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

Michael Hahn[a,1,2], **Daniel Jurafsky**[a], **and Richard Futrell**[b]

[a]Stanford University; [b]University of California, Irvine

This manuscript was compiled on February 7, 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 efficiency of information transfer, quantified as a tradeoff between incremental predictability and mutual information with latent tree structures. Applying these grammars to tree structures found in dependency corpora from 51 languages, we show that optimizing the grammar parameters for efficiency 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 framework in which certain syntactic universals can be explained through the statistical optimization of grammars for information-theoretic efficiency.

Natural languages vary a lot in the order in which they express information. Consider Figure 1, showing a sentence in Arabic (top) and Japanese (bottom), both translating to 'I wrote a letter to a friend.' Both sentences contain a verb meaning 'wrote', a noun expressing 'letter', and a phrase translating to 'to a friend'. In linguistic terminology, 'letter' is the object of the verb, whereas the phrases translating as 'to a friend' are known as adpositional phrases. While the two sentences containing words with the same meanings, the order of these words are entirely different in the two languages: In Arabic, the verb stands at the beginning, followed by both the object and the adpositional phrase. 'To' is expressed by a *preposition*, so named because it precedes the word expressing 'friend'. In Japanese, the verb stands at the end; object and adpositional phrase precede it; 'to' is expressed by a *postposition*, so named because it follows the word denoting 'friend'. It turns out that this variation reflects a deep and stable regularity: While languages ordering the objects before (Japanese) or after (Arabic) the verb are approximately equally common around the world, this is strongly correlated with the occurrence of pre- or postpositions: Languages ordering their objects the way Japanese does, have postpositions; languages ordering them as Arabic does have prepositions. This correlation has an extremely strong empirical basis: In a sample of about 1,100 languages from five contents (CITE Dryer WALS 95a), less than 7 % of languages provide clear exceptions to this generalization.

katabt	risāla	li	sadīq
VERB	NOUN	ADP	NOUN
wrote	letter	to	friend

tomodachi	ni	tegami-o	kaita
NOUN	ADP	NOUN	VERB
friend	to	letter	wrote

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.

This generalization falls in a group of language universals that were originally documented by Greenberg (3), known as **word order correlations**: These describe correlations between the relative positions of different types of expressions across languages. The example above documents that the position of the object ('letter') relative to the verb is **correlated** with the position of the adposition ('to'). Greenberg originally worked with 30 languages; the correlation universals have since been confirmed on the basis of much larger samples of languages. The authoritative study by Dryer (8) draws on

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.

MH and RF designed research. MH implemented models and experiments. MH and RF wrote the paper. MH, DJ, and RF provided comments on the paper.

The authors declare no conflict of interest.

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

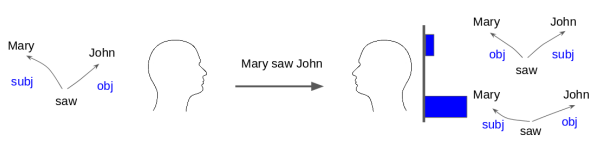

Fig. 2. Our model of communication: A speaker (left) expresses a dependency structure into a string of words forming a sentence. A listener probabilistically recovers a dependency structure. In this example, the grammar of English allows the listener to recover the correct dependency structure with very high confidence.

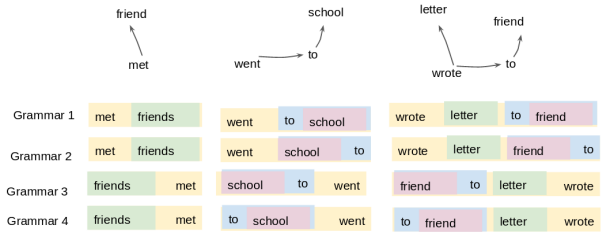

Fig. 3. Given a set of dependency structures, grammars define a set of sentences. For simplicity, we use English words. Each grammar defines a consistent ordering for the different kinds of syntactic relations, e.g., Grammars 1 and 2 order the object after the verb, Grammars 3 and 4 order the object before the verb.

625 languages and documents fifteen such correlations.

Explaining these patterns has been an important aim of linguistic research since Greenberg's seminal study.

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 that 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.

We formalize efficiency of language as a tradeoff between minimizing the

(2) argued that language optimizes a tradeoff between speaker effort (Force of Unification) and listener effort (Force of Diversification). According to Zipf, a language that minimizes speaker effort should reduce the number of different utterances to make production as easy as possible. Conversely, to minimize listener effort, the language should provide different utterances for different meanings, so that the listeenr can unambiguously identify the meaning from the utterance. Minimizing speaker and listener represent two opposing forces: A pressure to reduce the complexity of the language makes the utterances in a language more homogenous. A pressure to reduce ambiguity makes utterances more heterogeneous, so they can indicate different meanings. The idea that language results from the tension between these two pressures has a long and fruitful history. (**?**), RSA, Color naming

This idea has been shown to account for phenomena such as pragmatic inference (CITE) and color naming (CITE).

