters, incorrect tokenization, and part-of-speech errors.

We first extract all noun phrases where a noun is modified by multiple consecutive adjectives, i.e. all phrases consisting of an ordered set of consecutive adjectives $[a_1, \ldots, a_k], k \geq 2$, directly preceding a noun n. We then disqualify all noun phrases where more than six adjectives modify a noun, because we find that such samples tend to consist of bad data, such as "......." annotated as a sequence of adjectives. Finally, MUSE fastText embeddings are only released as word–embedding dictionaries, unlike standard fastText embeddings which are built from substrings of characters. Thus, unlike conventional fastText embeddings, they are unable to infer embeddings for unseen words. And so, we need to disqualify all noun phrases which include adjectives not in these dictionaries.

	Split by Token	
	# Phrases	# Adj Types
Training	10,000	2,695
Testing	1,000	806
Total	11,000	2,786
	Split by Type	
	# Phrases	# Adj Types
Training	9,165	2,514
Testing	1,835	890
Total	11,000	2,786

Table 1: English dataset summary.

We then randomly select 12,000 phrases. Of these, 1,000 are set aside as a development set. The remaining 11,000 phrases are split in two different ways: by **token** and by **type**. Splitting by token is done by randomly picking 10,000 phrases to form the training set and letting the remaining 1,000 phrases form the testing set. Splitting by type is done by randomly picking 90% of the unique adjective types in the data, letting all phrases where *all* their adjectives belong to this 90% form the training set, and letting the remaining phrases form the testing set. This ensures that every phrase in the testing set will contain at least one adjective not present in the training set. A summary of the English dataset can be found in Tab. 1.

4.2 Multi-Lingual Dataset

Because our multi-lingual models are trained on multiple languages at once, we do not need as many data per language and can afford to use much smaller corpora. We obtain the non-English data used to train ML and MF from UD. UD provides treebanks with annotated dependencies in many languages, which we use to determine which adjectives are modifying which nouns. The English portion of this dataset re-uses the ukWaC corpus.

For each language that we choose to include, we once again extract all noun phrases where a noun is modified by multiple consecutive adjectives. This time, however, we need to account for postnominal adjectives as well. We extract all phrases where an ordered set of consecutive adjectives $[a_1^{(\text{left})}, \ldots, a_j^{(\text{left})}], j \geq 2$, precedes n or an ordered set of consecutive adjec-

	Czech	
	# Phrases	# Adj Types
Training	5,000	2,065
Testing	1,000	820
Total	6,000	2,245
	English	
	# Phrases	# Adj Types
Training	5,000	1,930
Testing	1,000	806
Total	6,000	2,092
	German	
	# Phrases	# Adj Types
Training	5,000	1,835
Testing	1,000	743
Total	6,000	2,040
	Russian	
	# Phrases	# Adj Types
Training	5,000	1,814
Testing	667	602
Total	5,680	1,920

Table 2: Multi-lingual dataset summary.

Language	# Phrases	# Adj Types
Bulgarian	584	508
Catalan	503	515
Croatian	922	666
Danish	118	133
Dutch	321	328
Estonian	509	503
Finnish	250	254
French	621	612
Greek	104	132
Hebrew	147	170
Hungarian	228	321
Italian	397	419
Norwegian	756	543
Polish	408	508
Portuguese	222	275
Slovak	277	348
Slovenian	460	478
Spanish	1,000	947
Swedish	164	188
Ukrainian	373	472

Table 3: Additional languages dataset summary.

tives $[a_1^{(\text{right})}, \ldots, a_\ell^{(\text{right})}], \ell \geq 2$, follows n. We then once again disqualify all noun phrases which include adjectives not in the MUSE fastText dictionary. From the remaining pool, we randomly select 5,000 phrases to form our training set and 1,000 phrases to form our testing set, except for Russian, where we only have 667 phrases remaining to construct the testing set. A summary of the multi-lingual dataset can be found in Tab. 2.

Criteria for Choosing Languages. We have two criteria for choosing languages for this dataset. Firstly, the language must have MUSE fastText embeddings, as we require embeddings aligned in a common vector space. Secondly, the UD corpora for the language must contain over 5,000 usable multi-adjective noun phrases to provide a sufficiently large training set.

