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

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

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

This manuscript was compiled on May 31, 2019

The universal properties of human languages have been the subject of intense study across the language sciences. We report novel computational and corpus 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, trading off the need to reduce complexity with the need to reduce ambiguity. We formalize these two pressures with information-theoretic and neural network models of complexity and ambiguity, and simulate grammars optimizing their word order parameters on data from 51 languages. Evolution of grammars towards efficiency results in word order patterns that predict a large subset of the major word order correlations across languages.

language universals | language processing | computational linguistics

Understanding what is universal and what varies across human languages is a central goal of linguistics. Across theoretical paradigms, linguists have hypothesized that language is shaped by efficiency in computation (1–4) and communication (5–12). But formalizing how these pressures explain specific grammatical universals has proved difficult. Here we pair new computational models that measure the communicative efficiency of grammars with a new simulation framework for finding optimal grammars, and show that the most efficient grammars also exhibit a large class of language universals.

The language universals we study are the well-known Greenberg universals of word order (13). Human languages vary 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 the object 'letter', and a phrase translating to 'to a friend'. Yet, the order of these words are entirely different in the two languages: The verb stands at the beginning in Arabic, and at the end in Japanese. Arabic expresses 'to' by a *preposition* (*preceding* the noun 'friend'); Japanese uses a *postposition* (*following* it).

Yet 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 (Figure 1, top): Languages ordering their objects the way Japanese does, have postpositions; languages ordering them as Arabic does have prepositions.

This generalization lies in a group of language universals originally documented by Greenberg (13), 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 also found that the order of verb and object is correlated with other aspects of

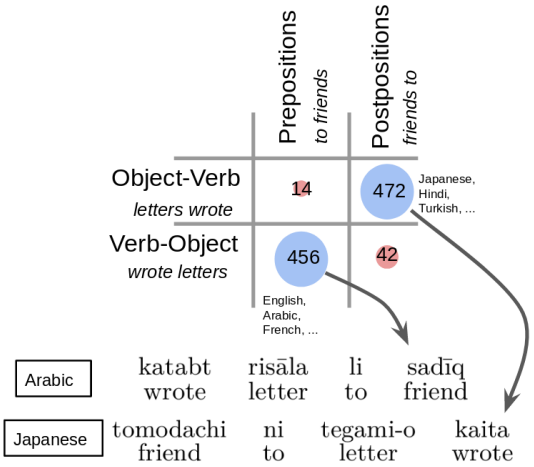


Fig. 1. One word order correlation: Languages can order the object after (Arabic) or before (Japanese) the verb, and have prepositions (Arabic) or postpositions (Japanese). For each combination, we indicate how many languages satisfy it, as documented in the World Atlas of Language Structures (14). Combinations on the diagonal are vastly more common than off-diagonal ones.

a language's word order (Table 1), such the order of verb and adpositional phrase ('wrote – to friend' in Arabic vs 'friend to – wrote' in Japanese), and that of noun and genitive ('book – of friend' in Arabic, 'friend of – book' in Japanese).

Supported by languages on all continents, these correlations are among the language universals with the strongest empirical support. Importantly, their validity is also independent from specific assumptions about theories of grammar.

Explaining these patterns has been an important aim of linguistic research since Greenberg's seminal study (4, 15–20).

Significance Statement

Human languages share many grammatical properties. We show that some of these properties can be explained by the need for languages to offer efficient communication between humans given our cognitive constraints. Grammars of languages seem to find a balance between two communicative pressures: to be simple enough to allow the speaker to easily produce sentences, but complex enough to be unambiguous to the hearer, and this balance explains well-known word order generalizations across our sample of 51 varied languages. Our results offer new quantitative and computational evidence that language structure, rather than being an outcome of arbitrary genetic constraints, is dynamically shaped by communicative and cognitive pressures.

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

The authors declare no conflict of interest.

