Explaining Syntactic Universals by Optimizing Grammars

(No author info supplied here, for consistency with TACL-submission anonymization requirements)

Abstract

Cross-linguistic universals of word order, and the search for explanations for these, have been a topic of intense study in linguistics. A prominent line of research has argued that these universals arise because languages are optimized for human communication and language processing, with a variety of proposals for what specific criteria that natural language might be optimized for. Here we test these explanations by formalizing and computationally implementing three such optimization criteria: dependency length minimization, predictability maximization, and parsing accuracy. We implement each criterion as an objective function to be applied to a word order grammar: a probabilistic grammar that generates word order given a fixed dependency tree topology. Using tree structures from 50 diverse languages in the Universal Dependencies corpora, we find that the optimized word order grammars recover most of the Greenberg word order correlation universals. Our results provide direct computational evidence that universals of word order can be explained as optimizing ease of human language processing.

1 Introduction

Cross-linguistic universals of word order, and the search for explanations for these, have been an important subject of research in linguistics. Greenberg (1963) first documented a set of systematic regularities in word orders across languages, such as correlations between the relative order of verbs and objects with the occurrence of pre- or postpositions. A prominent line of research has attempted to explains these regularities in terms of human communication and language processing: Properties of grammar are explained as making human language processing maximally efficient, and regularities across languages arise because these pressures originate from general principles of cognition and communication that hold across languages (**?????**). Under this view, the various human languages represent multiple solutions to the problem of efficient information transfer given human cognitive constraints.

Within this paradigm, linguists have proposed a range of criteria that are intended to quantify the ease or difficulty of processing that results from a given word order, and that are argued to be optimized by the word orders found in natural language. For example, it has been argued that minimizing the **length of syntactic dependencies** is one such overarching principle that can account for many crosslinguistic regularities, and that is grounded in psycholinguistic findings about language processing (Rijkhoff, 1986; Hawkins, 1990; Gibson, 1998; Temperley and Gildea, 2018). A second criterion is to maximize the **ease of incremental prediction**, i.e. to minimize the **surprisal** of each word in context. Surprisal is known to be a general predictor of human comprehension difficulty (Hale, 2001; Levy, 2008). A third criterion is the **ease of parsing**, i.e., the **identifiability of syntactic structure**, has also featured prominently as an explanation of word order patterns (Hawkins, 1990).

Computational work using cross-linguistic corpora has found strong evidence that languages are optimized for dependency length (Futrell et al., 2015) and predictability from local contexts (Gildea and Jaeger, 2015). However, there is a missing link between computational corpus studies and word order universals. While corpora have been shown to follow various optimization criteria, it has yet to be shown directly that these criteria lead to the proposed word order universals in practice. This link is nontrivial because each natural language contains a vast array of different phenomena that may interact in complicated ways.

We propose to evaluate optimization-based explanations of universals computationally by considering word order grammars that have been explicitly optimized for quantitative measures of language processing efficiency. First, we develop **word order grammars**: simple, interpretable probabilistic models of word order given unordered dependency tree structure. Second, we take dependency treebanks of natural languages and compute word order grammars for these trees which are optimized for objective functions representing processing efficiency. Third, we apply the optimized word order grammars to create counterfactual versions of each corpus, keeping everything else about the corpus constant, and evaluate which word order universals hold in the optimized languages.

Due to the possibility of interactions with other parts of grammar, the predictions may differ from language to language, and be impacted by other typological factors such as the presence of inflectional morphology. By carrying out this method across treebanks from many different languages with different genetic affiliations and typological properties, we can evaluate which predictions are made independently of other typological properties, and are thus predicted to be universal.

We model language processing using the neural methods that underlie state-of-the-art NLP systems, specifically neural language modeling and parsing.

Optimization of counterfactual grammars on real treebanks was previously done by Gildea and Temperley (2007), Gildea and Temperley (2010), and Gildea and Jaeger (2015), who optimized grammars for dependency length and trigram surprisal on data from several languages. Our contribution is an optimization method that is applicable more generally to such objective functions, and in particular still works when these optimization criteria are instantiated using powerful neural models.

We draw on data from the Universal Dependencies project (Nivre et al., 2017), which has been developing treebanks in a unified format for several dozens of diverse languages. This universal annotation scheme is particularly useful for studying word order universals, as many of them can be operationalized in terms of crosslinguistically valid syntactic notions.

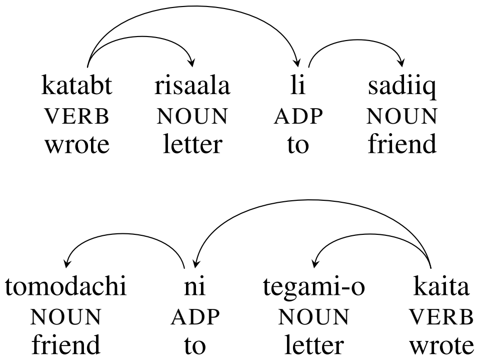

Figure 1: A sentence in Arabic (top) and Japanese (bottom), translating to ‘I wrote a letter to a friend.’ Note the reversal of word order: Arabic has verb-object order and prepositions, while Japanese has object-verb order and postpositions.

2 Explananda: Word Order Correlations

Working with a database of 30 languages, Greenberg (1963) proposed 45 universals of language that he believed to be true of most or all languages. 31 of them were concerned with word order, the remainder with morphology. Many of the word order universals state *correlations* between patterns: For instance, Universals 3 and 4 state that

‘*Languages with dominant VSO order are always prepositional.*’ (e.g. Arabic)

and

‘*With overwhelmingly greater than chance frequency, languages with normal SOV order are postpositional.*’ (e.g. Japanese)

These are illustrated in Figure 1: Arabic has basic order verb-object (VO) and prepositions (*li* ‘to’), while Japanese has order object-verb (OV) and postpositions (*ni* ‘to’). While some universals are absolute, many are *statistical* – stating that languages satsifying a correlation significantly outnumber those that do not.