4.3 Additional Languages Dataset

A glaring limitation of our multi-lingual dataset is that it is not typologically diverse: it contains two Germanic and two Slavic languages. Most critically, we note that in all four of its languages, adjectives predominantly precede the noun. While we are unable to train on more languages due to a lack of data, there is no reason why we cannot test on them. The additional languages dataset consists of phrases from 20 additional MUSE-supported languages using their UD corpora and the same preprocessing pipeline as described in §4.2. Among these are three Uralic languages (Estonian, Finnish, Hungarian) and one Afro-Asiatic language (Hebrew), while the rest are Indo-European. We do not include Arabic because its MUSE fastText embeddings seem to be incorrectly formatted. We also choose not to include Indonesian, Macedonian, Romanian, Turkish, or Vietnamese because they have too few (<50) phrases to construct a representative testing set. Meta-data describing the additional languages dataset can be found in Tab. 3.

5 Experimental Details and Hyperparameters

We split our experiments into English experiments (§6) and transfer learning experiments (§7). All of our models are trained for a single epoch of the relevant training data with a learning rate of 0.1 and a batch size of 32; we found a single epoch more than sufficient for our purposes in preliminary experimentation. We also set $|\mathcal{C}| = 15$ and $d = 300$

	EL	EF	Random
Token split	0.843	0.823	0.483
Type split	0.836	0.829	0.482

Table 4: English accuracy on different data splits. Comparing the two models on the same data split, the results do not differ significantly.

	EL	EF	Random
Scrambled	0.791	0.797	0.483
Unscrambled	0.784	0.797	0.483

Table 5: English accuracy with scrambled and unscrambled fastText vectors. Comparing different vectors for the same model, the results do not differ significantly.

for all models. We report the exact expectation of the random baseline. All significance testing is done with permutation tests following Dror et al. (2018), using 10,000 random permutations and significance at $\alpha = 0.05$. All differences between model performance and the corresponding random baselines are significant with $p < 0.01$.

6 English Experiments

Our English experiments serve to demonstrate the basic correctness of the model. We also provide a qualitative analysis of EF.

6.1 Predictive Accuracy

We train each of the English models on the token and type split English data described in §4.1. The token split allows us to evaluate the basic predictive accuracy of EL and EF, while the type split allows us to evaluate how well the EL and EF models generalize to unseen adjective types. Results are detailed in Tab. 4. We achieve high accuracy on both the token split and type split data, demonstrating the correctness of the model. Importantly, our strong performance on the type split data demonstrates that EL and EF generalize well to unseen adjective types. We also observe that EL and EF results are similar, suggesting that adjective ordering preferences naturally tend towards a total ordering, since learning $\mathbf{W}$ did not significantly improve results.

6.2 Validating Use of fastText

We now address a potential confounding influence of the pre-trained fastText embeddings. We are concerned that adjective ordering information may be pre-baked into the MUSE fastText embeddings that we use, since the embeddings were trained on text where adjectives were correctly ordered. To check this, we retrain two small fastText models on a subset of 12,500 sentences from ukWaC. The first model is trained on these sentences as they are, and the second model is trained on a version of these sentences where strings of consecutive adjectives have been randomly scrambled. We then retrain the EL and EF models on the token split data with both the scrambled and unscrambled fastText vectors. Results are detailed in Tab. 5.

That neither pair of scrambled and unscrambled results differs significantly indicates that adjective ordering information is *not* coming from the fastText embeddings. Otherwise, the unscrambled model should have outperformed the scrambled model. Due to the computational expense of retraining multi-lingual fastText, we do not repeat this validation with the multi-lingual models.

6.3 Qualitative Evaluation of EF

Perhaps the most convenient property of the EF model is that it is fully interpretable. We are able to, for any given adjective, extract information about which class it belongs to, and know from the model's design that classes follow a total ordering such that class 1 precedes class 2 precedes class 3, and so on. In this experiment, we first qualitatively analyze the 177 testing phrases in the token split data that EF orders incorrectly, making generalizations about what kinds of mistakes the model makes. We then make a qualitative comparison between the hierarchy that EF learns and the hierarchy proposed by Cinque (2010).

