Universals of word order result from optimization of grammars for efficient communication

immediate

This manuscript was compiled on January 22, 2019

The universal properties of human languages have been the subject of intense study across disciplines. We report novel computational and corpus-based evidence for the hypothesis that a prominent subset of these universal properties—those related to word order—result from a process of optimization for efficient communication among humans. We develop a probabilistic, differentiable model of word order grammars: the means by which different languages convert underlying hierarchical structures into strings of words. We show how the parameters of a word order grammar can be optimized for robustness of information transfer, as quantified using mutual information with latent tree structures, and efficiency of processing, as quantified using incremental predictability. Applying these grammars to tree structures found in dependency corpora, we show that optimizing the grammar parameters for efficiency and robustness results in word order patterns that reproduce a large subset of the major word order correlations reported in the linguistic typological literature, and reproduce the predictions of previous heuristic theories such as dependency length minimization.

language universals | language processing | computational linguistics

For decades, researchers in fields ranging from philology to cognitive science to statistical physics have been involved in documenting and trying to explain the universal syntactic and statistical properties of human language (1–4). An explanation for the universal properties of language would enable a deeper scientific understanding of what human language is and how to model it, with applications in psychology and natural language processing (5–7). In this work we examine this question from a computational perspective, demonstrating a fully formalized and implemented framework in which certain syntactic universals can be explained through the statistical optimization of grammars for information-theoretic efficiency.

We aim to explain the linguistic universals related to word order first documented by Greenberg (3). These universals are called **implicational** universals, and take the form of correlations between certain word order patterns and others across languages. For example, whether a language puts verbs before or after objects strongly determines whether it places adjectives before or after nouns, and whether it had prepositions or postpositions (8).

A prominent line of research has argued that universals arise for **functional** reasons: that is, because they make human communication and language processing maximally efficient, and regularities across languages hold because these efficiency constraints are rooted in general principles of communication and cognition (e.g., (2, 9–17)). Under this view, the various human languages represent multiple solutions to the problem of efficient information transfer given human cognitive constraints.

Researchers working in the functional paradigm have proposed a range of criteria which languages should meet in order to enable efficient communication and processing. These criteria are based on theories of online processing difficulty (see (17) for a review) and on information theoretic notions of efficiency and robustness in communication (18–20), and they have been formalized to varying degrees.

The contribution of this work is (i) to provide a principled formalization of the notion of communicative efficiency in word order, based on related work in information theory (21, 22), bounded rationality (23), and computational neuroscience (24); (ii) to introduce a computational framework in which we can produce simulated grammars directly optimized for our notion of efficiency, using state-of-the-art machine learning techniques (25–27); and (iii) to demonstrate that in doing so, we reproduce the major word order patterns reported in (3) and related work (8), as well as reproducing the predictions of previous heuristic theories of word order optimization (12, 28–31).

We formalize the notion of communicative efficiency using two terms: one having to do with minimizing the difficulty of linguistic communication, and one having to do with maximizing the information transferred in such communication. We quantify the difficulty of linguistic communication in terms of the word-by-word **predictability** of strings, which has been found to be a highly accurate and general predictor of human online comprehension difficulty (32–34), and which can be justified in terms of predictive coding theories of information processing in the brain (24) as well as the general algorithmic complexity of language generation and comprehension (35). We quantify the information transferred in language as **parseability**: the quantity of information contained in linguistic utterances about latent hierarchical tree structures. We also compare the results of optimizing for communicative efficiency to the results of optimizing for dependency length minimization, a previously proposed approximate metric of

Significance Statement

What explains the universal properties of human languages? We present evidence that a major subset of these properties can be explained by viewing languages as codes for efficient communication among agents with highly generic cognitive constraints. In doing so, we provide the first full formalization and computational implementation of ideas which have been stated informally in the functional linguistics literature for decades. The success of this approach suggests a new way to conceptualize human language in quantitative and computational work, as an information-theoretic code dynamically shaped by communicative and cognitive pressures. Our results argue against the idea that the distinctive properties of human language result from essentially arbitrary genetic constraints.

The authors declare no conflict of interest.

[2]To whom correspondence should be addressed. E-mail: mhahn2@stanford.edu

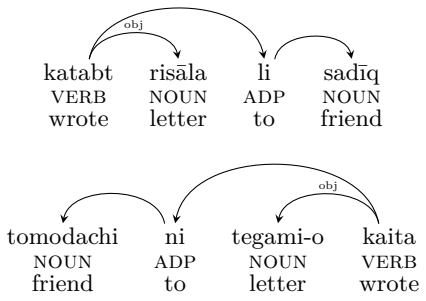

Fig. 1. A sentence in Arabic (top) and Japanese (bottom), translating to 'I wrote a letter to a friend.' Note the reversal of word order: Arabic has verb-object order and prepositions, while Japanese has object-verb order and postpositions.

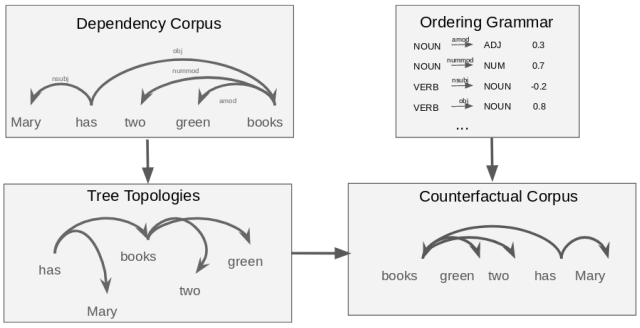

Fig. 2. Our method for constructing counterfactual versions of languages: We extract unordered dependency tree topologies from dependency corpora, and apply word order grammars to obtain counterfactually reordered corpora.

the amount of memory involved in language production and comprehension (36–40).