Since Greenberg’s work, researchers have both tried to explain the universals, and, drawing on data from more languages, refine them. For example, a prominent explanation is in terms of head-directionality: For instance, it is generally assumed that verbs are heads of clauses, and adpositions are heads of adpositional phrases. In languages such as Japanese and Arabic, the correlations can be understood as stating that phrases are head-final (Japanese) or head-initial (Arabic). This has been formalized in the idea of a head-direction parameter in the Principles and Parameters framework (Chomsky, 1981). However, later research based on more languages has found both correlations that are not explained by head-directionality, and correlations that would be expected but are not found (Dryer, 1992). Explanations of these universals in terms of human language processing typically refer to principles akin to Dependency Length Minimization (Rijkhoff, 1986; Hawkins, 1994, 2003): Inverting the order of adposition 'to' and the noun 'letter' in either of the sentences in Figure 1 would increase the overall length of dependencies.

3 Three Objective Functions

We consider three objective functions for optimizing grammars:

Short Dependencies Psycholinguistic research points to short syntactic dependencies easing linguistic processing (Gibson, 1998; Grodner and Gibson, 2005; Demberg and Keller, 2008; Bartek et al., 2011), and quantitative corpus evidence from many languages confirms that languages have shorter dependencies than would be expected at random (Futrell et al., 2015). White and Rajkumar (2012) show that dependency length minimization improves automated ordering of constituents. Some authors have argued that dependency length minimization explains several of the Greenberg word order correlations (Rijkhoff, 1986; Hawkins, 1994, 2003).

Ease of Incremental Prediction Psycholinguistic evidence shows that humans continuously engage in incremental prediction during language comprehension. The negative log-probability, or *surprisal* of words is a linear predictor of processing difficulty as reflected in reading times (Hale, 2001; Levy, 2008; Demberg and Keller, 2008; Smith and Levy, 2013). Therefore, word orders maximizing predictability should decrease processing load (Gildea and Jaeger, 2015; Ferrer-i Cancho, 2017).

Futrell and Levy (2017) introduce a model of incremental predictability given lossy memory representations, and argue that maximization of this predictability subsumes the predictions of dependency length minimization. In contrast, Gildea and Jaeger (2015) show that five natural languages optimize trigram surprisal, but that optimizing trigram surprisal does not shorten dependencies. Our work will determine the extent to which predictability maximization and dependency length minimization make overlapping or diverging predictions.

We propose to instantiate incremental prediction using recurrent neural language models, the basis of the state-of-the-art in language modeling (Jozefowicz et al., 2016). As they compute sentence probabilities incrementally word-by-word, and have bounded-dimensionality context representation, they can be viewed as general models of incremental prediction with lossy memory representations.

Ease of Syntactic Parsing Processing explanations of dependency length minimization and word order universals have often made reference to syntactic parsing (Hawkins, 1990). While the other two objective functions deal with the *difficulty* of language processing, we can also talk about its *accuracy*: whether the parse tree can actually be recovered from a sentence.

Using the Universal Dependencies treebanks, we can directly implement ease of dependency parsing as an objective function. We instantiate ease of parsing using a very generic graph-based parsing architecture (Kiperwasser and Goldberg, 2016; Dozat and Manning, 2016) that has shown superior performance across diverse languages (Dozat et al., 2017). Given its generic nature, we expect that its architecture minimizes the risk of introducing architectural biases beyond those given by the task of dependency parsing.

4 Approach: Optimizing Word Order

Our aim is to evaluate to what extent word order universals, in particular the Greenberg correlations, can be explained by optimizing for the three objective functions described in Section 3. Our approach is to optimize the word orders of languages, keeping the hierarchical syntactic structures of sentences unchanged. All natural languages have some degree of word order regularities, and often significantly rely on word order to disambiguate syntactic structure. Thus, it is not sufficient to optimize the word orders of individual sentences in the corpora – instead, we will optimize the *word order rules* of entire languages. That is, we construct counterfactual languages that have optimized but internally consistent grammatical regularities in the domain of word order, and

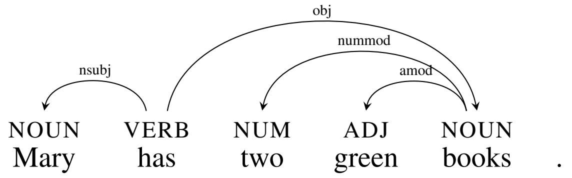

Figure 2: A Universal Dependencies tree

agree with an actual natural language in all other respects.

This requires specifying a space of possible *word order grammars* that specify how to order syntactic structures—dependency trees, in our case—into strings of words. To this end, we specify a parameteric model for linearizing dependency trees. We design our linearization model to be particularly simple to make its parameters interpretable and comparable across models.

4.1 Word Order Grammars

Our model linearizes dependency trees such as the one in Figure 2: Nodes are labeled with POS tags, and arcs are labeled with syntactic relations. POS tags and syntactic relations belong to a mostly language-independent inventory as specified by the Universal Dependencies project (Nivre et al., 2017). In our model, how a given node in the tree is linearized in relation to its head and to its siblings only depends on the POS tags of the head and the dependent, and on the syntactic relation.