Types of Mistakes. Two types of cases account for most of EF's mis-orderings. Firstly, many of the mis-ordered testing phrases deviate from typical adjective ordering tendencies because they contain adjective–noun collocates. Such phrases include "Italian [secret service]", "modern [good practice]", and "Japanese [popular culture]" (to illustrate how these are atypical, consider "secret Italian meatballs", "good modern ethics", and "popular Japanese restaurant"). We note that this tends to occur with adjectives that describe PROVENANCE: these, while typically placed near the noun, are also often prepended to collocates. We are largely unsurprised by this, as it mirrors the intuitive observations made regarding adjective-noun collocates illustrated in (3) and (4). An interesting direction for future work might be to model the likelihood of an adjective and a noun forming a collocate and integrate that into our current model.

Secondly, we observe that EF often mis-orders phrases containing adjectives describing ORDER (e.g. “next”, “first”, “other”) and QUANTITY (e.g. “few”, “many”). Examining EF’s adjective-class layer, we discover that it has placed these words together in the same class, when intuitively ORDER adjectives should precede QUANTITY adjectives (e.g. “next few lessons”, “first many partners”). Further experimentation would be necessary to determine why EF has done this, but we suspect intuitively that it may be because ORDER and QUANTITY adjectives are relatively small classes and are semantically similar. If they occur more often next to other classes than next to each other, there is only a weak pressure for the model to assign these words to distinct classes. A more rigorous error analysis would require a comprehensive dictionary of adjectives tagged with their semantic classes.[3] Unfortunately, constructing such a dictionary is beyond the scope of this paper.

Comparison with Cinque’s Hierarchy. We take the 100 most common adjectives in the English dataset and use EF’s adjective-class layer to determine their class memberships. We then compare these classes and their relative orderings to those proposed by Cinque (2010): VALUE → SIZE → SHAPE → COLOR → PROVENANCE.

We observe that EF follows most of Cinque’s rules. Most notably, EF clearly learns categories of adjectives describing SIZE, COLOR and PROVENANCE, and additionally learns that SIZE precedes COLOR precedes PROVENANCE. We perform a small-scale statistical verification of this observation by hand-constructing a testing set of Cinquean phrases and using it to evaluate the similarity of EF’s and Cinque’s predictions. To do this, we first select five common adjectives from each of the five Cinquean categories. We then construct a testing set using pairs of only these 25 adjectives based on Cinque’s hierarchy. This gives us $\binom{5}{2} * 5^2 = 250$ testing phrases. Since these are all pairs, the expected random baseline is simply 50%.

We then evaluate the predictive accuracy of EF on the Cinquean testing phrases. EF achieves an accuracy of 0.960 with $p < 0.01$, suggesting that EF agrees with most of Cinque’s rules. Importantly, this does *not* mean that EF is 96% accurate at ordering adjectives, but only that EF agrees with 96% of Cinque’s predictions on our test set. As discussed, many of EF’s mistakes on real corpus data are attributable to adjective ordering exceptions like adjective-noun collocates, which Cinque’s hierarchy does not address either.

While EF follows most of Cinque’s *existing* rules, we also observe that EF learns *additional* rules not described by Cinque. For instance, EF seems to learn a category of adjectives describing TYPE, which follows adjectives describing PROVENANCE and contains adjectives such as “financial”, “technical”, and “scientific”. This seems intuitively correct—to illustrate, consider “Russian financial burden”, “German technical wonder”, and “African scientific achievement”. This suggests that an accurate adjective ordering hierarchy may need to be more complex than described by Cinque. In particular, it seems that Cinque’s adjective classes are too broad. An alternate interpretation is that TYPE adjectives are defined by being capable of forming adjective-noun collocates with most of the nouns that they commonly modify.

But we must emphasize that this analysis is still anecdotal. The noted similarities and differences are difficult to quantify, and as far as we are aware there is no large-scale corpus of adjectives tagged with their Cinquean categories to enable a more reliable quantitative approach; we would ideally want such a corpus in a large number of languages. For now, we simply suggest that while Cinque’s hierarchy captures many truths about adjective ordering, it does not quite grasp the entire picture.