To evaluate whether communicative efficiency can explain word order universals, we apply the following method (41). First, we develop word order grammars: simple, interpretable models of word order, implemented as functions that take in an unordered syntactic dependency tree and output a distribution over possible word orders for that tree. Second, we take dependency treebanks of natural languages and compute word order grammars which are optimized for an objective functions representing communicative efficiency. Third, we evaluate whether the optimized word order grammars reproduce the desired word order universals when applied to attested dependency trees from the treebanks across many languages. Our favorable results suggest that it may be possible to generally model human languages as efficient codes for communication among agents with generic human-like cognitive and communicative constraints.

1. Background

A. Explananda: Word Order Correlations. Working with a database of 30 languages, Greenberg (3) proposed 45 universals of language that he determined 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'). Our goal is to explain Greenberg's reported word order correlations, as updated in (8).

B. Dependency trees. The sentences in Figure 1 are annotated with **dependency trees**: these are directed trees describing the grammatical relations among words. For example, the arcs labeled "obj" represent that the noun in question is the *direct object* if the verb, rather than e.g. the subject or an indirect object. A dependency arc is drawn from a **head** (e.g. *kaita* in Figure 1) to a **dependent** (e.g. *tegami-o*). Dependency trees can be defined in terms of many different syntactic theories (42). Although there are some differences in how different formalisms would draw trees for certain sentences, there is broad enough agreement about dependency trees that it has been possible to develop large-scale dependency-annotated corpora of text from dozens of languages (43).

In order to interpret a sentence correctly, a listener must be able to reconstruct the information expressed in the dependency tree and its labels. This is because, according to the principle of compositionality (44), the meaning of a sentence is a function of the meanings of the parts and how they are combined; the dependency tree (including the labels on the arcs) specifies how the meanings of words are combined.

2. Optimizing word order grammars

Here we describe our model of word order and how we derive optimal grammars within that model in terms of communicative efficiency.

Let t be an unordered dependency tree, i.e. a structure as in Figure 1 but without any specified ordering of the words. We think of an unordered dependency tree t as representing the compositional meaning to be conveyed by a sentence. We define a **word order grammar** as a function $L_\theta(t)$, whose behavior is specified by parameters θ, which takes an unordered dependency tree t as input and produces as output a probability distribution over ordered sequences of words $\mathbf{w}$ linearizing the tree, or equivalently, a probability distribution over orders on the nodes of the tree. This process is visualized in Figure 2.

Next we define how a word order grammar linearizes trees in terms of its parameters θ, and then we describe how θ can be set to optimize objective functions related to communicative efficiency.

A. Specification of word order grammars. All natural languages have some degree of word order regularity. 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.

Our goal is to specify word order grammars that transduce unordered dependency trees into strings of words. We have three additional goals: first, we would like our grammars to be able to operate over dependency trees as specified in large human-annotated treebank corpora, such as those developed by the Universal Dependencies project (43). Second, we would like our grammars to be simple and easily interpretable, so we can determine whether they exhibit word order universals.

Third, we would like our grammars to be differentiable, so that it is easy to optimize their parameters for efficiency.

Our model linearizes dependency trees such as the ones in Figure 1. We assume nodes are labeled with part-of-speech (POS) tags, and arcs are labeled with syntactic relations. The parameters of a word order grammar are as follows. For each dependency label type τ, we have (1) a **Direction Parameter** $a_\tau \in [0, 1]$, and (2) a **Distance Parameter** $b_\tau \in \mathbb{R}$. Each dependent is ordered on the left of its head with probability a_τ and to the right with probability $1 - a_\tau$. Then for each set of co-dependents $\{s_1, \ldots, s_n\}$ placed on one side of a head, their order outward from the head is determined by iteratively sampling from the distribution $\text{softmax}(b_{\tau_1}, \ldots, b_{\tau_n})$ ((45), p. 184) without replacement.

This simple parameterization lets us directly test whether a given grammar follows word order universals. For example, we can test whether a Greenbergian word order correlation holds in a grammar by checking whether $a_\tau > \frac{1}{2}$ for different syntactic relations τ.

B. Objective functions. Let T be a distribution over unordered dependency trees and let L_θ be a word order grammar. Our goal is to find parameters θ which maximize an objective function J in expectation over utterances:

$$J_T(\theta) = \mathbb{E}_{t\sim T}\, \mathbb{E}_{\mathbf{w}\sim L_\theta(t)}[R(t; \mathbf{w})], \qquad [1]$$

where the function R specifies a score for an individual utterance $\mathbf{w}$ expressing a tree t. In practice, T will be the empirical distribution over unordered trees observed in a dependency treebank. Below we discuss possible values of R, which may implement predictability, parseability, or a weighted combination of the two, representing overall efficiency.

Predictability Predictability represents the incremental difficulty of online language processing. We formalize predictability as the **surprisal** (negative log probability) of each word given its preceding context (32–34). Optimization of predictability has been proposed as an explanation for certain word order patterns by (46).