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

	Arabic (English, ...)		**Japanese** (Turkish, ...)	
	Correlates with...		Correlates with...	
	Verb	Object	Object	Verb
	kataba *wrote*	rasaa'il *letters*	tegami-o *letter*	kaita *wrote*
①	li *to*	sadiiq *a friend*	tomodachi *friend*	ni *to*
②	kaana *was*	sadiiq *a friend*	tomodachi *friend*	datta *was*
③	sawfa *will*	yaktub *write*	kak *write*	udesho *will*
④	sadiiq *friend*	John *of John*	John no *John of*	tomodachi *friend*
⑤	kutub *books*	taqra'uha *that you read*	anata-ga yonda *that you read*	hon *book*
⑥	'an *that*	tusil *she arrives*	toochaku suru *has.arrived*	koto *that*
⑦	dhahabt *went*	'ila lmadrasa *to school*	gakkoo ni *school to*	ittekimashita *went*
⑧	'uriid *wants*	'an 'ughaadir *to leave*	yom *to leave*	itai *want*

Table 1. Greenberg's word order correlations, exemplified by Arabic (left) and Japanese (right) examples: Across the world, the orders of different constituents are strikingly correlated with that of verb and object. Selection is based on a more recent typological study by Dryer (15), restricted to those correlations that are annotated in available corpus data. See SI Section S1 for more on Greenberg correlations.

Prominent among this research is the argument that language 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., (4, 5, 8, 21–28)). Under this view, the various human languages represent multiple solutions to the problem of efficient information transfer given human cognitive constraints.

In an early and influential functional framework, Zipf (5) argued that language optimizes a tradeoff between two pressures: to reduce complexity and to reduce ambiguity. What Zipf called the Force of Unification is a pressure to reduce the complexity of the language to make production and processing as easy as possible. His Force of Diversification favors languages that provide different utterances for different meanings, so that the listener can unambiguously identify the meaning from the utterance. These two forces act in opposing directions: Producing and processing simple utterances incurs little cost, but more complex and diverse utterances are required to provide enough information. The idea that many properties of language arise from the tension between these two pressures has a long and fruitful history in linguistics (21, 24, 29–32).

Recent work has drawn on information theory to computationally test this "dual pressures" idea in various domains of language, showing that it predicts both basic statistical properties of languages (33) and language development (9), and sophisticated aspects of language, such as pragmatic inference (34), and the distribution of color words (35) and kinship categories (36) across many languages. While it has been suggested that the dual pressure should also apply to grammar (24), testing these accounts is more difficult, as this requires large amounts of data representative of language use across languages, computational methods for estimating the efficiency of entire languages, and a simulation methodology for comparing different possible grammars.

In this work, we address these challenges by combining large-scale text data from 51 languages with machine learning techniques to estimate both aspects of the communicative efficiency of grammar: complexity and ambiguity. We use machine learning models based on neural networks to model the evolution of grammars towards efficiency. We apply this approach to the problem of explaining Greenberg's word order correlation universals.

In Study 1 we compare the word order of actual grammars of 51 languages with alternative "counterfactual" grammars parameterized by different word orders. We use our model to measure the communicative efficiency of each possible grammar, showing that the grammars of real languages are more efficient than alternative grammars. The fact that real grammars lie at the Pareto frontier of the efficiency space of possible grammars suggests that the word order of languages has evolved to optimize communicative efficiency.

In Study 2 we test whether efficiency optimization accounts for the Greenbergian word order correlations. For each of the 51 languages, we create hypothetical grammars optimized for efficiency. We then test statistically whether these optimized grammars exhibit the Greenberg correlations, using a Bayesian mixed-effects logistic regression to control for language and language family. Efficiency optimization indeed predicts all eight Greenberg collections. Our results show how general properties of efficient communication give rise to these universal word order properties of human language.

Grammars and Grammar Data

Following a long tradition in theoretical and computational linguistics, we formalize the grammatical structure of languages using dependency trees (37–40). This linguistic formalism represents grammatical dependencies as directed arcs between syntactically related words, annotated with grammatical relations like subject or object (Figure 2). While syntactic formalisms vary, the dependency grammar community has an agreed representation format which has been used to annotate corpora of text from dozens of languages (41), and there are computational methods for mapping syntactic structures in this representation to other standard linguistic formalisms (42).

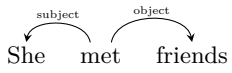


Fig. 2. An English sentence with annotated syntactic relations.