Comparison with Functional Theories The bulk of the existing work on statistically modelling adjective ordering can be broadly separated into two categories: that which is *theoretically-motivated* (e.g. Wulff, 2003; Futrell et al., 2020), and that which is *empirically-motivated* (e.g. Malouf, 2000). The theoretically-motivated approach attempts to deduce the source of adjective ordering preferences by fitting adjective ordering data to pre-determined features derived from more fundamental functional pressures. The empirically-motivated approach attempts to fit adjective ordering data as accurately as possible by learning features from data. This paper falls into the empirically-motivated category because a hierar-

[3] Specifically, these would have to be semantic classes comparable with those learned by EF.

	ML (Learned $\mathbf{W}$)			MF (Fixed $\mathbf{W}$)			Random
	Transfer	Mono-ling	Joint	Transfer	Mono-ling	Joint	
Czech	$0.851^{\dagger\ddagger*}$	$0.886^{\dagger*}$	0.899	$0.817^{\ddagger*}$	$0.831^{\dagger*}$	0.888	0.483
English	$0.803^{\dagger\ddagger}$	0.820	0.820	0.800	0.811	0.808	0.487
German	$0.695^{\dagger\ddagger*}$	0.802	0.807	$0.732^{\dagger\ddagger*}$	0.796	0.807	0.488
Russian	$0.840^{\dagger\ddagger}$	$0.893^{\dagger}$	0.911	$0.859^{\ddagger}$	$0.873^{\dagger}$	0.892	0.485

Table 6: Multi-lingual accuracy. A † denotes that a result differs significantly from the result to its right. A ‡ denotes that a result differs significantly from the result two to its right. A * denotes that a result differs significantly from its ML/MF counterpart. The terminology used to describe the columns is defined in §7.1.

chical model like ours or Cinque's is in no way functional – it postulates that a particular hierarchy exists but does not explain why it exists in that particular order. Importantly, this means that a hierarchical theory is not necessarily at odds with the functional theories. Rather, it is very possible that one or more functional theories might serve to explain the empirically observed hierarchies.

Interestingly, there seems to be a gap in predictive accuracy between theoretically-motivated and empirically-motivated models. For example, Wulff (2003) and Futrell et al. (2020) achieve accuracies in the low 70s, while Malouf (2000) and this paper achieve accuracies in the 80s. While these results are hard to compare directly as they were achieved on different datasets, this suggests that there are some ordering preferences not yet captured by any existing functional theory.

7 Transfer Learning Experiments

An important claim of the hierarchical theory for adjective ordering is that the hierarchy applies universally across languages. If this is the case, then we should be able to accurately order adjectives from languages that we have not trained on.

7.1 Predictive Accuracy

We evaluate each of the multi-lingual models on the multi-lingual dataset in three different scenarios. The first scenario (henceforth the **mono-lingual** scenario) addresses single-language training and testing. For this, we train one model on each of the four languages in the dataset by itself. Each model is then tested on the language that it was trained on. The second scenario (henceforth the **transfer** scenario) addresses the model's ability to generalize to unseen languages by holding out the language in question. For this, we train four models, each on every language but the one we want to test (e.g. on Czech, English, German, but not Russian). Each model is then tested on the language that was held out during training. The third scenario (henceforth the **joint** scenario) addresses the potential for augmenting single-language training with additional data from other languages. For this, we train a single model on all four languages together. The model is then tested on each of the four languages individually. Results are detailed in Tab. 6.

We observe that the model performs much better than chance on the transfer scenario. This confirms the theory that universal hierarchical adjective ordering tendencies generalize across languages. Otherwise, we would expect chance level performance. We also observe that for all languages, performance on the joint scenario is better than or equal to performance on the mono-lingual scenario, which is in turn better than or equal to performance on the transfer scenario. This upward trend of transfer $\leq$ mono-lingual $\leq$ joint suggests that while training on additional languages can help performance, the most important single factor is to train on the language that is being tested. In fact, given that the multi-lingual models did not achieve the same performance on the joint scenario as the English models did on the English dataset (§6.1), we predict that performance on the mono-lingual scenario would have been the best for all languages if there had been more training data. Finally, we observe that for the most part, corresponding ML and MF results do not differ significantly, suggesting once again that adjective ordering preferences tend towards a total ordering. Taken together, these observations suggest that a universal hierarchy of adjective ordering tendencies exists, though individual languages may also feature additional unique tendencies not shared by the others.