We formalize predictability as follows. Given a language model P_ϕ with parameters ϕ, the log probability of a sentence $\mathbf{w}$ is:

$$R^\phi_{Pred}(t; \mathbf{w}) = \sum_{i=1}^{\#\mathbf{w}} \log P_\phi(w_{i+1}|w_{1\ldots i}),$$

where the language model parameters ϕ are chosen predict the language induced by tree distribution T and word order grammar $Ł_\theta$ as accurately as possible (maximizing predictability):

$$\phi(\theta) = \arg\max_\phi \mathbb{E}_{t\sim T}\, \mathbb{E}_{\mathbf{w}\sim L_\theta(t)}[R^\phi_{Pred}(\mathbf{w}, \phi)].$$

We implement the language model P_ϕ using an LSTM neural network architecture (25), currently the state of the art for this task (26). Note that the language model parameters ϕ and the word order grammar parameters θ are recursively dependent on each other. See the SI for details on how θ and ϕ are jointly optimized, as well as on the neural network architecture.

Parseability Parseability is the extent to which the syntactic and semantic relationships among words can be recovered from the string of words. We specify parseability as the mutual information between strings and trees, or equivalently the quantity of information provided by strings about trees. Formally, we say the parseability score of an utterance $\mathbf{w}$ is the pointwise mutual information of the $\mathbf{w}$ and the tree t it was generated from, calculated using a parser with parameters ψ:*

$$R^\psi_{Pars}(t; \mathbf{w}) = \log \frac{P_\psi(t|\mathbf{w})}{P(t)} = \sum_{i=1}^{\#\mathbf{w}} \log P_\psi \begin{pmatrix} head(w_i) \\ label(w_i) \end{pmatrix} |\mathbf{w} - \log P(t),$$

where the parser parameters ψ are chosen to maximize parsing accuracy:

$$\psi(\theta) = \arg\max_\phi \mathbb{E}_{t\sim T}\, \mathbb{E}_{\mathbf{w}\sim L_\theta(t)}[R_{Pars}(\mathbf{w}, \psi)].$$

We implement P_ψ using a highly generic graph-based parsing architecture (27, 47) which recovers graph structures from word strings by solving a minimal spanning tree problem. Due to data sparsity, we train parsers to predict dependency trees from POS tags alone, not from full wordforms, therefore we do not capture the extent to which morphological wordforms are informative about dependency tree structure.

See the SI (appendix section 6-7) for further details on the joint optimization of θ and ψ and on the neural parser architecture, and see SI (section 8) for robustness to different parsing models.

Efficiency The two scoring functions—predictability and parseability—can be combined into an overall equation of the form:

$$R_{Efficiency} = R_{Pars} + \lambda R_{Pred}, \qquad [2]$$

with an interpolation weight $\lambda \in [0, 1)$. In all experiments in this paper we use $\lambda = .9$ (see SI appendix section 5 for mathematical justification). This combined equation represents efficiency: bits of information transferred about dependency trees minus the cost of the transmission. In expectation over utterances (substituting Eq. 2 into Eq. 1), the overall objective function to be maximized by a word order grammar can be written in information-theoretic terms as:

$$J_T(\theta) = I[L_\theta(T); T] - \lambda H[L_\theta(T)], \qquad [3]$$

where $I[L_\theta(T); T]$ is the **mutual information** between strings and trees, and $H[L_\theta(T)]$ is the **entropy** of strings (21). Maximizing Eq. 3 is equivalent to maximizing mutual information between strings and trees subject to a constraint on the maximum entropy of strings.

The two terms of Eq. 3 implement two opposing pressures. Parseability (the mutual information term) pushes to make strings more heterogeneous, so that they can indicate different tree structures unambiguously. Predictability (the entropy term) pushes to make strings more homogeneous, to reduce the complexity of using the language. The idea that natural language arises from a tension between these two forces goes back over a century (9). Eq. 3 is a special case of the objective function proposed in (48–50) as a general objective for communicative systems, taking unordered dependency trees T as the underlying meanings to be conveyed. Eq. 3 can also be seen as a simplified form of the Information Bottleneck (22),

*Note that $P(t)$ is a constant term which will be ignored in the optimization process.

a general objective function for lossy compression which has recently been applied to explain linguistic phenomena such as color naming systems (51) (see SI section 4 for the precise relationship).

Dependency length In previous studies of word order optimization, it has been common to propose that word orders are chosen to minimize **dependency length**, the linear distance between words linked in head–dependent relationships (31, 39, 40). Dependency length has been proposed as a metric of processing difficulty on the reasoning that long dependencies create high memory requirements for parsing and generation (37, 38), and minimization of dependency length is known to recover a large subset of the Greenbergian universals (12–14). We demonstrate that the minimization of dependency length can be formalized in our framework, reproducing the finding that optimizing grammars for dependency length recovers several word order universals (30).

To quantify dependency length, we sum the lengths of syntactic dependencies in the sentence (28–31):

$$R_{DepL}(t;\mathbf{w}) = -\sum_{i=1}^{\#\mathbf{w}} |i - head(t;w_i)|, \quad [4]$$

where $head(t;i) \in \{1,...,n\}$ is the index of the head of w_i, and the sum is negative because dependency length is seen as a cost. Dependency length quantified in this manner is a heuristic measure of complexity: the actual processing complexity induced by long dependencies is not a linear function of length and depends crucially on the types of dependencies involved (52) and the specific elements intervening between the head and dependent (37, 38, 53).

C. Discussion and Related Work. Our work builds on previous work on optimizing word order grammars for various objectives (30, 41, 54). Previous work has used deterministic word order grammars optimized by simple hill-climbing. Our work departs from previous work in three ways. First, we use word order grammars that are probabilistic and differentiable, enabling greater expressivity and optimization by stochastic gradient descent. Second, we use state-of-the-art neural language models for predictability, jointly optimizing parameters of word order grammars and language models. Third, we implement parseability and efficiency as a novel objective functions, which would not be feasible with non-differentiable word order grammars.