TODO somewhere start calling this Efficiency

TODO explain why surprisal. some production/speaker work. (plan reuse)

In this work, we show computationally, using corpus data from 51 languages, that efficiency optimization accounts for the Greenbergian word order correlations. That is, we show that the word orders in natural languages have evolved to optimize efficiency, and that this optimization accounts for the prevalence of the word order correlations.

1. Formalizing Efficiency in Word Order

We model the process where a speaker communicates an utterance to a listener (Figure 2) in Shannon's framework of information theory. In Shannon's model, a transmitter encodes a message into a signal. The receiver decodes the signal, attempting to reconstruct the original message. Applying this model to word order, we take the message to consist of a set of syntactic and semantic relations. Following a long tradition in formal and computational liguistics, we formalize these in the format of Dependency Grammar: This linguistic formalism draws directed arcs between syntactically related words, annotated with syntactic relations. In Figure 2, the message consists of a dependency structure with the words 'saw', 'Mary', 'John', where Mary is the subject of the event denoted by 'saw', and John is the object.

When uttering, the speaker needs to choose an ordering in which to order the words in this tree to generate a string of words. The listener receives this string of words. By the principle of compositionality (25), the meaning of a sentence is a function of the meanings of the parts and how they are combined. The dependency structure (including the labels on the arcs) specifies how the meanings of words are combined. Therefore, a listener needs to recover the information provided in the dependency structure in order to understand a sentence correctly. Consistent with Shannon's model, we assume a Bayesian listener who decodes dependency structures probabilistically (Figure 2). If communication is successful, the listener can identify the intended structure with high confidence (Figure 2).

All natural languages have some degree of word order regularities that are specified in their grammar. For instance, English places subjects before and objects after the verb, allowing the listener to unambiguously decode the sentence in Figure 2. That is, speakers and listeners have shared knowledge of a **grammar** that specifies how dependency structures are encoded into sentences. Natural languages differ in the rules they apply: Some place the object after the verb, some place it before the verb. This is illustrated in Figure 3: Grammars specify how dependency structures are encoded into strings of words. For instance, Grammar 1 – corresponding to Arabic in Figure 1 – orders objects ('friends', 'letter') after verbs and has prepositions ('to friend'). Grammar 2 orders objects after verbs but has postpositions ('friend – to'). Grammars 3 and 4 place the object before the verb, and one of them (Grammar 3) corresponds to Japanese order.

In Information Theory, the usefulness of a communication channel depends on (1) how costly encoding and transmission

Correlates with... verb	object	Real	DepL	Pred	Pars	Efficiency
adposition	NP	*to*	*a friend*			
copula	NP					
is	*a friend*					
auxiliary	VP					
has	written					
noun	genitive					
friend	of John					
noun	relative clause					
books	that you read					
complementizer	S					
that	Mary					
verb	PP					
went	to school					
want	VP					
wants	to to leave					
verb	subject					
(there) entered	a tall man					
verb	manner adverb					
ran	quickly					

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 provide an example. Given the statistical nature of the correlations, not every language satisfies every one of them. Not all correlations are satisfied by every natu

is, and (2) how precisely messages can be recovered from codes. In our model, these will depend on the grammar: Grammars can differ in the degree to which a listener can unambiguously recover the dependency structure from an utterance (TODO can we find an example for this?)

The degree to which listeners can reconstruct dependency structures from an utterance is formalized as the amount of information that utterances provide about their underlying tree structures:

$$R_{Pars} := I[L,T] = \sum_{t,l} p(t,l) \log \frac{p(t|l)}{p(t)} \quad [1]$$

where the sum runs over all possible pairs of sentences l and dependency structures t in the language. This quantity describes the degree to which dependency structures can be unambiguously recovered from sentences. It is large when they can mostly be recovered, and small if sentences have high amounts of ambiguity about the dependency structure. This quantity formalizes Zipf's principle of listener effort and his Force of Diversification: It can be maximized when trees can be decoded fully unambiguously from utterances. This formalization is a standard formalization of Zipf's principle of listener economy (CITE RSA, Regier, ...).

The counteracting Force of Unification is formalized by the cost of encoding the dependency structure and producing the utterance. TODO motivate surprisal here.

Do languages optimize their word order for efficiency? Does this optimization explain the observed language universals?