Language	Family	Accuracy	Random
Bulgarian	Slavic	0.851	0.487
Catalan	Romance	0.763	0.494
Croatian	Slavic	0.850	0.487
Danish	Germanic	0.791	0.492
Dutch	Germanic	0.819	0.488
Estonian	Finnic	0.673	0.493
Finnish	Finnic	0.702	0.493
French	Romance	0.802	0.490
Greek	Greek	0.832	0.490
Hebrew	Semitic	0.868	0.493
Hungarian	Ugric	0.839	0.466
Italian	Romance	0.740	0.493
Norwegian	Germanic	0.797	0.480
Polish	Slavic	0.779	0.500
Portuguese	Romance	0.722	0.491
Slovak	Slavic	0.770	0.475
Slovenian	Slavic	0.818	0.485
Spanish	Romance	0.771	0.491
Swedish	Germanic	0.769	0.492
Ukrainian	Slavic	0.833	0.487

Table 7: MF accuracy on additional languages.

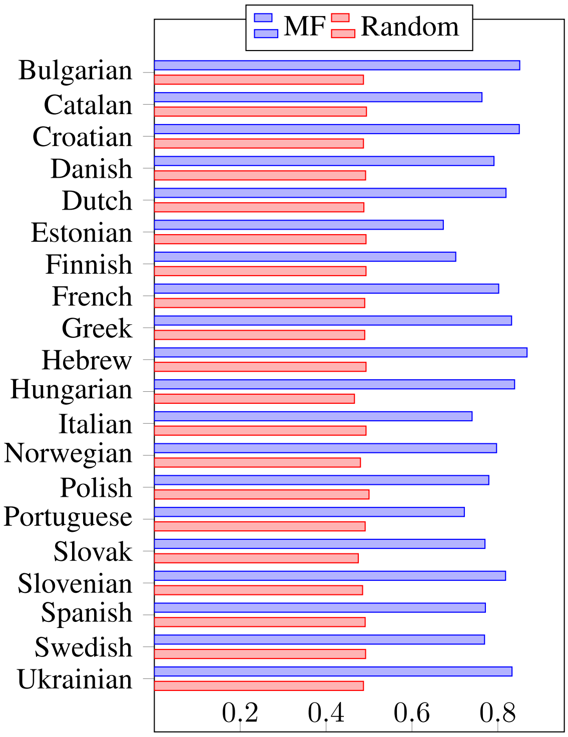

Figure 1: MF accuracy on additional languages.

7.2 Testing on Additional Languages

To build confidence that our findings truly generalize widely across typologically diverse languages, we train the MF model on Czech, English, German, and Russian, and test it on each of the languages in the additional languages dataset. We choose to test only the MF model as the ML model would not have the data to learn a correct $\mathbf{W}^{(\text{right})}$ matrix (as Czech, English, German, and Russian tend not to have postnominal adjectives) and would thus understandably under-perform on the languages which predominantly feature postnominal adjectives (i.e. Catalan, French, Hebrew, Italian, Portuguese, and Spanish). This experiment is conceptually identical to the multi-lingual transfer scenario. Results are detailed in Tab. 7 and visualized in Fig. 1.

MF performs much better than chance on every language, with similar accuracies as those achieved in the transfer scenario. This gives us confidence that the conclusions drawn in §7.1 do generalize widely across typologically diverse languages.

8 Conclusion

We built an interpretable, multi-lingual latent-variable model of hierarchical adjective ordering that directly enforces a hierarchy of semantic classes and is trained entirely using corpus data. We found that our fixed-W variants, which enforce total orderings of semantic classes, perform similarly to our learned-W variants, suggesting that adjective ordering preferences naturally tend towards total orderings. We also found that our model is able to accurately order adjectives from 24 different languages, regardless of whether it was directly trained on them, although it does benefit from having been trained on the language on which it is tested. Interestingly, we were able to achieve high predictive accuracy on languages predominantly featuring postnominal adjectives (e.g. French, Spanish), despite having only trained on languages predominantly featuring prenominal ones (Czech, English, German, Russian), by simply reversing the prenominal adjective ordering rules for postnominal ones.