In the space of word order grammars, more complex models than ours have been defined (55–57). 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 phenomena that are conditioned on the larger context—e.g., differences in word order between embedded and main clauses. Also like previous models, our model only generates word orders that are **projective**, meaning that dependency lines do not cross. There are two reasons for constraint. First, there is currently no easily interpretable probabilistic model of non-projective word ordering to build on from the natural language processing literature (cf. (58)). Second, natural language dependency trees are overwhelmingly projective (59). Note that the preponderance of projectivity in natural language is itself a linguistic universal, equivalent to context-freeness in formal language theory (59, 60), and explanations have been offered for it in terms of dependency length minimization (61).

We test efficiency metrics based both on information theory (Eq. 3) and dependency length minimization (Eq. 4). We do not see these as competing theories. Rather, we see the information-theoretic objective as a broad description of efficiency in general, and dependency length as a heuristic description of a particular component of complexity. There is reason to believe that dependency length minimization might emerge as a byproduct of optimization for parseability and predictability under resource constraints (50): in particular, short dependencies are optimal for predictability given lossily compressed representations of context (62, 63). In the results below, we demonstrate that optimizing for our Eq. 3 does indeed produce dependency length minimization as a byproduct.

3. Results

A. Relative efficiency of languages. We first demonstrate that real languages are relatively efficient compared to random baselines. For each language, we generated ten baseline word order grammars for the language by choosing all word order grammar parameters randomly at uniform from $[0, 1]$. Figure 3 shows the predictability and parseability for each real language relative to its baseline grammars. In order to control for limitations due to our word order grammar formalism, we represent real languages in the figure by maximum likelihood fits of word order grammars to the real language data. For the calculation of predictability and parseability, we make all (baseline and real) word order grammars deterministic by always choosing the highest-probability linearization of each tree; by making the grammars deterministic in this way we eliminate an anticonservative bias toward low predictability in the baseline languages, which are highly nondeterministic. The majority of real languages in Figure 3 are below and to the left of their baseline equivalents, demonstrating that they are relatively high in predictability and/or parseability.

Figure 3 also shows the average position of optimized languages. Languages appear to be attracted toward these points and away from the region of the baseline languages. We also see that several languages actually end up *more* efficient than the computationally optimized languages.

These results are shown in more detail for two typologically distinct languages (English and Japanese) in Figure 4. Here we include also the predictability and parseability of the original language data and of the word order grammars optimized for dependency length. We see that real languages are even more predictable than any optimized or baseline language, while optimization for either parseability or dependency length suffices to raise parseability to the level of the real languages. This result is in line with the finding that languages with longer dependencies are harder to parse in general (64).

Finally we demonstrate the relationship between dependency length minimization and the maximization of efficiency. Figure 5 shows average dependency length per sentence length for four typologically distinct languages, showing real languages, random baselines, and languages optimized for dependency length, parseability, predictability, and efficiency. We see that optimizing for efficiency lowers dependency length

Correlates with... verb	object	Operationalization	Real	DepL	Pred	Pars	Efficiency
adposition	NP	lifted_case	86	**81*****	47	**76*****	**68*****
copula	NP	lifted_cop	94	**81*****	53	**79*****	**61****
auxiliary	VP	aux	88	**74*****	**84*****	55	**69****
noun	genitive	nmod	80	**82*****	55	**74*****	**70*****
noun	relative clause	acl	80	**85*****	48	**77****	**73*****
complementizer	S	lifted_mark	76	**85*****	**59****	**80*****	**74****
verb	PP	obl	88	**78*****	**72*****	59	**69****
want	VP	xcomp	88	**90*****	**78****	**92*****	**92*****
verb	subject	nsubj	33	**29****	51	**8*****	**13*****
verb	manner adverb	advmod	35	51	**21*****	51	**32*****

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

Table 1. Greenbergian Correlations. Following (8), each correlation is stated in terms of a pair of a 'verb patterner' and an 'object patterner', whose relative order correlate with that of verbs and objects. For each correlation, we give our operationalization in terms of UD. For each correlation we report what percentage of the languages in our sample satisfied it ('Real'). We then report, for each correlation and each objective function, how many (in %) of the optimized grammars satisfy the correlation, with the significance level in a logistic mixed-effects analysis across language families.

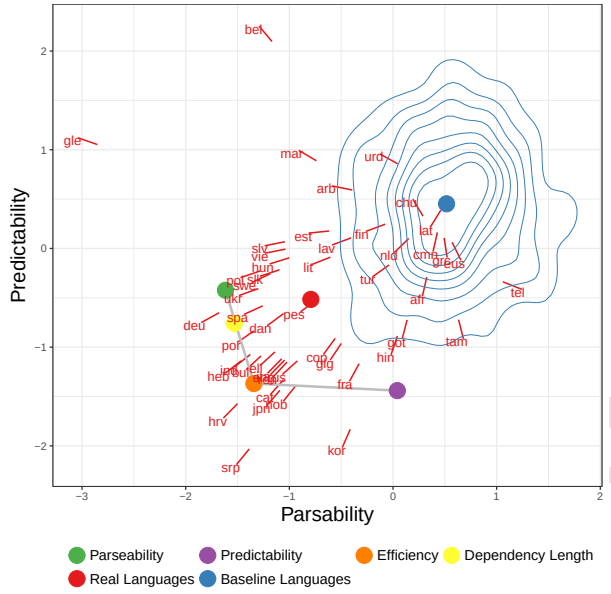

Fig. 3. Predictability and parseability of 51 UD languages (red), indicated by ISO codes, compared to ten baseline word order grammars per language (green). Predictability and parseability scores are z-scored within language. Each point for a real language has a line pointing in the direction of the center of mass of its baselines. The green contour shows the density of baseline languages. Unlabeled dots represent the centroid for real languages (red), baseline languages (green), and languages optimized for predictability (yellow), parseability (pink), efficiency (blue), and dependency length (red). When a language is to the bottom–left of its baselines, this indicates that it is relatively optimal for efficiency.

relative to random baselines, in keeping with the suggestion that dependency length minimization is a by-product of efficiency maximization (50). In 80% of the languages, optimizing explicitly for dependency length produces dependencies that overshoot the depenndency length of the real language; in 3/4 of the lanuages shown, the real language is best matched by efficiency optimization.