The **predictability** of the language is the negative entropy of the sentences:

$$R_{Pred} := H[L] = \sum_{l} p(l) \log p(l) \quad [2]$$

where the sum runs over all possible sentences l that belong to the language. This quantity describes how homogeneous the language is, i.e., it is larger if the distribution over sentences is concentrated on a smaller number of frequent sentences.

The opposing pressure is **parsability**, which is the mutual information beteen sentences and dependency structures:

$$R_{Pars} := I[L,T] = \sum_{t,l} p(t,l) \log \frac{p(t|l)}{p(t)} \quad [3]$$

where the sum runs over all possible pairs of sentences l and dependency structures t in the language. This quantity describes the degree to which dependency structures can be unambiguously recovered from sentences. It is large when they can mostly be recovered, and small if sentences have high amounts of ambiguity about the dependency structure.

The **efficiency** of a language is a weighted combination

$$R_{Eff} := R_{Pars} + \lambda R_{Pred} \quad [4]$$

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).

Eq. **??** is a special case of the objective function proposed in (26–28) as a general objective for communicative systems, taking unordered dependency trees T as the underlying meanings to be conveyed. Eq. **??** can also be seen as a simplified form of the Information Bottleneck (29), a general objective function for lossy compression which has recently been applied to explain linguistic phenomena such as color naming systems (30) (see SI section 4 for the precise relationship).

A. Optimizing Grammars for Efficiency. We now ask: (1) Do the grammars of natural languages evolve towards optimizing efficiency of communication? (2) Does this process of optimization account for Greenberg's order correlation universals?

Answering these questions requires a sample of dependency structures as actually used by speakers of difefrent languages.

Such samples have recently become available with the Universal Dependencies project, which has collected and created dependency annotations for several dozens of different languages. We use data from 51 languages. These corpora represent a typologically and genetically diverse group of languages.

To answer whether languages evolve to have efficient word order, we compare the efficiency of the actual grammars of these 51 languages to randomly constructed baseline grammars. To show that this process of efficiency optimization accounts for Greenberg's correlation universals, we computationally construct grammars that optimize efficiency. We then show that these grammars mostly exhibit the Greenbergian order correlation universals. We furthermore show that this process of optimization also *explains* DLM, providing a first-principles explanation of this heuristic generalization.

For every one of the 51 corpora, we computationally construct eight optimal grammars. Note that the original orders of the actual languages do not enter this objective. See SI for our method for creating optimized grammars.

2. 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 4 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 4 are below and to the left of their baseline equivalents, demonstrating that they are relatively high in predictability and/or parseability.

Figure 4 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.

B. Efficiency Explains DLM. 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 relative to random baselines, in keeping with the suggestion that dependency length minimization is a by-product of efficiency maximization (28). 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.

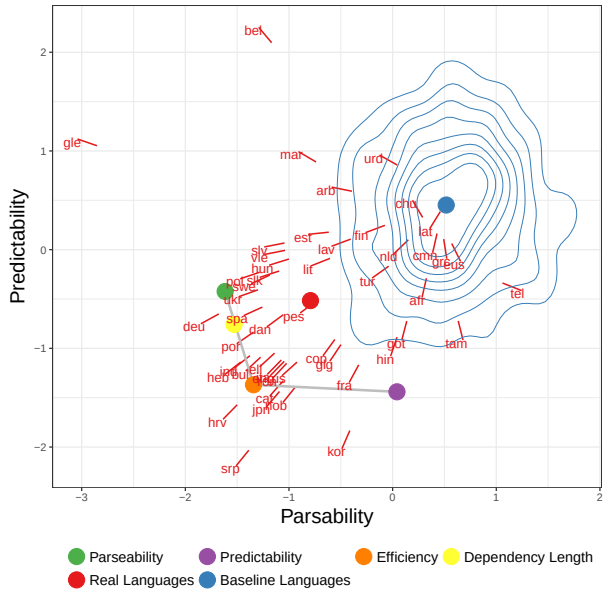

Fig. 4. Predictability and parseability of 51 UD languages (red), indicated by ISO codes, compared to ten baseline word order grammars per language (green). Predictability and parseability scores are z-scored within language. Each point for a real language has a line pointing in the direction of the 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.

C. 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. Most of them can be straightforwardly implemented in UD, allowing us to check which correlations a word order grammar satisfies.

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

*We coded language families according to universaldependencies.org.