Our models require a sample of syntactic structures as actually used by speakers across different languages, for which we draw on the recent Universal Dependencies project (41), which has collected and created syntactic annotations for several dozens of different languages. 51 languages had sufficient data for our purposes. These corpora represent a typologically and genetically diverse group of languages. We obtained a total of 11.7M words in 700K sentences annotated with syntactic structures; with a median of 117K words and 7K sentences for each individual language.

Study 1: Relative efficiency of languages

We first ask whether the grammars of human languages evolve towards optimizing efficiency of communication. To do this we compare the efficiency of the actual grammars of the 51 languages from the Universal Dependencies datasets to randomly constructed baseline grammars.

The grammars of natural languages specify how the different words in a syntactic structure are ordered into a sentence, i.e., a string of words (43). This is illustrated in Figure 3: We show how four different grammars order objects, adpositional phrases, and adpositions. 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.

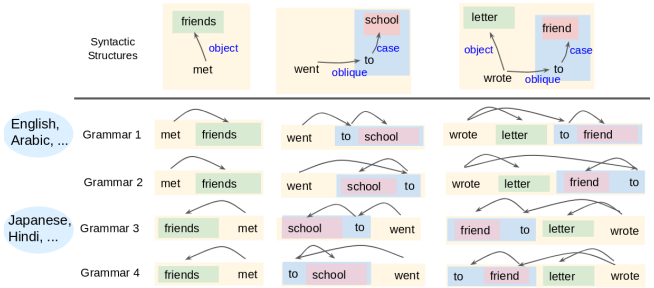


Fig. 3. Grammars define consistent ordering rules for syntactic structures. Here, Grammars 1 and 2 order the object after the verb, Grammars 3 and 4 order the object before the verb. Grammars 1 and 3 conform to the Greenberg correlations and are common around the world; Grammars 2 and 4 are rare or impossible.

Beyond the syntactic relations exemplified in Figure 3, human languages have further types of syntactic relations. The Universal Dependencies project, the source of our data, defines a total of 38 syntactic relations. We adopt the grammar model of Gildea and colleagues (44–46); a grammar assigns a weight from $[-1, 1]$ to each of these 38 syntactic relations, and orders words according to the weights assigned to their relations (See *Materials and Methods* for details).

Given a database of syntactic structures (such as those at the top of Figure 3), obtained from a corpus of some real language L , we can apply a grammar to order the structures in the database into a dataset of counterfactual sentences belonging to a hypothetical language defined by that grammar (Figure 3). This hypothetical language has identical syntactic structures and grammatical relations as the true language L, but different word order.

We create baseline grammars by randomly sampling the weights for each syntactic relation. These baseline grammars have systematic word order rules similar to natural language, but do not exhibit any correlations among the orderings of different syntactic relations. All four grammars in Figure 3 are equally likely under this baseline distribution.

For every one of the 51 languages, we construct 50 counterfactual versions by randomly creating 50 baseline grammars, applying these to obtain counterfactual orderings for all syntactic structures that were available for that language.

Following the information-theoretical literature on language processing, we formalize the efficiency of a language as a weighted combination of two terms: the amount of information that utterances contain about the underlying messages,

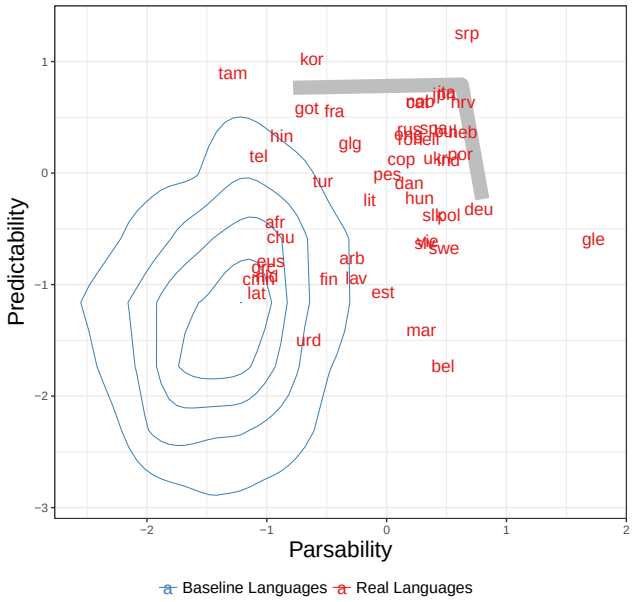


Fig. 4. Predictability and parseability of 51 languages (red), indicated by ISO codes, compared to baseline word order grammars (blue distribution). Predictability and parseability scores are z-scored within language. The gray curve indicates the Pareto frontier of computationally optimized grammars, averaged over the 51 languages.