B. Greenbergian Word Order Correlations. We now examine to what extent we can recover Greenberg's word order correlations in optimized grammars. Dryer (8) presents a comprehensive updated version of the word order correlations, drawing on 625 languages, which we take as the basis of our evaluation. In (8), all word order correlations are relative to the position of the direct object wrt the main verb of a sentence. This dependency corresponds to dependencies of the form $\xrightarrow{obj}$ in UD. Similarly, most of the other dependencies in Dryer's formulation can be formalized: for instance, the dependency between nouns and relative clauses can be formalized as $\xrightarrow{acl}$. Testing whether an ordering model satisfies the correlation between relative clauses and objects reduces to checking whether these two relation types 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).

Dryer (8) presents three correlations which do not correspond to dependencies annotated in UD: the dependencies between question particles and verbs, those between nouns and plural words, and those between nouns and articles. Two pairs of Dryer's correlations, namely those for the dependencies between complementizers and adverbial subordinators and their complement clauses,and those for the dependencies between verbs and adpositional phrases, and adjectives and their standard of comparison, had to be collapsed into two correlations in UD. From Dryer's 15 correlations, we obtained 10 formalized correlations, roughly covering nine of Greenberg's original universals.

In order to test whether an objective function predicts a correlation, we selected all word order grammars created for the given function, and counted the percentage of grammars satisfying the correlation. We conducted, for each correlation, a mixed-effects logistic regression model predicting whether a_τ show the same direction for the correlating dependency and for the verb-object dependency, with random effects for languages and language families.[†] 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.

We compare the prevalence of the word order correlations in simulated languages to their prevalence in the real languages. To do evaluate their presence in real languages, we tested

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

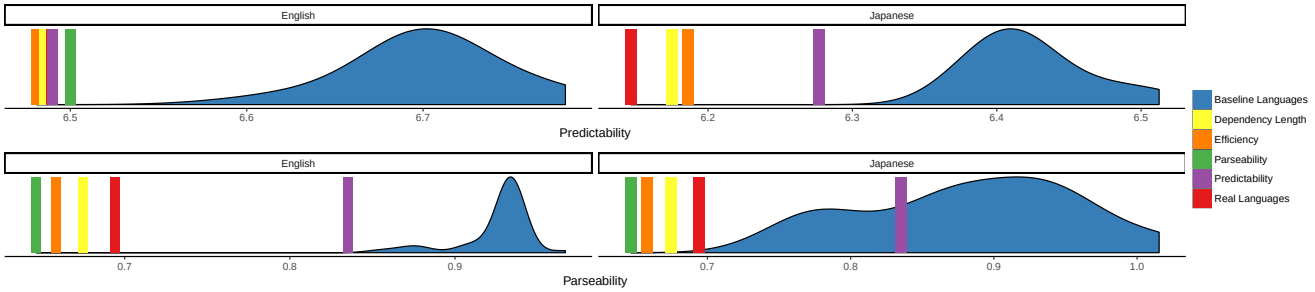

Fig. 4. Predictability (top) and parseability (bottom) of real, optimized, and random languages, for English (left) and Japanese (right). We provide the entropy $H[L_\theta(T)]$ for predictability (lower is better) and the conditional entropy $H[T|L_\theta(T)]$ for parseability (lower is better). Both measures are normalized by sentence length.

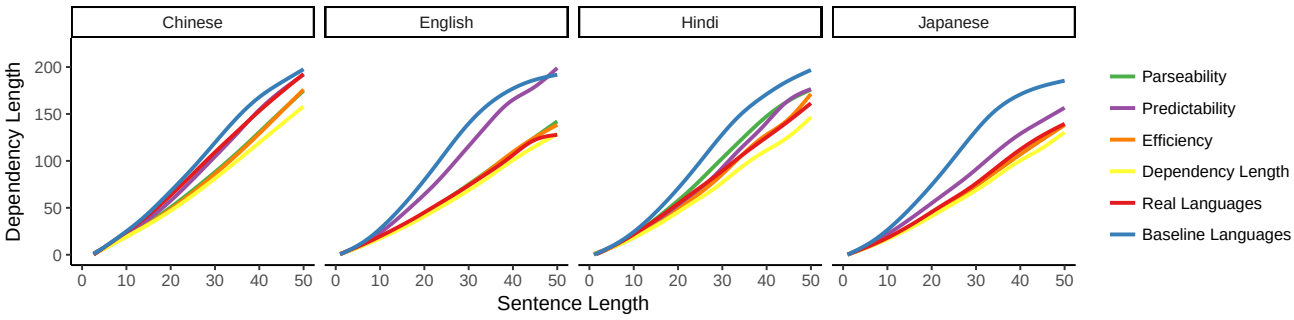

Fig. 5. Average dependency length as a function of sentence length in four languages. Across languages, real and optimized languages have shorter dependencies than random baseline orderings.

for the correlations in word order grammars fit by maximum likelihood to actual orderings from treebanks. The word order correlations detected this way match linguistic descriptions compiled in the World Atlas of Linguistic Structures (WALS, (65)) to the extent that they are documented in WALS.