More formally, for each triple τ consisting of (i) the head's POS tag, (ii) a syntactic relation, and (iii) the dependent's POS tag, we have (1) a Direction Parameter $a_\tau \in [0, 1]$, and (2) a Distance Parameter $b_\tau \in \mathbb{R}$. The Direction Parameter describes the probability that the dependent is placed before the head. The Distance Parameter encodes the distance of the dependent from the head, relative to other dependents: Siblings with greater distance parameters are likely to be less close to the head than those with smaller distance parameters. Formally: Let $\{s_1, \dots, s_n\}$ be siblings placed on the same side of the head. For each of these siblings s_i, its POS, the POS of the head, and the label of the dependency together determine a distance parameter b_{τ_i}. We compute $\mathrm{softmax}(b_{\tau_1}, \dots, b_{\tau_n})$ to obtain a distribution over the siblings. From this distribution, we iteratively sample siblings s_i without replacement, until none are left. We then linearly order the siblings in the order in which they were sampled, from left to right. For dependents following heads, we invert the sign of the b_τ's entering the softmax, approximately reversing the order in which they will be sampled.

This parameterization will make it possible to directly test word order universals such as those we discussed above: By comparing a_τ for different dependencies, we can test whether the Greenberg word order correlations hold in a counterfactual language.

4.2 Discussion and Related Work

A similar model was defined by Gildea and Temperley (2007) and subsequent work, where each dependent has a single weight in $[-1, 1]$ defining deterministically its position relative to the head and the other dependents. Ours model is more general by being probabilistic. This makes it possible to model some amount of word order freedom, and enables optimization by stochastic gradient descent.

Probabilistic ordering models more complex than ours have been defined by Futrell and Gibson (2015), Wang and Eisner (2017), and others (see e.g. Belz et al., 2011).These models rely on feature extraction or n-gram counts and are more powerful than ours. However, the simple parametric form of our model enables easy interpretation across languages.

Like these previous models, our model makes simplifying assumptions and will not fit the rules of natural languages perfectly: For instance, none of the models accounts for word order variation that is conditioned on the larger context—e.g., differences in word order between embedded and main clauses. Also, like many ordering models in the literature, our method only generates projective linearizations (that is, dependency lines are constrained not to cross).

5 Optimizing Ordering Grammars

5.1 Formalizing Objective Functions

In Section 4.1, we introduced a parametric family of ordering models linearizing dependency trees into sentences. We are interested in finding ordering models θ_d that minimize one of the three objective functions defined in Section 3. The first step will be to formalize these objective functions.

Dependency length: We sum the lengths of syntactic dependencies in the sentence:

$$R_{DepL}(\mathbf{w}) := \sum_{i=1}^{\#\mathbf{w}} |i - head(w_i)|, \tag{1}$$

where $head(w_i) \in \{1, ..., n\}$ is the index of the head of w_i.[1]

Predictability: Given a neural language model $P_{\theta_{LM}}$, with parameter vector θ_{LM}, the surprisal of a sentence $\mathbf{w}$ is:

$$R_{Pred}(\mathbf{w}, \theta_{LM}) := -\sum_{i=0}^{\#\mathbf{w}} \log P_{\theta_{LM}}(w_{i+1}|w_{1\ldots i}). \tag{2}$$

An ordering model θ_d induces a distribution over linearized sentences. The language model appropriate to θ_d is the one that models this distribution best—equivalently, the one that achieves the lowest cross-entropy on sentences linearized by θ_d:

$$\theta_{LM}(\theta_d) := \arg\min_{\theta_{LM}} \mathbb{E}_{\mathbf{w}\sim\theta_d} R_{Pred}(\mathbf{w}, \theta_{LM}). \tag{3}$$

Parsability: We express the difficulty of parsing a sentence as the log-loss of predicting the heads and dependency labels for the words in the sentence:

$$R_{Pars}(\mathbf{w}, \theta_P) := -\sum_{i=1}^{\#\mathbf{w}} \log P_{\theta_P} \left(\begin{matrix} head(w_i) \\ label(w_i) \end{matrix} | \mathbf{w} \right). \tag{4}$$

The appropriate parsing model θ_P is the model $\theta_P(\theta_d)$ that is optimally capable of parsing sentences under the distribution induced by θ_d:

$$\theta_P(\theta_d) := \arg\min_{\theta_{LM}} \mathbb{E}_{\mathbf{w}\sim\theta_d} R_{Pars}(\mathbf{w}, \theta_P). \tag{5}$$

5.2 Optimizing Grammars

For each of the three objective functions, we seek an ordering grammar θ_d that minimizes the average loss over the sentences in the counterfactual language obtained by linearizing trees according to θ_d:

$$\min_{\theta_d} \mathbb{E}_{\mathbf{w}\sim\theta_d} R_{DepL}(\mathbf{w}) \tag{6}$$

$$\min_{\theta_d} \mathbb{E}_{\mathbf{w}\sim\theta_d} R_{Pred}(\mathbf{w}, \theta_{LM}(\theta_d)) \tag{7}$$

$$\min_{\theta_d} \mathbb{E}_{\mathbf{w}\sim\theta_d} R_{Pars}(\mathbf{w}, \theta_P(\theta_d)), \tag{8}$$

where linearized sentences $\mathbf{w}$ are obtained by sampling trees from the treebank and stochastically ordering them according to the ordering model θ_d. Note that the original word orders from the treebank do not enter this objective – they in particular do not influence the language model and the parser, which are entirely determined by θ_d.

In the case of dependency length, coordinate descent on a_τ, b_τ can be applied to solve this problem (Gildea and Temperley, 2007). However, the case of prediction and parsing is more challenging: θ_d enters the objective twice, once in the distribution over linearized sentences, and once through the prediction or parsing loss.

We address this problem by simultaneously optimizing θ_d and the parameters θ_{LM}, θ_P of the language model or parser for a single objective. In the case of prediction, the objective takes the form

$$\min_{\theta_d, \theta_{LM}} \mathbb{E}_{\mathbf{w}\sim\theta_d} R_{Pred}(\mathbf{w}, \theta_{LM}). \tag{9}$$

Solving (9) is equivalent to solving the original problem (7). The objective is analogous for the parser, with θ_P instead of θ_{LM} and R_{Pars} instead of R_{Pred}.