and the cost or difficulty of communication (33–36, 47, 48). We model the informativity term as the degree to which listeners can reconstruct syntactic structures from an utterance, i.e., the **parseability** of the language. We model the cost or complexity term as the **predictability**, or negative entropy, of the utterances. We then estimate predictability and parseability using standard neural network models, and combine them linearly to produce a single model for efficiency. See *Materials and Methods* for details.

For every one of the 51 languages, we computationally construct grammars that are *optimized* for parseability, predictability, and efficiency (see *Materials and Methods*). This optimization problem is challenging because both the parseability and predictability of a sentence can only evaluated globally, in the context of an entire language. We address this challenge by using neural network methods to estimate parseability and predictability, and by introducing a simple, differentiable computational formalism for describing grammatical regularities. This setup means that it is possible to find optimal grammars by standard methods such as stochastic gradient descent (see SI section S5). For each grammar, we report predictability and parseability as estimated on the data resulting from ordering the syntactic structures from the corpus according to the grammar.

In Figure 4, we plot predictability and parseability of the grammars of 51 languages, together with the distribution of random baseline grammars, and the Pareto frontier defined by computationally optimized grammars. Languages are attracted towards the Pareto frontier and away from the region of the baseline languages. The majority of real languages are above and to the right of their baseline equivalents, demonstrating that they are relatively high in predictability and/or parseability. 100 % of languages improve on either predictability or parseability ($p < 0.05$, by one-sided t-test, with Bonferroni correction and Hochberg's step-up procedure (49)). 90 % of languages improve over the baselines in parseability

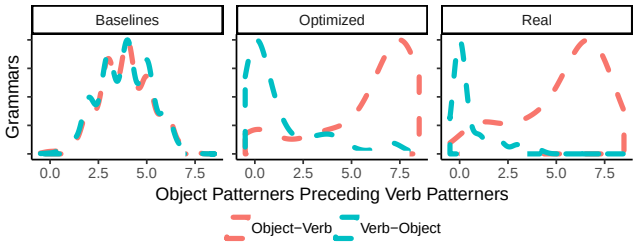


Fig. 5. Efficiency optimization produces grammars where the orders of the eight relations in Table 1 are strongly correlated with the order of verb and object. We arrange grammars (baseline, optimized, real) by the number of relations where the language patterns with Japanese (as opposed to with Arabic), and plot a kernel density estimate. Object-Verb order leads to grammars where object patterners precede (like Japanese); verb-object order leads to verb patterners preceding (like Arabic). Baseline grammars show no such correlation.

($p < 0.05$), 80 % improve in predictability ($p < 0.05$). See SI Section S3 for additional analyses.

Study 2: Greenberg Word Order Correlations

We have found that the grammars of human languages concentrate along the Pareto frontier of parseability and predictability. This does not yet, however, say anything about what grammatical properties characterize Pareto-optimal languages in general, or which properties of human languages make them efficient.

We show that all languages close to the Pareto frontier – both real and counterfactual ones – are highly likely to satisfy Greenberg's correlation universals. That is, optimizing for efficiency produces languages that satisfy these correlations. In contrast, the baseline grammars are constructed without any correlations between the ordering of different syntactic relations, and will therefore not satisfy any of those universals.

We first considered this question for the 51 real languages. Among the real grammars of the 51 languages, the number of satisfied correlations is strongly correlated with efficiency ($\rho = 0.61$, $p < 0.0001$).

We next examine those grammars from Study 1 that we had computationally optimized for efficiency. We controlled for variation across different optima by creating eight optimized grammars for each of the 51 datasets of syntactic structures from real languages. For each language, we created four grammars with verb-object order, and four object-verb grammars. We test whether the process of efficiency optimization produces the Greenberg correlations.