In summary, our work presents converging evidence that adjectives exhibit universal hierarchical ordering tendencies, with the added observations that individual languages feature additional unique tendencies not shared by others, and that adjective ordering is symmetric with respect to the noun.

References

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Sarah Benor and Roger Levy. 2006. The chicken or the egg? A probabilistic analysis of English binomials. *Language*, 82:233–278.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(January):993–1022.

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.

Guglielmo Cinque. 1994. On the evidence for partial N-movement in the Romance DP. In Guglielmo Cinque, Jan Koster, Jean-Yves Pollock, and Rafaella Zanuttini, editors, *Paths towards universal grammar: essays in honor of Richard S. Kayne*. Georgetown University Press.

Guglielmo Cinque. 2010. *The Syntax of Adjectives. A Comparative Study*. MIT Press.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*.

Ann Copestake and Aurélie Herbelot. 2011. Exciting and interesting: issues in the generation of binomials. In *Proceedings of the UCNLG+Eval: Language Generation and Evaluation Workshop*, pages 45–53, Edinburgh, Scotland. Association for Computational Linguistics.

R.M.W. Dixon. 1982. *Where have all the adjectives gone? And other essays in semantics and syntax*. Janua linguarum: Series maior. Mouton.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

S. T. Dumais, G. W. Furnas, T. K. Landauer, S. Deerwester, and R. Harshman. 1988. Using latent semantic analysis to improve access to textual information. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '88, page 281–285, New York, NY, USA. Association for Computing Machinery.

Aaron Dunlop, Margaret Mitchell, and Brian Roark. 2010. Prenominal modifier ordering via multiple sequence alignment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 600–608, Los Angeles, California. Association for Computational Linguistics.

Richard Futrell, William Dyer, and Greg Scontras. 2020. 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, Online. Association for Computational Linguistics.

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

Felix Hill. 2012. Beauty before age? Applying subjectivity to automatic English adjective ordering. In *Proceedings of the NAACL HLT 2012 Student Research Workshop*, pages 11–16, Montréal, Canada. Association for Computational Linguistics.

Anna Ivanova and Roger Levy. 2018. Pragmatic inference of intended referents from binomial word order. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society*, pages 1862–1867.

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, Hong Kong. Association for Computational Linguistics.

Margaret Mitchell. 2009. Class-based ordering of prenominal modifiers. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, pages 50–57, Athens, Greece. Association for Computational Linguistics.

Margaret Mitchell, Aaron Dunlop, and Brian Roark. 2011. Semi-supervised modeling for prenominal modifier ordering. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 236–241, Portland, Oregon, USA. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Gregory Scontras, Judith Degen, and Noah D. Goodman. 2017. Subjectivity predicts adjective ordering preferences. *Open Mind*, 1:53–66.

James Shaw and Vasileios Hatzivassiloglou. 1999. Ordering among premodifiers. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 135–143, College Park, Maryland, USA. Association for Computational Linguistics.

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: Essays in Honor of S.-Y. Kuroda*.

Alexandra Teodorescu. 2006. Adjective ordering restrictions revisited. In *Proceedings of the 25th West Coast Conference on Formal Linguistics*, pages 399–407. Somerville, MA: Cascadilla Proceedings Project.

Eva Maria Vecchi, Roberto Zamparelli, and Marco Baroni. 2013. Studying the recursive behaviour of adjectival modification with compositional distributional semantics. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 141–151, Seattle, Washington, USA. Association for Computational Linguistics.