We solve (9) by performing stochastic gradient descent over both sets of parameters θ_d, θ_{LM} simultaneously (analogously for θ_P): In each step, we sample a dependency tree t from the treebank, then sample an ordering from the current setting of θ_d to obtain a linearized sentence $\mathbf{w} \sim P_{\theta_d}(\cdot|t)$. Then we do a gradient descent step using the estimator

$$\left(\begin{matrix} \partial_{\theta_d} \left(\log P_{\theta_d}(\mathbf{w})\right) \cdot R_{Pred}(\mathbf{w}, \theta_{LM}) \\ \partial_{\theta_{LM}} R_{Pred}(\mathbf{w}, \theta_{LM}) \end{matrix} \right) \tag{10}$$

for the gradient

$$\left(\begin{matrix} \partial_{\theta_d} \\ \partial_{\theta_{LM}} \end{matrix} \right) \mathbb{E}_{\mathbf{w}'\sim\theta_d} R_{Pred}(\mathbf{w}', \theta_{LM}).$$

This unbiased estimator is the ordinary gradient estimator for θ_{LM}, and the REINFORCE estimator for θ_d (Williams, 1992).

We apply the same method to minimizing dependency length. In this case, only θ_d needs to be estimated, so only the first component of the gradient estimator (10) remains.

In addition, we estimated maximum-likelihood ordering grammars on the original ordered dependency trees with SGD.[2]

[1] For simplicity, we assume that $head(w_i) = i$ if w_i is the root of the sentence, so that the root does not contribute to dependency length.

[2] For nonprojective trees, we ignored discontinuities.

6 Implementation

Language Model We choose a standard LSTM (Hochreiter and Schmidhuber, Jürgen, 1997) language model with vocabulary size restricted to the most frequent 50,000 words in the treebanks for a given language. Given the small size of the corpora, this limit is only attained for few languages. In each time step, the input is embeddings for the word, for language-specific POS tags, and for universal POS tags. The model predicts both the next word and its POS tags in each step.

Parser We choose the biaffine attention parser by Dozat and Manning (2016). This parser has an extremely generic architecture introduced by a number of authors (Kiperwasser and Goldberg, 2016; Zhang et al., 2016), and was used in the winning entry of the ConLL 2017 shared task on multilingual dependency parsing (Dozat et al., 2017), indicating that it is well-suited to representing the syntax of typologically diverse languages. To reduce overfitting on small corpora, we choose a delexicalized setup, parsing only from POS tags.[3]

Optimization Details We employ two common variance reduction methods to improve the estimator (10), while keeping it unbiased: For dependency length and predictability, note that the loss incurred for a specific word only depends on ordering decisions made up to that word (and its head, in the case of dependency length). We represent the process of linearizing a tree as a dynamic stochastic computation graph, and use these independence properties to apply the method described in Schulman et al. (2015) to obtain a version of (10) with lower variance. Second, we use a word-dependent moving average of recent per-word losses as control variate (Williams, 1992). For stability, we represent $a_\tau \in [0, 1]$ via its logit $\in \mathbb{R}$. To encourage exploration of the parameter space, we add an entropy penalty (Xu et al., 2015) for each Direction Parameter a_τ.

We update θ_d using SGD with momentum. For the language model θ_{LM}, we use plain SGD. Following Dozat and Manning (2016), we use Adam (Kingma and Ba, 2014) for the parser θ_P.

For each language and objective function, we apply early stopping using a Monte-Carlo estimate of the respective objective functions on the development set.

Hyperparameters For dependency length minimization, we selected hyperparameters to give stable results across languages. For language modeling and parsing, we selected hyperparameters on the respective objectives for selected languages on the provided development partitions, and averaged them to obtain two sets of hyperparameters (one for language modeling, one for parsing) that were applied across all languages.

All word and POS embeddings are randomly initialized. While pretrained embeddings from unlabeled data could improve performance of language models and parsers, they could also introduce confounds from the languages' actual word orders as found in the unlabeled data.

7 Data

We base our experiments on the Universal Dependenciesi 2.1 treebanks (Nivre et al., 2017). We use all languages for which at least one treebank with a training partition was available, a total of 50 languages. For each language where multiple treebanks with training sets were available, we pooled their training sets; similarly for development sets. Punctuation was removed.

Universal dependencies represents as dependents some words that are typically classified as heads in syntactic theory. This particularly applies to the *cc*, *case*, *cop*, and *mark* dependencies. Following prior work studying dependency length minimization (Futrell et al., 2015), we modified each treebank by inverting these dependencies, promoting the dependent to the head position. We report results on this modified version of UD.

8 Results

We first evaluate the counterfactual ordering grammars for dependency length, predictability, and parsability, comparing both with the corresponding real languages and with random baselines. The purpose is (1) to verify that our optimization method yields good results on all three objective functions, and (2) to assess to what extent optimizing on one objective function increases or decreases performance on the other ones.

8.1 Dependency Length

For each optimized ordering grammar and for each language, we created a counterfactual corpus by

[3] Preliminary experiments showed that a parser incorporating word forms overfitted long before the ordering model had converged. Parsing from POS tags reduces early overfitting.

sampling a linearization for each tree, and computed per-sentence aggregate dependency lengths as in (1). Similar to Futrell et al. (2015), we considered random projective orderings as baselines.[4] For each model, we computed the average dependency length per word.

Results Results from four diverse languages are shown in Figure 3. In every one of the 50 languages, all of the optimized models show shorter dependencies than the random baseline languages. Furthermore, optimizing for dependency length achieves shorter average dependency length than 86% of the real languages (Chinese, Hindi, and Japanese in Figure 3). This suggests that 14% of the languages—including English in Figure 3—make use of word order freedom, or word order rules finer than allowed by our formalism, to achieve even shorter dependency lengths. Optimizing for parsing or predictability also results in dependencies shorter than the real language in 26% and 14%, respectively (including Chinese and Hindi in Figure 3). Optimizing for parsing results in dependencies shorter than predictability in 78% of the languages (Chinese in Figure 3).