For each grammar (baseline, optimized, and real), we computed how many of the eight relations in Table 1 had the same order as Japanese (in contrast to Arabic). Figure 5 shows the results, separately for grammars with verb-object and object-verb orders. In optimized grammars, the order of the eight relations is strongly correlated with the placement of the object, similar to the 51 real languages in our sample. In contrast, baseline languages show no correlation.

So far, our results show that languages are efficient in part because they mostly satisfy the Greenberg correlations. We asked whether efficiency optimization predicts every one of the eight correlations to hold in most languages. We constructed a Bayesian multivariate logistic regression model predicting which of the eight correlations an optimized grammar satisfies. We controlled for variation between the syntactic structures used in different languages and language families by entering the language and language family as random effects. See SI Section S4.3 for details.

	Correlates with... verb *wrote*	object *letters*	Real	Optimized
①	adposition *to*	NP *a friend*		
②	copula *is*	NP *a friend*		
③	auxiliary *has*	VP *written*		
④	noun *friend*	genitive *of John*		
⑤	noun *books*	relative clause *that you read*		
⑥	complementizer *that*	S *she has arrived*		
⑦	verb *went*	PP *to school*		
⑧	want *wants*	VP *to leave*		

Table 2. Efficiency optimization accurately predicts the Greenbergian correlations. For each correlation, we provide the prevalence among the actual grammars of the 51 languages (left), and the posterior distribution of the prevalence among grammars optimized for efficiency (right). Efficiency optimization predicts all eight correlations to hold in the majority of grammars. make visually clear that it's 51 languages

In Table 2, we compare the prevalence of the eight correlations in real and optimized languages. For the real languages, we indicate how many of the 51 languages satisfy a correlation. For the optimized languages, we indicate the posterior distribution of the proportion of satisfying languages, obtained from the mixed-effects analysis. Grammars optimized for efficiency accurately predict all eight correlations to hold at prevalences significantly greater than 50%, similar to actual human languages. In the multivariate mixed-effects analysis, efficiency optimization predicts all eight correlations to hold across languages (posterior probability 0.9911). Optimizing for only predicability or only parseability does not predict all of the correlations (See SI Section S4.4).

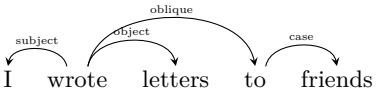
Discussion

We found that the grammars of natural languages are more efficient than baseline grammars, and that a large subset of the Greenbergian word order correlations can be explained in terms of optimization of grammars for efficient communication.

Our work makes crucial use of neural network models for estimating the efficiency of languages. This method currently requires large computational resources; it still takes about three weeks to create optimized grammars for 51 languages, even with specialized hardware. We believe that further advances in machine learning will reduce the computational cost, making this approach more widely applicable.

What makes the grammars of human languages efficient? Study 2 shows that Greenberg's correlations are one key property that real languages share with optimal grammars. Prior work has suggested *Dependency Length Minimization* as another characteristic of efficient word order. This is the idea that word order minimizes the average distance between syntactically related words. It is known that human languages reduce this distance compared to random baselines (50–52).

Indeed, 100% of grammars optimized for efficiency also reduce average distance between related words compared to baselines ($p < 0.05$, by one-sided t-test). To some extent, the Greenberg correlations help reduce the distance between related words. Consider again the sentence 'I wrote letters to friends' (cf. Figures 1 and 3). Both real and optimized grammars of English linearize its syntactic structure as follows:



Note that this ordering exhibits correlations (1) and (7) from Table 1. Among all possible ways of ordering this syntactic structure, this ordering also minimizes the average distance between any two syntactically related words: E.g., inverting 'to' and 'friends' would increase the distance between 'wrote' and 'to'. Understanding mathematically why dependency length minimization improves efficiency is an important problem for future research.

An idea related to, but distinct from, functional optimization is the idea that grammars are biased towards simplicity (9). It has been proposed that languages have a single head-directionality parameter and that this accounts for the Greenberg correlations (20). As an explanation of correlations, this idea makes strong predictions, and turns out to overpredict correlations (15). Nevertheless, future research should examine whether there are more principled connections between communicative efficiency and grammar simplicity.

