Supplementary Information: Crosslinguistic word order variation reflects evolutionary pressures of dependency and information locality

Michael Hahn
Department of Linguistics, Stanford University

Yang Xu
Department of Computer Science, Cognitive Science Program, University of Toronto

mhahn2@stanford.edu

Contents

Part I

Detailed Formulations of Efficiency and Grammar

S1 Information Locality

S1.1 Formalizing Information Locality

Information Locality is motivated by the interaction of two prominent psycholinguistic perspectives on what determines human comprehension difficulty in processing syntactic structure. **Memory-based theories** [1, 2, 3] propose that comprehension difficulty arises from the difficulty of retrieving and integrating information from preceding context. **Expectation-based theories** [4, 5] states that difficulty arises at points in a sentence that are hard to anticipate from the preceding context. Jointly considering both perspectives leads to the prediction that words should be easy to process when they are easy to predict from preceding context (as predicted by expectation-based accounts) unless the relevant predictive information has been affected by memory decay or interference (as predicted by memory-based accounts) [6, 7, 8, 9] (see also [10, 11, 12] for closely related proposals). Under this perspective, word order enables efficient processing when predictive information about a word is concentrated in its recent past, so that it can be utilized before it suffers memory decay or interference. This idea has been formalized using the term **Information Locality** by Futrell et al. [8] and Hahn et al. [9], though it is closely related to proposals from preceding work on the role of efficiency in language [13], of relations between usage statistics and conceptual structure in language [14], information-theoretic studies of language [15], and also to classical experimental findings about the role of contextual constraint on the occurrence of words [16, 17].

The key formal notion is the *conditional mutual information* between two words X_i, X_{i+t} at a distance t (Figure S1 A):

$$I_t := I[X_i, X_{i+t} | X_{i+1} \dots X_{i+t-1}] = \mathbb{E}_X \left[\log \frac{P(X_{i+1} | X_i, X_{i+1} \dots X_{i+t-1})}{P(X_{i+t} | X_{i+1} \dots X_{i+t-1})} \right] \quad (1)$$