Results Results are shown in Table 1. All correlations but two are confirmed in the models estimated from the real orderings. The exceptions are the subject–verb dependency and the verb–adverb dependency, which typically go in the opposite direction from the standard description. We will discuss these exceptions further below.

In keeping with previous work, we see that optimizing for dependency length correctly accounts for nine word order correlations, missing only the verb–adverb dependency. Predictability and parseability predict five and seven correlations, respectively, making largely complementary predictions. Efficiency significantly predicts all the word order correlations, each in the same direction as attested in the dependency corpora.

We now address the two word order correlations whose direction in the dependency corpora is opposite from what would be expected in the typological literature. The first is the correlation of the order of verb–subject and verb–object dependencies. Our sample of mainly European languages highly over-represents languages with the general order subject–verb–object (such as English), in which the order of the verb–subject and verb–object dependencies are anti-correlated. Surprisingly, given the sample of tree structures of these languages, it turns out that the optimal languages tend to have anti-correlated orders for subjects and objects order similar to the real languages.

The second anomalous dependency is the verb–manner adverb dependency. We believe the anti-correlation in the UD corpora arises 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. Nevertheless, the languages optimized for efficiency reproduce the anti-correlation of the orders of verb–object and verb–adverb at around the same rate as the real languages.

We further evaluate the word order predictions of efficiency, showing that efficiency is most successful in predicting correlations in the direction found in the UD corpora. We constructed a single logistic model predicting, for each of the ten dependencies, whether it is correlated or anti-correlated with the *obj* dependency in languages optimized for efficiency, with random effects for language and language family, correlated across the ten dependencies. We conducted the same analysis for predictability, parseability, and dependency length. We used this model to estimate the posterior distribution of the number of correlations that an objective function predicts to be in the same direction as found in the UD treebanks. The resulting distributions are shown in Figure 6. The estimated posterior probability that efficiency predicts less than all ten dependencies to correlate in the same direction as in the UD treebanks is 0.0242. The probability that it predicts less than nine of the correlations is $3 \cdot 10^{-4}$. For dependency length, the posterior puts much of the probability mass on predicting only nine of the correlations; predictability and parseability predict significantly less correlations.

C. Discussion. In Section A, we found that word order grammars in the majority of UD languages have better efficiency than baseline word order grammars. Furthermore, in Section B

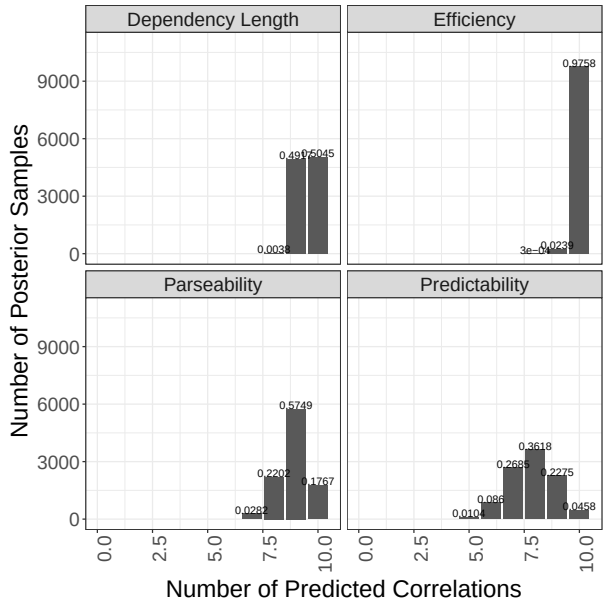

Fig. 6. Posterior of the number of correlations predicted in the direction found in the UD treebanks, computed from a mixed-effects logistic regression jointly modeling all ten dependencies. Efficiency predicts all ten correlations with high posterior probability.

we found that explicitly optimizing grammars for efficiency reproduces 10 major word order correlations reported in the typological literature, and with greater accuracy than the typological literature when it comes to the languages studied.

While the efficiency objective succeeds in predicting the presence of the most word order correlations, the dependency length objective seems to predict a greater prevalence for those correlations that it does predict. We believe this discrepancy is a result of noise in the optimization process. Optimizing for efficiency is computationally challenging, requiring stochastic gradient descent on word order grammar, language model, and parser parameters, while optimizing for dependency length is simpler and thus more optimal grammars can be reached easily (See SI section 2 for supporting data). Inasmuch as dependency length minimization is computationally easier to implement than efficiency maximization, we believe it will remain useful as an easy-to-use heuristic for predicting and explaining word order universals.

4. Conclusion

We found that a large subset of the Greenbergian word order correlations can be explained in terms of optimization of grammars for efficient communication, as defined by information theoretic criteria and implemented using state-of-the-art machine learning methods. We defined a space of word order grammars, as well as objective functions reflecting communicative efficiency, and found that the word order grammars that maximize communicative efficiency reproduce the word order universals. Beyond our present results, we provide a complete formalization and computational framework in which theories of the functional optimization of languages can be tested. Other objective functions could be hypothesized and tested within our framework; furthermore, future advances in machine learning will enable the optimization of richer and richer models of grammar.

A major question for functional explanations for linguistic universals is: *how* do languages come to be optimized? Do speakers actively seek out new communicative conventions that allow better efficiency? Or do languages change in response to biases that come into play during language acquisition (66, 67)? Our work is neutral toward such questions. To the extent that language universals arise from biases in learning or in the representational capacity of the human brain, our results suggest that those biases tilt toward communicative efficiency. Our work does provide against the idea that word order universals are best explained in terms of learning biases that are irreducibly arbitrary and genetic in nature (68).