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

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

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Noëmi Aepli, Željko Agić, Lars Ahrenberg, Gabrielė Aleksandravičiūtė, Lene Antonsen, Katya Aplonova, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Victoria Basmov, Colin Batchelor, John Bauer, Sandra Bellato, Kepa Bengoetxea, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Kristina Brokaitė, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Savas Cetin, Fabricio Chalub, Jinho Choi, Yongseok Cho, Jayeol Chun, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Elvis de Souza, Arantza Diaz de Ilarraza, Carly Dickerson, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Hanne Eckhoff, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Aline Etienne, Wograine Evelyn, Richárd Farkas, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Griciūtė, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Nizar Habash, Jan Hajič, Jan Hajič jr., Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Kim Harris, Dag Haug, Johannes Heinecke, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Jena Hwang, Takumi Ikeda, Radu Ion, Elena Irimia, Ọlájídé Ishola, Tomáš Jelínek, Anders Johannsen, Fredrik Jørgensen, Markus Juutinen, Hüner Kaşıkara, Andre Kaasen, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Boris Katz, Tolga Kayadelen, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Elena Klementieva, Arne Köhn, Kamil Kopacewicz, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Sookyoung Kwak, Veronika Laippala, Lorenzo Lambertino, Lucia Lam, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phng Lê H`ông, Alessandro Lenci, Saran Lertpradit, Herman Leung, Cheuk Ying Li, Josie Li, Keying Li, KyungTae Lim, Maria Liovina, Yuan Li, Nikola Ljubešić, Olga Loginova, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Niko Miekka, Margarita Misirpashayeva, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Keiko Sophie Mori, Tomohiko Morioka, Shinsuke Mori, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Lng Nguy˜ên Thị, Huy`ên Nguy˜ên Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Adédayọ Olúòkun, Mai Omura, Petya Osenova, Robert Östling, Lilja Øvrelid, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Angelika Peljak-Łapińska, Siyao Peng, Cenel-Augusto Perez, Guy Perrier, Daria Petrova, Slav Petrov, Jason Phelan, Jussi Piitulainen, Tommi A Pirinen, Emily Pitler, Barbara Plank, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Peng Qi, Andriela Rääbis, Alexandre Rademaker, Loganathan Ramasamy, Taraka Rama, Carlos Ramisch, Vinit Ravishankar, Livy

Real, Siva Reddy, Georg Rehm, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Mykhailo Romanenko, Rudolf Rosa, Davide Rovati, Valentin Roșca, Olga Rudina, Jack Rueter, Shoval Sadde, Benoît Sagot, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Dage Särg, Baiba Saulīte, Yanin Sawanakunanon, Nathan Schneider, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Muh Shohibussirri, Dmitry Sichinava, Aline Silveira, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Isabela Soares-Bastos, Carolyn Spadine, Antonio Stella, Milan Straka, Jana Strnadová, Alane Suhr, Umut Sulubacak, Shingo Suzuki, Zsolt Szántó, Dima Taji, Yuta Takahashi, Fabio Tamburini, Takaaki Tanaka, Isabelle Tellier, Guillaume Thomas, Liisi Torga, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Sowmya Vajjala, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Lars Wallin, Abigail Walsh, Jing Xian Wang, Jonathan North Washington, Maximilan Wendt, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Mary Yako, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Zhuoran Yu, Zdeněk Žabokrtský, Amir Zeldes, Manying Zhang, and Hanzhi Zhu. 2019. Universal Dependencies 2.5. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

A Reproducibility

In the interest of fostering reproducibility, we provide the following additional information about our data, models, and computing infrastructure.

A.1 Data

We use Universal Dependencies (UD) 2.5 and ukWaC, which can be found at http://hdl.handle.net/11234/1-3105 and https://wacky.sslmit.unibo.it/doku.php, respectively. Note that UD has since been updated to version 2.6.

A.2 Model Parameters and Runtime

The learned-W models (EL, ML) have $2 * |\mathcal{C}|^2 + d * |\mathcal{C}| = 4,950$ learnable parameters. The fixed-W models (EF, MF) have $d * |\mathcal{C}| = 4,500$ learnable parameters. The time taken to train each model varies based on the number of training samples—as a rule of thumb, training the learned models takes about 1.5-2 hours per 10,000 samples, while training the fixed models takes about 1 hour per 10,000 samples. Training all of the model variants necessary to reproduce this paper in full takes about 24 hours. Testing either model type takes only several minutes per 1,000 samples.

A.3 Computing Infrastructure

All our development, training, and testing was done on a personal computer with the following specifications:

- Operating System: Windows 10 Pro (64-bit)
- CPU: Intel Core i7-7700k @ 4.20 GHz
- GPU: None
- RAM: 64GB DDR4
- Storage Used: Approximately 200GB

A.4 Other Notes

We did not use validation sets as we saw little value to extensively tuning the model, since we were trying to explore the properties of a natural phenomenon rather than aiming to achieve the highest possible accuracy. All reported results are from the first time each model variant was tested.