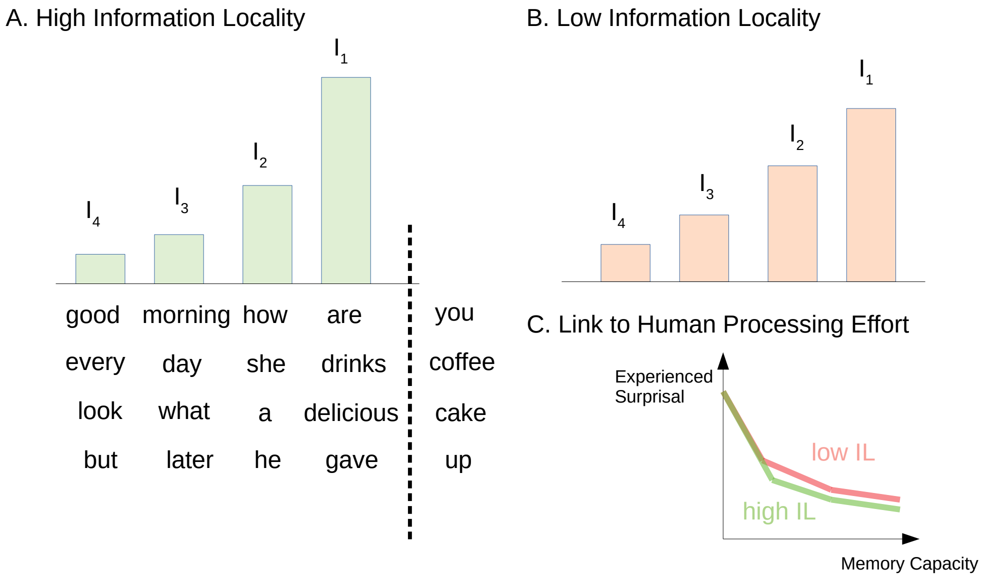


Figure S1: Information Locality, Mutual Information, and their links to psycholinguistic processing effort. (A) The conditional mutual information I_t measures how much predictive information a word t words in the past provides about the next word, on average across a corpus. While we only show values up to $t = 4$, t runs through all integers up to the length of the longest sentence in the corpus. In human language, I_t is largest at $t = 1$ and quickly decays as t increases. (B) Another possible situation, where the predictive information is spread out more widely over the past context. Here, I_1 is lower and I_t decays more slowly. Such a situation corresponds to a lower degree of Information Locality than in A. (C) The decay of I_t is linked to two aspects of psycholinguistic processing: memory and surprisal. For an individual comprehender, there is a tradeoff whereby a higher memory capacity lowers surprisal on average. The shape of this tradeoff depends on I_t and thus on word order: If I_t decays more quickly (green), a comprehender can achieve lower surprisal at the same memory budget, i.e., the tradeoff is more efficient. The efficiency of the tradeoff can be measured by its area under the curve (AUC), which is lower for the green curve.

where the expectation X runs over all sequences of words in the statistics of the language. The conditional mutual information I_t measures how much predictive information words that are t words apart provide about each other's identity, controlling for information that is redundant with the $t-1$ intervening words, and averaging across all such word pairs in a corpus.

Mutual information is closely related to two other well-studied information quantities [e.g. 18, 19, 20]: the *entropy rate* $H[X_t|\dots,X_{t-2},X_{t-1}]$ measuring how unpredictable words are in context on average, and the *unigram entropy* $H[X_t]$ measuring the diversity of the distribution over individual words, i.e., how unpredictable a word is without context. The difference between the two turns out to be[1]

$$H[X_t]-H[X_t|\dots,X_{t-2},X_{t-1}] = \sum_{t=1}^{\infty} I_t \tag{2}$$

which measures the total average amount of predictive information contained in the preceding context.

Formalizing Information Locality Broadly speaking, Information Locality asserts that language favors orderings where a higher fraction of the overall predictive information (2) is contained at words in the recent context, and only a small fraction is contained in words farther in the past. This is equivalent to stating that I_t is high for small distances t and decays relatively steeply as t increases (Figure S1 A–B).

In this paper, we choose maximization of the mutual information between adjacent words $I[X_i,X_{i+1}]$ as a particularly simple operationalization of Information Locality:

$$I_1 = I[X_i,X_{i+1}] = \mathbb{E}_X\left[\log \frac{P(X_{i+1}|X_i)}{P(X_{i+1})}\right] \tag{3}$$

If this quantity is high, a larger fraction of (2) is provided by the immediately preceding word. A smaller fraction of the overall predictive information from the past is then contained in context further in the past. Conversely, if I_1 is small, a larger fraction of (2) must be contained further in the past.

We next discuss how this relates to proposals from prior work.

Area under Memory-Surprisal Tradeoff Curve Hahn et al. [9] provide a mathematical derivation of information locality in terms of a memory-surprisal tradeoff, combining the expectation-based and memory-based perspectives with a general information-theoretic analysis. This is formalized by the following theorem about comprehenders processing a stream of words using some (otherwise arbitrary) memory representations M (Figure S1 C): If $T \geq 0$ is an integer chosen so that the information-theoretic capacity of the listener's memory representation M satisfies

$$M \leq \sum_{t=1}^{T} t \cdot I_t \tag{4}$$

then this comprehender's average surprisal S satisfies

$$S \geq H[X_t|\dots X_{t-1}] + \sum_{t=T+1}^{\infty} I_t \tag{5}$$

They showed that comprehenders can achieve a lower surprisal at the same memory capacity when I_t decays faster. This happens because of the factor t in (4), which creates a higher memory cost due to predictive information I_t at higher distances t. Our chosen formalization (3) emerges in the limit of small memory capacities: For $T=1$, the surprisal bound precisely equals $H[X_t]-I_1$. A higher value of I_1 thus guarantees a lower (i.e., more favorable) surprisal at low memory budgets.

Hahn et al. [9] proposed to quantify information locality in terms of the area under the memory-surprisal tradeoff curve (AUC): a lower AUC corresponds to a faster decay of surprisal as memory capacity increases, and thus higher IL. In Section S23, we compare to results obtained when quantifying IL in terms of this AUC measure as estimated by Hahn et al. [9].

[1] $H[X_t]-H[X_t|\dots,X_{t-2},X_{t-1}] = I[X_t:(\dots,X_{t-2},X_{t-1})]$, which is $\sum_{t=1}^{\infty} I_t$ by the chain rule of mutual information.