8.2 Ease of Language Modeling and Parsing

We now evaluate for ease of prediction and parsing on English and Japanese, two typologically very different languages. For each language, we created language models and parsers (1) for the actual language, (2) for each optimized counterfactual grammar, and (3) for 20 random baselines with random $a_\tau, b_\tau \in [0, 1]$. We fitted two language models and parsers for each language in (1) and (2), and averaged the resulting values, controlling for variation due to random initialization of LSTMs.

We treat ordering grammars as deterministic in this experiment, replacing all sampling operations by *max* operations when linearizing trees. This controls for the impact that an ordering grammar's degree of nondeterminism has on cross-entropy and parser accuracy.

The cost of creating language models and parsers for baseline languages prevented conducting this experiment on the full set of languages; we chose Japanese and English as a typologically very diverse sample: English is isolating with SVO word order, while Japanese is agglutinative and has very consistent head-final order.

Results We evaluate predictability using cross-entropies, and parsability with unlabeled attachment scores (UAS). Results are shown in Figure 4. In both languages, the random ordering grammars (blue) are hardest for language modeling and parsing. Languages optimized for the respective measures showed far better performance than any of the random grammars, confirming that our method produces good results that cannot be achieved easily by random sampling of grammars. Languages optimized for dependency length (green) also achieved very strong performance in all four cases.[5] In contrast, optimizing on one of prediction (orange) or parsing (violet) does not yield such good results on the other measure. One possible interpretation is that surprisal can be lowered when different syntactic trees are mapped to the same surface string, while this makes parsing harder.

While the real languages (red) are slightly surpassed by optimized languages in parsability, they achieve even better predictability than any of the other languages. This again suggests that real languages make use of word order freedom, or more specialized ordering rules, to optimize predictability even further than allowed by our ordering model.

8.3 Greenbergian Word Order Correlations

We now examine to what extent optimized counterfactual languages satisfy the Greenberg word order correlations, allowing us to evaluate whether the correlations are predicted by our three objective functions. Since Greenberg's original work, which was based on 30 languages, the universals have been refined on the basis of many more languages. Dryer (1992) presents a comprehensive updated version of the word order correlations, drawing on 625 languages, which we take as the basis of our evaluation. In Dryer (1992), all word order correlations are relative to the position of the direct object. In Universal Dependencies, this dependency can be operationalized as VERB $\xrightarrow{obj}$ NOUN. Similarly, most of the other dependencies

[4] More precisely, we take $a_\tau = 0.5$, $b_\tau = 0$ for each τ.

[5] Figure 4 shows that minimizing dependency length achieved slightly stronger results on predictability than optimizing for predictability in Japanese, and slightly stronger results on parsability than optimizing for parsability in English. Note that, when optimizing for predictability and parsability, the target of optimization changes as the neural model is trained, making these objectives harder to optimize for than minimizing dependency length.

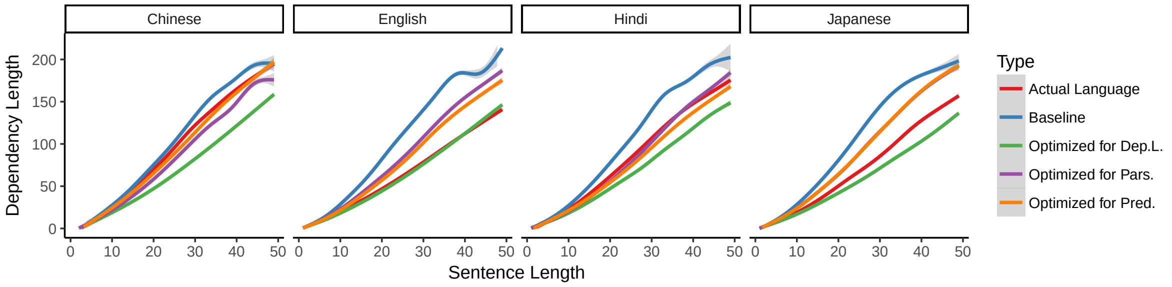

Figure 3: Aggregate dependency length as a function of sentence length in four languages. Across languages, real and optimized languages have shorter dependencies than random baseline orderings.

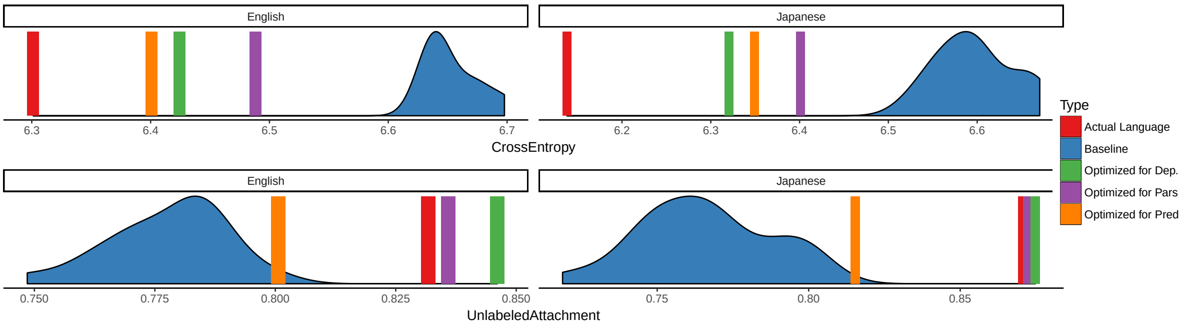

Figure 4: Predictability (top) and parsability (bottom) of real, optimized, and random languages, for English (left) and Japanese (right). We provide cross-entropy for predictability (lower is better) and unlabeled attachment scores for parsability (higher is better).