A major question for functional explanations for linguistic universals is: *how* do languages end up 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 (53, 54)? 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 evidence against the idea that word order universals are best explained in terms of learning biases that are irreducibly arbitrary and genetic in nature (55).

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 trees of syntactic relations. The question arises of whether these more basic properties themselves might be explainable in terms of efficient communication.

Beyond those universals that we examined, Greenberg (13) described other universals; some of those are syntactic and could be tested with more specific annotation of datasets; others are not syntactic and are beyond the scope of this study. Beyond Greenberg's work, there are remaining word order universals not captured by our model, such as the trade-off of rich morphological marking and flexibility in word order (56–58). Our models do not capture this trade-off because the grammar model assumes fixed word order. Future work can investigate whether these and other remaining universals can be explained using more sophisticated models of how syntactic structures are transduced into strings of words, based on larger databases with richer annotations.

These questions notwithstanding, our work provides strong evidence that the grammatical structure of languages is shaped by the need to support efficient communication. Beyond our present results, we provide a computational framework in which theories of the efficiency optimization of languages can be tested. While our study has focused on syntax, our results suggest that this method can be fruitfully applied to testing efficiency explanations in other domains of language structure.

Materials and Methods

Corpus Data. We base our experiments on the Universal Dependencies 2.1 treebanks (41). We use all languages for which at least one treebank with a training partition was available, a total of 51 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 (50), we applied automated conversion to a more standard formalism, modifying each treebank by inverting these dependencies, and promoting the dependent to the head position.

Word Order Grammars. We adapt the grammar model of (44) to UD. A grammar assigns a parameter $x_\tau \in [-1, 1]$ to every relation τ belonging to the 38 universal syntactic relations defined by UD. A syntactic structure, consisting of a set of words and syntactic relations between them, is then ordered into a string of words recursively starting from the root; the dependents of a word then are ordered around the head according to the values x_τ corresponding to their syntactic relations; those dependents where $x_\tau < 0$ are ordered before the head; the others after the head. See SI Section S5.2 for the methodology used to extract the languages' actual grammars from datasets, and for validation against expert judgments.

Formalizing Efficiency. We adopt the formalization of language efficiency of (33), equivalent to a deterministic version of the Information Bottleneck (35). Very similar formalizations of Zipf's ideas have been proposed across the information-theoretic literature on language (34, 36, 47). See SI Section S2 for discussion.

In this framework, the overall efficiency of language is a weighted combination of terms representing the amount of information that utterances contain about the underlying messages, and the cost of communication (33–36, 47). We model the first term as the degree to which listeners can reconstruct syntactic structures from an utterance, i.e., the **parseability** if the language. This is formalized as the amount of information that utterances u provide about their underlying syntactic structures t:

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

where the sum runs over all possible pairs of utterances u and syntactic structures t in the language.

Again following (33), we formalize the complexity of a language as its entropy (33, 59, 60). This corresponds to the average word-by-word surprisal, the degree to which sentences are unpredictable from the general statistics of the language. Surprisal has been found to be a highly accurate and general predictor of human online processing difficulty (61–63). In expectation over all utterances u in a language, the negative surprisal describes the **predictability**, or negative entropy, of the utterances:

$$R_{Pred} := -H[\mathcal{U}] = \sum_{u} p(u) \log p(u) \quad [2]$$

where the sum runs over all possible sentences u that belong to the language.

Maximizing one of the two scoring functions under a constraint on the other function (e.g., maximizing parseability under a constraint

on the minimal predictability) amounts to maximizing a weighted combination of the two scoring functions (33, 35):

$$R_{Eff} := R_{Pars} + \lambda R_{Pred} \quad [3]$$

with an interpolation weight $\lambda \in [0, 1)$ that controls the relative strength of the two pressures. When optimizing grammars for efficiency, we set $\lambda := 0.9$ in Equation 3 in order to give approximately equal weight to both components. See SI section S2.2 for mathematical discussion of λ, and robustness to other choices.