While our work has shown that certain word order universals can be explained by efficiency in communication, we have made a number of basic assumptions about how language works in constructing our word order grammars: for example, that sentences can be syntactically analyzed into dependency trees. The question arises of whether these more basic properties themselves might be explainable in terms of efficient communication. Relatedly, there are many remaining word order universals not captured by our model, such as the trade-off of rich morphological marking and flexibility in word order (69–71). Our models cannot capture this trade-off for technical reasons: the parseability objective does not operate over wordforms, only POS tags, and so it does not take advantage of morphological cues to syntactic structure. Future work can investigate whether these and other remaining universals can be explained using more sophisticated models of how meaning representations are transduced into strings of words, based on larger databases with richer annotations.

Materials and Methods

We base our experiments on the Universal Dependencies 2.1 treebanks (43). 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 (31), we modified each treebank by inverting these dependencies, promoting the dependent to the head position. We report results on this modified version of UD.

The efficiency optimization results from Table **??** were preregistered: http://aspredicted.org/blind.php?x=8gp2bt.

See the SI for details on the neural language models and parsers used, and on the optimization procedures.

1. Behaghel O (1909) Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern. *Indogermanische Forschungen* 25:110–142.
2. Zipf GK (1949) *Human behavior and the principle of least effort.* (Addison-Wesley Press, Oxford, UK).
3. Greenberg JH (1963) Some universals of grammar with particular reference to the order of meaningful elements in *Universals of Language*, ed. Greenberg JH. (MIT Press, Cambridge, MA), pp. 73–113.
4. Lin HW, Tegmark M (2017) Critical behavior in physics and probabilistic formal languages. *Entropy* 19(7):299.
5. Hawkins JA (2007) Processing typology and why psychologists need to know about it. *New Ideas in Psychology* 25(2):87–107.
6. Bender EM (2009) Linguistically naïve!= language independent: Why nlp needs linguistic typology in *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?* (Association for Computational Linguistics), pp. 26–32.
7. Bender EM (2013) *Linguistic fundamentals for natural language processing: 100 essentials from morphology and syntax*, Synthesis Lectures on Human Language Technologies. (Morgan & Claypool Publishers) Vol. 6.