Correlate	Operationalization	Real	Dep.L.	Pred.	Pars.
adjective + std. of comp.	ADJ $\xrightarrow{obl}$ NOUN	80	**89****	55	**59****
adposition + NP	ADP $\xleftarrow{case}$ NOUN	85	**87****	42	**61*****
copula + NP	AUX $\xleftarrow{cop}$ NOUN	84	**92*****	**58*****	**60****
auxiliary + VP	AUX $\xleftarrow{aux}$ VERB	88	**76***	**73***	50
article + N’	DET $\xleftarrow{det}$ NOUN	64	**70***	49	47
noun + genitive	NOUN $\xrightarrow{nmod}$ NOUN	77	**92*****	49	**60****
noun + relative clause	NOUN $\xrightarrow{acl}$ VERB	85	**94*****	42	**58***
complementizer + S	SCONJ $\xleftarrow{mark}$ VERB	85	**91*****	**53****	57
verb + PP	VERB $\xrightarrow{obl}$ NOUN	92	71	**74****	52
want + VP	VERB $\xrightarrow{xcomp}$ VERB	92	**91***	**67****	**63*****
verb + subject	VERB $\xrightarrow{nsubj}$ NOUN	28	36	**76*****	41
verb + adverb	VERB $\xrightarrow{advmod}$ ADV	32	47	**39*****	50

Significance levels: *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$

Table 1: Greenbergian Correlations. Following Dryer (1992), each correlation is stated in terms of a pair of a ‘verb patterner’ and an ‘object patterner’ (the ‘Correlate’ column), whose relative order correlates with that of verbs and objects. For clarity, we organize the correlations by the category of the ‘verb patterner’. For each correlation, we give our operationalization in terms of UD. A rightwards arrow indicates that the dependency is predicted to have the same direction as the VERB $\xrightarrow{obj}$ NOUN dependency; a leftwards arrow indicates an inverse correlation. For each correlation we report how many of the languages in our sample satisfied it (‘Real’). We then report, for each correlation and each objective function, how many of the optimized grammars satisfy the correlation, with the significance level in a logistic mixed-effects analysis across language families.

in Dryer's formulation can be formalized: For instance, the dependency between nouns and relative clauses can be formalized as NOUN $\xrightarrow{acl}$ VERB. Testing whether an ordering model satisfies the correlation between relative clauses and objects reduces to checking whether these two triples have direction parameters a_τ that are either both > 0.5 (both kinds of dependents precede their heads) or both < 0.5 (they both follow their heads). In some cases, such as the dependency between adpositions and nouns, the formalization will differ between Universal Dependencies and the conversion that has function word heads – for instance the dependency NOUN $\xrightarrow{case}$ ADP will be represented as ADP $\xrightarrow{case}$ NOUN in our conversion of the treebanks.

There were three correlations for which an approximate operationalization was not feasible on the basis of UD: the dependencies between question particles and verbs, and those between nouns and plural words, both of which require information more fine-grained than encoded in UD. Two of Dryer's correlations, namely those for the dependencies between complementizers and adverbial subordinators and their complement clauses, had to be collapsed into a single correlation in UD. From Dryer's 15 correlations, we obtained 12 formalized correlations, roughly covering nine of Greenberg's original universals.

Validating Model Parameters We first validated that the Direction Parameters a_τ reflect typological judgments by comparing the sign of the coefficient for the verb-object dependency in the per-language maximum-likelihood fits to the actually observed orderings with the judgments in the World Atlas of Language Structures (WALS) (Dryer, 2013). The prediction is that a_τ is > 0.5 for languages with OV basic word order, and < 0.5 for those with VO order. Among the languages for which we had treebanks, WALS had no entry for 4 languages, and 'no dominant order' for 3 languages. For the remaining 43 languages, there was perfect agreement between a_τ and the WALS entry.

Testing Correlations In order to test whether an objective function predicts a correlation, we selected all ordering grammars created for the given function, and counted the percentage of grammars satisfying the correlation. Certain language families, such as Germanic and Romance languages, are more strongly represented in the UD treebanks than many other families. To control for variation between language families, we conducted, for each correlation, a mixed-effects logistic regression model predicting whether $a_\tau > 0.5$ holds for the correlating dependency depending on whether $a_\tau > 0.5$ for the verb-object dependency, with per-language-family random intercept and slope.[6] We are interested in the direction and significance of this effect: If the effect is significant, in the positive direction, we can conclude that a correlation is predicted across corpora from languages belonging to different language families.

For comparison, we also tested these correlations on the ordering models fitted to actual orderings from treebanks, evaluating in how many of the languages these models satisfied the correlations. This will help evaluate whether our formalizations of the universals reflect cross-linguistic patterns as reflected in the 50 languages of our sample.

Results Results are shown in Table 1. All correlations but two are confirmed by the models estimated from the real orderings. The exceptions are the subject-verb dependency and the verb-adverb dependency. We will discuss these exceptions further below.

Dependency Length correctly accounts for nine universals. Predictability and parsability each correctly predict six universals.

While predictions made by the three functions do not conflict (for the subset that are statistically significant), they are clearly complementary. For instance, predictability is successful at predicting correlations involving auxiliaries and verbs, while dependency length and parsability are more successful for the other groups of correlations.

The position of adverbs is predicted in the opposite direction from the typological judgment, but turns out to be in agreement with the data from the real orderings. We believe this is because the *advmod* dependency does not distinguish between manner adverbs—the subject of the typological judgment—and various other types of modifiers such as sentence-level adverbs.