N-Gram Surprisal [21] showed that the word orders of five languages minimize trigram surprisal (i.e., $H[X_t|X_{t-2}, X_{t-1}]$), compared to most other possible orderings. While they justified trigram surprisal as an approximation to surprisal as considered in expectation-based models of processing, it can also be justified as a formalization of information locality: trigram surprisal equals $H[X_t] - I_1 - I_2$; it is thus low if and only if $I_1 + I_2$ is high.

Decay of Unconditional Mutual Information A line of prior work has considered the *unconditional mutual information* J_t [15, 8, 14]:

$$J_t := I[X_i, X_{i+t}] = \mathbb{E}_X \left[\log \frac{P(X_{i+1}|X_i)}{P(X_{i+1})} \right] \tag{6}$$

This differs from I_t in that it does not factor out information redundant with intervening information. Note that $I_1 = J_1$; thus, our formalization (3) equivalently states that J_t decays quickly as t increases.

Information locality was stated in terms of unconditional mutual information by Futrell et al. [8], who provided an approximate mathematical derivation in terms of minimizing surprisal under a certain class of memory loss models. While they did not provide a full operationalization of Information Locality, they proposed that language favors that words are close together when they have a high (unconditional) mutual information, i.e., J_t decays quickly.

Further related to the principle of Information Locality, Culbertson et al. [14] show that the typologically most frequent relative orderings of noun phrase modifiers are such that modifiers are closer to the noun if they have higher mutual information with the noun. While they interpreted mutual information as reflecting statistical properties of the world that correlate with conceptual structure, their account is fully compatible with the principle of Information Locality as derived from theories of psycholinguistic processing effort.

Decaying Cue Effectiveness Relatedly, Qian and Jaeger [13] argue that the effectiveness of past predictive information in language production decreases over distance. They studied the overall predictive information (2) (their "cumulative discourse informativity", Formula (4) in their paper) and the decay of I_t (their "cue effectiveness", Formula (3) in their paper), proposing that I_t decays over distances t due to, among other factors, limitations of human memory. They fitted a power law to the decay of cue effectiveness; in this framework, a steep decay is reflected in the coefficients of the power law. Information locality can also be linked to classical findings that most predictive information about a word, at least as utilized by humans, comes from a few preceding words [16, 17].

S1.2 Estimating Mutual Information

Mutual information is defined in terms of an idealized statistical distribution over all possible sentences; it is thus necessary to approximate it using the available finite corpus data. We follow the approach of Gildea and Jaeger [21] and Study 3 of Hahn et al. [9], drawing on long-standing techniques in natural language processing (see Section S23 for a second estimation method, used in Study 2 of Hahn et al. [9]).

We split each dataset into a training set and a held-out set. While the UD datasets have predefined splits, those vary substantially in the train/held-out ratio across languages. We therefore, for each language, randomly sampled a subset whose size was the greater of 100 sentences and 5% of all sentences, and used those as held-out data, and the remainder as training data. We estimate the probabilities $p(x_t|x_{t-1})$ using counts from the training set, and estimate the entropies $H[X_t]$, $H[X_t|X_{t-1}]$ as cross-entropies on the held-out data:

$$H[X_t] \approx - \sum_{i=1}^{|HeldOut|} \log p(x_i) \tag{7}$$

$$H[X_t|X_{t-1}] \approx - \sum_{i=1}^{|HeldOut|} \log p(x_i|x_{i-1}) \tag{8}$$

I_1 is then estimated as the difference of these cross-entropies. This approach of estimating mutual information as a difference of cross-entropies is a well-established method with theoretical guarantees [22], avoiding an overestimation bias that would result from naively applying the definition of mutual information to the full dataset.

The method for estimating probabilities $p(x_t|x_{t-1})$ exactly follows Study 3 of Hahn et al. [9] and is based on Kneser-Ney Smoothing [23], which we describe here for completeness. First, the unigram probabilities are estimated using Laplace smoothing as

$$p(w_t) := \frac{N(w_t)+1}{|Train|+|V|\cdot 1} \tag{9}$$

where $N(w_t)$ is the number of occurrences of w_t in the training data. Here $|Train|$ is the number of tokens in the training set, $|V|$ is the number of types occurring in train or held-out data.