8. Dryer MS (1992) The Greenbergian word order correlations. *Language* 68(1):81–138.
9. Gabelentz Gvd (1901) *Die Sprachwissenschaft, ihre Aufgaben, Methoden, und bisherigen Ergebnisse*. (Weigel, Leipzig).
10. Hockett CF (1960) The origin of language. *Scientific American* 203(3):88–96.
11. Givón T (1991) Markedness in grammar: distributional, communicative and cognitive correlates of syntactic structure. *Stud Lang* 15:335–370.
12. Hawkins JA (1994) *A performance theory of order and constituency*. (Cambridge University Press, Cambridge).
13. Hawkins JA (2004) *Efficiency and complexity in grammars*. (Oxford University Press, Oxford).
14. Hawkins JA (2014) *Cross-linguistic variation and efficiency*. (Oxford University Press, Oxford).
15. Croft WA (2001) Functional approaches to grammar in *International Encyclopedia of the Social and Behavioral Sciences*, eds. Smelser NJ, Baltes PB. (Elsevier Sciences, Oxford), pp. 6323–6330.
16. Haspelmath M (2008) Parametric versus functional explanations of syntactic universals in *The Limits of Syntactic Variation*, ed. Biberauer T. (John Benjamins, Amsterdam), pp. 75–107.
17. Jaeger TF, Tily HJ (2011) On language 'utility': Processing complexity and communicative efficiency. *Wiley Interdisciplinary Reviews: Cognitive Science* 2(3):323–335.
18. Ferrer i Cancho R, Solé RV (2001) Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited. *Journal of Quantitative Linguistics* 8(3):165–173.
19. Piantadosi ST, Tily H, Gibson E (2011) Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences* 108(9):3526–3529.
20. Gibson E, et al. (2013) A noisy-channel account of crosslinguistic word-order variation. *Psychological Science* 24(7):1079–1088.
21. Cover TM, Thomas J (2006) *Elements of Information Theory*. (John Wiley & Sons, Hoboken, NJ).
22. Tishby N, Pereira F, Bialek W (1999) The information bottleneck method in *Invited paper to The 37th annual Allerton Conference on Communication, Control, and Computing*.
23. Genewein T, Leibfried F, Grau-Moya J, Braun DA (2015) Bounded rationality, abstraction, and hierarchical decision-making: An information-theoretic optimality principle. *Frontiers in Robotics and AI* 2:27.
24. Friston K, Kiebel S (2009) Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364:1211–1221.
25. Hochreiter S, Schmidhuber J (1997) Long short-term memory. *Neural Computation* 9(8):1735–1780.
26. Goldberg Y (2017) Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies* 10(1):1–309.
27. Dozat T, Qi P, Manning CD (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* pp. 20–30.
28. Ferrer i Cancho R (2004) Euclidean distance between syntactically linked words. *Physical Review E* 70(5):056135.
29. Liu H (2008) Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science* 9(2):159–191.
30. Gildea D, Temperley D (2010) Do grammars minimize dependency length? *Cognitive Science* 34(2):286–310.
31. Futrell R, Mahowald K, Gibson E (2015) Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences* 112(33):10336–10341.
32. Hale JT (2001) A probabilistic Earley parser as a psycholinguistic model in *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics and Language Technologies*. pp. 1–8.
33. Levy R (2008) Expectation-based syntactic comprehension. *Cognition* 106(3):1126–1177.
34. Smith NJ, Levy R (2013) The effect of word predictability on reading time is logarithmic. *Cognition* 128(3):302–319.
35. Li M, Vitányi P (2008) *An introduction to Kolmogorov complexity and its applications*. (Springer-Verlag, New York).
36. Hudson RA (1995) Measuring syntactic difficulty. Unpublished manuscript.
37. Gibson E (1998) Linguistic complexity: Locality of syntactic dependencies. *Cognition* 68(1):1–76.
38. Gibson E (2000) The dependency locality theory: A distance-based theory of linguistic complexity in *Image, Language, Brain: Papers from the First Mind Articulation Project Symposium*, eds. Marantz A, Miyashita Y, O'Neil W. pp. 95–126.
39. Liu H, Xu C, Liang J (2017) Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*.
40. Temperley D, Gildea D (2018) Minimizing syntactic dependency lengths: Typological/cognitive universal? *Annual Review of Linguistics* 4:1–15.
41. Gildea D, Temperley D (2007) Optimizing grammars for minimum dependency length in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*. (Prague, Czech Republic), pp. 184–191.
42. Corbett GG, Fraser NM, McGlashan S (1993) *Heads in Grammatical Theory*. (Cambridge University Press, Cambridge).
43. Nivre J, et al. (2017) Universal dependencies 2.0 – CoNLL 2017 shared task development and test data. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.
44. Frege G (1892) Über sinn und bedeutung. *Zeitschrift für Philosophie und philosophische Kritik* 100(1):25–50.
45. Goodfellow I, Bengio Y, Courville A (2016) *Deep Learning*. (MIT Press). http://www.deeplearningbook.org.
46. Ferrer i Cancho R (2017) The optimality of attaching unlinked labels to unlinked meanings. *Glottometrics* 36:1–16.
47. Kiperwasser E, Goldberg Y (2016) Simple and accurate dependency parsing using bidirectional lstm feature representations. *Transactions of the Association of Computational Linguistics* 4:313–327.
48. Ferrer i Cancho R, Solé R (2002) Zipf's law and random texts. *Advances in Complex Systems* 5(1):1–6.
49. Ferrer i Cancho R, Díaz-Guilera A (2007) The global minima of the communicative energy of natural communication systems. *Journal of Statistical Mechanics: Theory and Experiment* 2007(06):P06009.
50. Futrell R (2017) Ph.D. thesis (Massachusetts Institute of Technology, Cambridge, MA).
51. Zaslavsky N, Kemp C, Regier T, Tishby N (2018) Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences* 115(31):7937–7942.
52. Demberg V, Keller F (2008) Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition* 109(2):193–210.
53. Lewis RL, Vasishth S (2005) An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science* 29(3):375–419.
54. Gildea D, Jaeger TF (2015) Human languages order information efficiently. *arXiv* 1510.02823.
55. Belz A, et al. (2011) The first surface realisation shared task: Overview and evaluation results in *Proceedings of the 13th European Workshop on Natural Language Generation*. (Association for Computational Linguistics), pp. 217–226.
56. Futrell R, Gibson E (2015) Experiments with generative models for dependency tree linearization in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. (Association for Computational Linguistics, Lisbon, Portugal), pp. 1978–1983.
57. Wang D, Eisner J (2016) The Galactic Dependencies Treebanks: Getting more data by synthesizing new languages. *Transactions of the Association for Computational Linguistics* 4:491–505.
58. Bohnet B, Björkelund A, Kuhn J, Seeker W, Zarrieß S (2012) Generating non-projective word order in statistical linearization in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. (Association for Computational Linguistics), pp. 928–939.
59. Kuhlmann M (2013) Mildly non-projective dependency grammar. *Computational Linguistics* 39(2):355–387.
60. Marcus S (1965) Sur la notion de projectivité. *Mathematical Logic Quarterly* 11(2):181–192.
61. Ferrer i Cancho R (2006) Why do syntactic links not cross? *Europhysics Letters* 76(6):1228.
62. Futrell R, Levy R (2017) Noisy-context surprisal as a human sentence processing cost model in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. (Valencia, Spain), pp. 688–698.
63. Futrell R, Levy R, Gibson E (2017) Generalizing dependency distance: Comment on "dependency distance: A new perspective on syntactic patterns in natural languages" by haitao liu et al. *Physics of Life Reviews* 21:197–199.
64. Gulordava K, Merlo P (2016) Multi-lingual dependency parsing evaluation: a large-scale analysis of word order properties using artificial data. *Transactions of the Association for Computational Linguistics* 4:343–356.
65. Dryer MS, Haspelmath M (2013) *WALS Online*. (Max Planck Institute for Evolutionary Anthropology, Leipzig).
66. Fedzechkina M, Jaeger TF, Newport EL (2012) Language learners restructure their input to facilitate efficient communication. *Proceedings of the National Academy of Sciences* 109(44):17897–17902.
67. Culbertson J, Smolensky P, Legendre G (2012) Learning biases predict a word order universal. *Cognition* 122(3):306–329.
68. Chomsky N (2010) Some simple evo devo theses: How true might they be for language? in *The Evolution of Human Language*. (Cambridge University Press, Cambridge), pp. 45–62.
69. Jespersen O (1922) *Language: Its nature, development, and origin*. (Henry Holt and Co., New York).
70. McFadden T (2003) On morphological case and word-order freedom in *Proceedings of the Berkeley Linguistics Society*.
71. Futrell R, Mahowald K, Gibson E (2015) Quantifying word order freedom in dependency corpora in *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*. (Uppsala, Sweden), pp. 91–100.