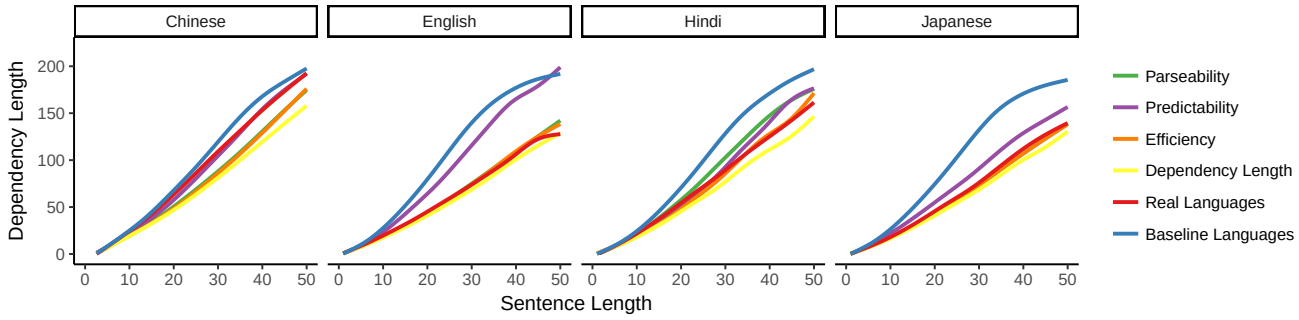

Fig. 5. Average dependency length as a function of sentence length in four languages. Across languages, real and optimized languages have shorter dependencies than random baseline orderings.

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

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

Table 2. 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.

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 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, (32)) to the extent that they are documented in WALS.

Results Results are shown in Table 2. 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.

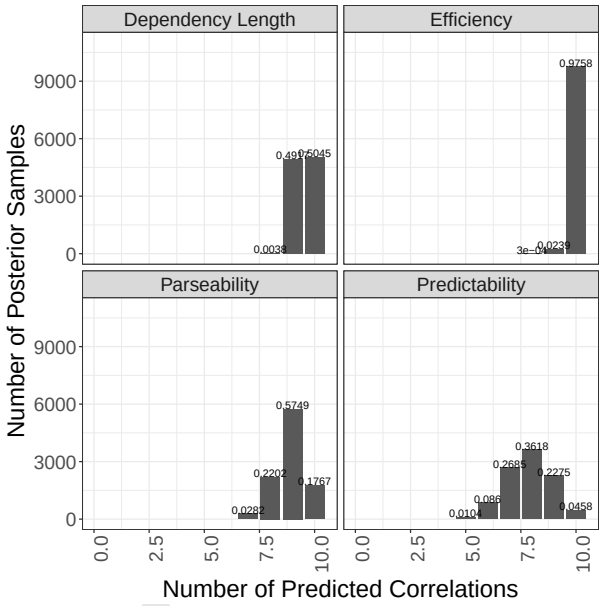

Fig. 6. Posterior of the number of correlations predicted in the direction found in the UD treebanks, computed from a mixed-effects logistic regression jointly modeling all ten dependencies. Efficiency predicts all ten correlations with high posterior probability.

3. DLM TODO put these at the appropriate place

We compare Efficiency to the predictions of Dependency Length Minimization. This theory states that natural languages order information in such a way that the distances between syntactically linked words are minimized. For instance, in Figure 1, there would be syntactic links between the adposition (to) and the noun (friend), and between the verb and both the object (letter) and the adposition (to). There is strong evidence that natural language minimizes the length of these dependencies (21–24). Theoretical work in the functional linguistics literature has proposed that this principle explains the correlations. However, the principle itself is stipulative and not a first-principles explanation. (12) theoetically argues that it increases parsability.

However, these arguments have been made on a theoretical basis. We will show that (1) Dependency Length Minimization indeeds predicts most of the correlations, and (2) Dependency Length Minimization is explained by efficiency optimization.

These computational results confirm theoretical ideas that have been stated informally by authors in the functional linguistics literature at least since the 1980s.

ACKNOWLEDGMENTS. We thank Ted Gibson, Michael C. Frank, Judith Degen, Chris Manning, and audiences at CAMP 2018 for helpful discussion.

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. Ferrer i Cancho R (2004) Euclidean distance between syntactically linked words. *Physical Review E* 70(5):056135.
22. Liu H (2008) Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science* 9(2):159–191.
23. Gildea D, Temperley D (2010) Do grammars minimize dependency length? *Cognitive Science* 34(2):286–310.
24. 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.
25. Frege G (1892) Über sinn und bedeutung. *Zeitschrift für Philosophie und philosophische Kritik* 100(1):25–50.
26. Ferrer i Cancho R, Solé R (2002) Zipf's law and random texts. *Advances in Complex Systems* 5(1):1–6.
27. 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.
28. Futrell R (2017) Ph.D. thesis (Massachusetts Institute of Technology, Cambridge, MA).
29. 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*.
30. 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.
31. 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.
32. Dryer MS, Haspelmath M (2013) *WALS Online.* (Max Planck Institute for Evolutionary Anthropology, Leipzig).