We estimate predictability using LSTM recurrent neural networks (64), general sequence models that are the strongest known predictors of the surprisal effect on human processing effort (65, 66). We estimate parseability using a generic neural network architecture that casts recovery of syntactic structures as a minimum spanning tree problem (67, 68). All parseability and predictability values are reported on the held-out (dev) partitions from the predefined split for each UD corpus. Grammars are optimized for efficiency by simultaneous gradient descent on the parameters of the grammar and these neural models. See SI Sections S5-S7 for details and for robustness of our results to these choices.

Data Availability. The efficiency optimization results from Table 2 were preregistered: http://aspredicted.org/blind.php?x=ya4qf8.

Code and results are available at https://github.com/m-hahn/grammar-optim.

ACKNOWLEDGMENTS. We thank Ted Gibson, Michael C. Frank, Judith Degen, Paul Kiparsky, and audiences at CAMP 2018 and CUNY 2019 for helpful discussion.

1. Chomsky N (2005) Three factors in language design. *Linguistic Inquiry* 36(1):1–61.
2. Hauser M, Chomsky N, Fitch W (2002) The faculty of language: What is it, who has it, and how did it evolve? *Science* 298(5598):1569.
3. Berwick RC, Weinberg A (1984) *The grammatical basis of linguistic performance.* (Cambridge, MA: MIT Press).
4. Hawkins JA (1994) *A performance theory of order and constituency.* (Cambridge University Press, Cambridge).
5. Zipf GK (1949) *Human behavior and the principle of least effort.* (Addison-Wesley Press, Oxford, UK).
6. Croft W, Cruse A (2004) *Cognitive Linguistics.* (Cambridge University Press, Cambridge, UK).
7. Goldberg A (2005) *Constructions at work: The nature of generalization in language.* (Oxford University Press, Oxford, UK).
8. Pinker S, Bloom P (1990) Natural language and natural selection. *Behavioral and brain sciences* 13:707–784.
9. Smith K, Tamariz M, Kirby S (2013) Linguistic structure is an evolutionary trade-off between simplicity and expressivity in *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 35.
10. Nowak MA, Krakauer DC (1999) The evolution of language. *Proceedings of the National Academy of Sciences* 96:8028–8033.
11. Nowak MA, Komarova NL (2001) Towards an evolutionary theory of language. *Trends in cognitive sciences* 5(7):288–295.
12. Nowak MA, Komarova NL, Niyogi P (2002) Computational and evolutionary aspects of language. *Nature* 417(6889):611.
13. 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.
14. Dryer MS, Haspelmath M (2013) *WALS Online.* (Max Planck Institute for Evolutionary Anthropology, Leipzig).
15. Dryer MS (1992) The Greenbergian word order correlations. *Language* 68(1):81–138.
16. Lehmann WP (1973) A structural principle of language and its implications. *Language* 49:47–66.
17. Vennemann T (1974) Theoretical word order studies: Results and problems. *Papiere zur Linguistik* 7:5–25.
18. Jackendoff R (1977) *X-bar syntax: A study in phrase structure.* (MIT Press).
19. Frazier L (1985) Syntactic complexity. *Natural language parsing: Psychological, computational, and theoretical perspectives* pp. 129–189.
20. Chomsky N (1988) *Language and Problems of Knowledge: The Managua Lectures.* (MIT Press, Cambridge, MA).
21. Gabelentz Gvd (1901) *Die Sprachwissenschaft, ihre Aufgaben, Methoden, und bisherigen Ergebnisse.* (Weigel, Leipzig).
22. Hockett CF (1960) The origin of language. *Scientific American* 203(3):88–96.
23. Givón T (1991) Markedness in grammar: distributional, communicative and cognitive correlates of syntactic structure. *Stud Lang* 15:335–370.
24. Hawkins JA (2004) *Efficiency and complexity in grammars.* (Oxford University Press, Oxford).
25. Hawkins JA (2014) *Cross-linguistic variation and efficiency.* (Oxford University Press, Oxford).
26. 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.
27. 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.
28. Jaeger TF, Tily HJ (2011) On language 'utility': Processing complexity and communicative efficiency. *Wiley Interdisciplinary Reviews: Cognitive Science* 2(3):323–335.
29. Horn L (1984) Toward a new taxonomy for pragmatic inference: Q-based and r-based implicature. *Meaning, form, and use in context: Linguistic applications* 11:42.