By the criteria of Dryer (1992), the position of subject and object relative to the verb are correlated; however, this is a weak correlation. The subject and object on the same side of the verb in ba-

[6] We coded language families according to universaldependencies.org.

sic word orders in only about 55% of the world's languages. This is reflected in Table 1, where only 28% of the languages in the sample satisfy the correlation. [7]

We found that predictability prefers the subject and the object to be ordered on the same side of the verb, while dependency length and parsability numerically prefer the opposite. This is understandable: Predictability increases if nouns are expected on the same side of the verb, while this makes it harder to distinguish between subjects and objects, decreasing parsability.

Our results support two theoretical accounts of the variation between SVO and SOV order. Ferrer-i Cancho (2017) argues that SOV order is favored by predictability maximization while SVO order is favored by dependency length minimization, precisely as we find. In contrast, Gibson et al. (2013) argue that the advantage of SVO order is in parsability, which is also in line with our findings. Our results represent the first empirical evidence that predictability favors SOV and that parseability favors SVO.

8.4 Discussion

In Section 8.1 and 8.2, we found that, for all three objective functions, the optimization method produces results that far exceed the random baselines, confirming the effectiveness of our method. We have also found that natural languages yield strong results on all three measures, supporting the idea that our formalized objective functions reflect real pressures affecting natural languages.

One surprising finding was that optimizing dependency length produces strong results on both predictability and parsability. This result is in contrast to Gildea and Jaeger (2015), who found a tradeoff between optimizing dependency length and trigram surprisal, with optimization on one hurting performance on the other. We believe the discrepancy arises because of our more powerful language modeling and optimization techniques.

In Section 8.3, we found that none of the three objective functions alone explains all of the correlation universals, but—with the exception of the position of adverbial modifiers—all universals are predicted by at least one objective function. Many of the word order correlations have previously been subject to explanations in terms of dependency length minimization; in addition to confirming this computationally on the basis of crosslinguistic data, we show that many of these universals can also be explained in terms of predictability and parsability. It is remarkable that predictability predicts six of the correlations even though we implement it using generic sequence prediction models that do not make explicit reference to syntactic structure.

We have seen that optimizing for short dependencies yields good parsability and predictability, and also that optimizing for either of these shortens dependencies. In the case of parsability, this agrees with previous results showing that short dependencies ease parsing with different non-neural architectures (Gulordava and Merlo, 2016). Why does language modeling favor short dependencies? A possible explanation is that sequential language models will find it easier to make predictions on the basis of the more recent past. Indeed, Futrell and Levy (2017) argued that ease of incremental prediction under such memory limitations explains dependency length minimization. In our experiments, the typological facts explained by dependency length minimization were not all replicated by optimizing predictability. However, it is possible that this discrepancy is due to the size of the corpora: It may be possible that better language models would be needed to derive all the predictions of dependency length minimization from predictability alone.

9 Conclusion

We have tested explanations of crosslinguistic word order regularities in terms of efficiency of human language processing, taking the Greenbergian word order correlations as a prominent representative of word order regularities. Complementing prior theoretical work, we used corpora from 50 languages to create counterfactual languages optimized for three measures of processing efficiency. We found that most of the Greenbergian word order correlations as documented by Dryer (1992) can be derived from functional pressure to shorten dependencies, increase predictability, or increase parsabilty. Our results provide clear evidence in favor of functional explanations of syntactic universals.

[7] The divergence between this number and the worldwide 55% can be attributed to the overrepresentation of certain language families in existing treebanks. Our mixed-effects analysis helps control for this.

References

Brian Bartek, Richard L Lewis, Shravan Vasishth, and Mason R Smith. 2011. In search of on-line locality effects in sentence comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(5):1178.

Anja Belz, Mike White, Dominic Espinosa, Eric Kow, Deirdre Hogan, and Amanda Stent. 2011. The First Surface Realisation Shared Task: Overview and Evaluation Results. In *Proceedings of the Generation Challenges Session at the 13th European Workshop on Natural Language Generation*, pages 217–226, Nancy, France. Association for Computational Linguistics.

Ramon Ferrer-i Cancho. 2017. The Placement of the Head that Maximizes Predictability. An Information Theoretic Approach. page 34.

Noam Chomsky. 1981. *Lectures on government and binding: The Pisa lectures*. 9. Walter de Gruyter.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109:193–210.

Timothy Dozat and Christopher D Manning. 2016. Deep biaffine attention for neural dependency parsing. *arXiv preprint arXiv:1611.01734*.

Timothy Dozat, Peng Qi, and Christopher D Manning. 2017. Stanford's Graph-based Neural Dependency Parser at the CoNLL 2017 Shared Task. *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 20–30.

Matthew S. Dryer. 1992. The Greenbergian Word Order Correlations. *Language*, 68(1):81–138.

Matthew S. Dryer. 2013. Order of Object and Verb. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Richard Futrell and Edward Gibson. 2015. Experiments with Generative Models for Dependency Tree Linearization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1978–1983, Lisbon, Portugal. Association for Computational Linguistics.

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

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341.

E. Gibson. 1998. Linguistic complexity: locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Edward Gibson, Steven T. Piantadosi, Kimberly Brink, Leon Bergen, Eunice Lim, and Rebecca Saxe. 2013. A Noisy-Channel Account of Crosslinguistic Word-Order Variation. *Psychological Science*, 24(7):1079–1088.

Daniel Gildea and T. Florian Jaeger. 2015. Human languages order information efficiently. *arXiv:1510.02823 [cs]*. ArXiv: 1510.02823.

Daniel Gildea and David Temperley. 2007. Optimizing Grammars for Minimum Dependency Length. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 184–191, Prague, Czech Republic. Association for Computational Linguistics.

Daniel Gildea and David Temperley. 2010. Do Grammars Minimize Dependency Length? *Cognitive Science*, 34(2):286–310.