30. Lindblom B (1990) Explaining phonetic variation: A sketch of the h&h theory in *Speech production and speech modelling.* (Springer), pp. 403–439.
31. Schwartz JL, Boë LJ, Vallée N, Abry C (1997) The dispersion-focalization theory of vowel systems. *Journal of phonetics* 25(3):255–286.
32. Haspelmath M (2006) Against markedness (and what to replace it with). *Journal of linguistics* 42(1):25–70.
33. Ferrer i Cancho R, Solé RV (2003) Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences* 100(3):788–791.
34. Frank MC, Goodman ND (2012) Predicting pragmatic reasoning in language games. *Science* 336(6084):998–998.
35. 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.
36. Kemp C, Regier T (2012) Kinship categories across languages reflect general communicative principles. *Science* 336(6084):1049–1054.
37. Hays DG (1964) Dependency theory: A formalism and some observations. *Language* 40:511–525.
38. Melčuk IA (1988) *Dependency syntax: Theory and practice.* (SUNY Press).
39. Corbett GG, Fraser NM, McGlashan S (1993) *Heads in Grammatical Theory.* (Cambridge University Press, Cambridge).
40. Tesnière L, Kahane S (2015) *Elements of structural syntax.* (John Benjamins Publishing Company New York).
41. Nivre J, et al. (2017) Universal dependencies 2.1. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.
42. Boston MF, Hale JT, Kuhlmann M (2009) Dependency structures derived from minimalist grammars in *The Mathematics of Language.* (Springer), pp. 1–12.
43. Adger D (2015) Syntax. *Wiley Interdisciplinary Reviews: Cognitive Science* 6(2):131–147.
44. 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.
45. Gildea D, Temperley D (2010) Do grammars minimize dependency length? *Cognitive Science* 34(2):286–310.
46. Gildea D, Jaeger TF (2015) Human languages order information efficiently. *arXiv* 1510.02823.
47. Regier T, Kemp C, Kay P (2015) Word meanings across languages support efficient communication in *The Handbook of Language Emergence.* (Wiley-Blackwell, Hoboken, NJ), pp. 237–263.
48. Goodman ND, Stuhlmüller A (2013) Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science* 5(1):173–184.
49. Hochberg Y (1988) A sharper bonferroni procedure for multiple tests of significance. *Biometrika* 75(4):800–802.
50. 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.
51. Liu H, Xu C, Liang J (2017) Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews.*
52. Temperley D, Gildea D (2018) Minimizing syntactic dependency lengths: Typological/cognitive universal? *Annual Review of Linguistics* 4:1–15.
53. 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.
54. Culbertson J, Smolensky P, Legendre G (2012) Learning biases predict a word order universal. *Cognition* 122(3):306–329.
55. 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.
56. Jespersen O (1922) *Language: Its nature, development, and origin.* (Henry Holt and Co., New York).
57. McFadden T (2003) On morphological case and word-order freedom in *Proceedings of the Berkeley Linguistics Society.*
58. 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.
59. 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.
60. Futrell R (2017) Ph.D. thesis (Massachusetts Institute of Technology, Cambridge, MA).
61. 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.
62. Levy R (2008) Expectation-based syntactic comprehension. *Cognition* 106(3):1126–1177.
63. Smith NJ, Levy R (2013) The effect of word predictability on reading time is logarithmic. *Cognition* 128(3):302–319.
64. Hochreiter S, Schmidhuber J (1997) Long short-term memory. *Neural Computation* 9(8):1735–1780.
65. Frank SL, Bod R (2011) Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science* 22(6):829–834.
66. Goodkind A, Bicknell K (2018) Predictive power of word surprisal for reading times is a linear function of language model quality in *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018).* (Association for Computational Linguistics, Salt Lake City, UT), pp. 10–18.
67. 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.
68. Zhang X, Cheng J, Lapata M (2017) Dependency parsing as head selection in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers.* (Valencia, Spain), pp. 665–676.