Joseph H Greenberg. 1963. Some universals of grammar with particular reference to the order of meaningful elements. *Universals of language*, 2:73–113.

Daniel Grodner and Edward Gibson. 2005. Consequences of the serial nature of linguistic input for sentenial complexity. *Cognitive science*, 29(2):261–290.

Kristina Gulordava and Paola Merlo. 2016. Multilingual dependency parsing evaluation: a large-scale analysis of word order properties using artificial data. *Transactions of the Association for Computational Linguistics*, 4:343–356.

John Hale. 2001. A Probabilistic Earley Parser as a Psycholinguistic Model. In *Proceedings of NAACL*, volume 2, pages 159–166.

John A. Hawkins. 1990. A Parsing Theory of Word Order Universals. *Linguistic Inquiry*, 21(2):223–261.

John A Hawkins. 1994. *A performance theory of order and constituency*, volume 73. Cambridge University Press.

John A. Hawkins. 2003. Efficiency and complexity in grammars: Three general principles. *The nature of explanation in linguistic theory*, pages 121–152.

Sepp Hochreiter and Schmidhuber, Jürgen. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the Limits of Language Modeling. *arXiv:1602.02410 [cs]*. ArXiv: 1602.02410.

Diederik Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*. ArXiv: 1412.6980.

Eliyahu Kiperwasser and Yoav Goldberg. 2016. Simple and Accurate Dependency Parsing Using Bidirectional LSTM Feature Representations. *arXiv:1603.04351 [cs]*. ArXiv: 1603.04351.

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

Joakim Nivre, Željko Agić, Lars Ahrenberg, Lene Antonsen, Maria Jesus Aranzabe, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, John Bauer, Kepa Bengoetxea, Riyaz Ahmad Bhat, Eckhard Bick, Victoria Bobicev, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Aljoscha Burchardt, Marie Candito, Gauthier Caron, Gülşen Cebiroğlu Eryiğit, Giuseppe G. A. Celano, Savas Cetin, Fabricio Chalub, Jinho Choi, Silvie Cinková, Çağrı Çöltekin, Miriam Connor, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Arantza Diaz de Ilarraza, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Marhaba Eli, Ali Elkahky, Tomaž Erjavec, Richárd Farkas, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta Gonzáles Saavedra, Matias Grioni, Normunds Gr¯ uzītis, Bruno Guillaume, Nizar Habash, Jan Hajič, Jan Hajič jr., Linh Hà My, Kim Harris, Dag Haug, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Radu Ion, Elena Irimia, Tomáš Jelínek, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşıkara, Hiroshi Kanayama, Jenna Kanerva, Tolga Kayadelen, Václava Kettnerová, Jesse Kirchner, Natalia Kotsyba, Simon Krek, Veronika Laippala, Lorenzo Lambertino, Tatiana Lando, John Lee, Phuong Le Hong, Alessandro Lenci, Saran Lertpradit, Herman Leung, Cheuk Ying Li, Josie Li, Keying Li, Nikola Ljubešić, Olga Loginova, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Gustavo Mendonça, Niko Miekka, Anna Missilä, Cătălin Mititelu, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Shinsuke Mori, Bohdan Moskalevskyi, Kadri Muischnek, Kaili Müürisep, Pinkey Nainwani, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Luong Nguyen Thi, Huyen Nguyen Thi Minh, Vitaly Nikolaev, Hanna Nurmi, Stina Ojala, Petya Osenova, Robert Östling, Lilja Øvrelid, Elena Pascual, Marco Passarotti, Cenel-Augusto Perez, Guy Perrier, Slav Petrov, Jussi Piitulainen, Emily Pitler, Barbara Plank, Martin Popel, Lauma Pretkalniņa, Prokopis Prokopidis, Tiina Puolakainen, Sampo Pyysalo, Alexandre Rademaker, Loganathan Ramasamy, Taraka Rama, Vinit Ravishankar, Livy Real, Siva Reddy, Georg Rehm, Larissa Rinaldi, Laura Rituma, Mykhailo Romanenko, Rudolf Rosa, Davide Rovati, Benoît Sagot, Shadi Saleh, Tanja Samardžić, Manuela Sanguinetti, Baiba Saulīte, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Dmitry Sichinava, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Antonio Stella, Milan Straka, Jana Strnadová, Alane Suhr, Umut Sulubacak, Zsolt Szántó, Dima Taji, Takaaki Tanaka, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Sowmya Vajjala, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Lars Wallin, Jonathan North Washington, Mats Wirén, Taksum Wong, Zhuoran Yu, Zdeněk Žabokrtský, Amir Zeldes, Daniel Zeman, and Hanzhi Zhu. 2017. *Universal Dependencies 2.1*.

Jan Rijkhoff. 1986. Word order universals revisited: The Principle of Head Proximity.

John Schulman, Nicolas Heess, Theophane Weber, and Pieter Abbeel. 2015. Gradient estimation using stochastic computation graphs. In *Advances in Neural Information Processing Systems*, pages 3528–3536.

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

David Temperley and Daniel Gildea. 2018. Minimizing Syntactic Dependency Lengths: Typological/Cognitive Universal? *Annu. Rev. Linguist*, 4:1–15.

Dingquan Wang and Jason Eisner. 2017. The Galactic Dependencies Treebanks: Getting More Data by Synthesizing New Languages. *arXiv:1710.03838 [cs]*. ArXiv: 1710.03838.

Michael White and Rajakrishnan Rajkumar. 2012. Minimal dependency length in realization ranking. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 244–255. Association for Computational Linguistics.

Ronald J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *International Conference for Machine Learning*. ArXiv: 1502.03044.

Xingxing Zhang, Jianpeng Cheng, and Mirella Lapata. 2016. Dependency parsing as head selection. In *EACL*.