Then, conditional probabilities $p_2(w_t|w_{t-1})$ are estimated as follows. For a sequence w_1w_2, let $N(w_1w_2)$ be the number of times w_1w_2 occurs in the training set. If $N(w_{t-1}w_t) = 0$, set

$$p(w_t|w_{t-1}) := p(w_t) \tag{10}$$

Otherwise, we interpolate between second-order and first-order estimates:

$$p(w_t|w_{t-1}) := \frac{\max(N(w_{t-1}w_t)-1, 0.0) + \#\{w : N(w_{t-1}w) > 0\}\cdot p(w_t)}{N(w_{t-1})} \tag{11}$$

Kneser and Ney [23] show that this definition results in a well-defined probability distribution, i.e., $\sum_{w\in V} p(w|w_{t-1}) = 1$. This method can be justified as approximate Bayesian inference assuming a Zipfian-like distribution over words [24].

S2 Ordering Grammars

S2.1 Ordering Grammar Formalism

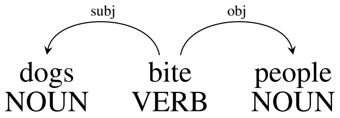
We adopt the word order grammar formalism of [25, 26, 21] to Universal Dependencies. The original grammar formlism of [25] is defined for constituency treebanks; it defines weights for each combination of parent and child constituent category (e.g., "NP→JJ" for the position of the adjective within the noun phrase). We adapt this to Universal Dependencies by defining weights for dependency relation labels (e.g. *amod* for the noun-adjective dependency).

Dependents of a head are ordered in ascending order by their weights, so that dependents with negative weights appear before the head and dependents with positive weights appear after the head.

For instance, a grammar might define the weights (among others)

nsubj : -0.8
obj : 0.3

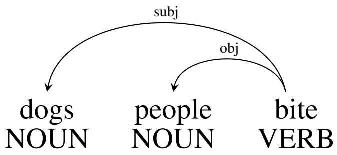
Applying this to a simple transitive sentence would result in SVO order:



In contrast, the following grammar, where both weights are negative, results in SOV order:

nsubj : -0.8
obj : -0.3

as in the following example:



S2.2 Optimization Methods

Hill-Climbing Method The hill-climbing method is adopted from the method of Gildea and Temperley [25]. It first randomly initializes the weights of the grammar. In every iteration, it then randomly chooses one relation and changes the grammar by moving this relation to a randomly selected new position. If the objective function (a linear combination of IL and DL) improves, the new grammar is adopted, else it is discarded. We iterate this until the grammar remains stable for 2K iterations, for at most 10K iterations.

Gradient Descent Method We further use the gradient-based optimization method of [27] to optimize DL, which converges more quickly than the hill-climbing method, in particular on larger datasets.[2] This method considers a probabilistic extension of the grammar formalism where each grammar defines a distribution over possible linearizations of a tree; grammars as defined above correspond to the special case where the distribution is always concentrated on one linearization (i.e., it is deterministic). We refer to [27] for the precise definition of this extension. This extension makes the average dependency length a *differentiable* function of the grammar parameters, opening the door to the use of gradient-based optimization algorithms for ordering grammars. The optimization method then applies stochastic gradient descent using the REINFORCE estimator [28] to optimize the average dependency length across the trees in the corpus and the possible linearizations of each tree. Over the course of optimization, the probabilistic grammars converge to essentially deterministic ones that approximately minimize average dependency length across the trees in the corpus.

S3 Interpolating Efficiency Plane and Pareto Frontier

Here, we describe how we interpolated subject-object position congruence throughout the efficiency plane, and how we approximated the Pareto frontier. We made all choices before evaluating the hypotheses tested in the paper. Results do not depend on the smoothing method: See Figure S18 for an analysis of coadaptation based on the raw samples that do not depend on the smoothing method, showing equivalent results.

Distribution of Subject-Object Position Congruence Given the set of grammar samples (obtained through approximate optimization or random generation) $\xi_i = (x_i, y_i)$ ($x_i =$ IL, $y_i =$ DL) with associated subject-object position congruences z_i, for each point $\xi = (x, y)$ in the plane spanned by DL and IL,

[2] While this method is also applicable to optimizing mutual information (IL), it does not offer an efficiency advantage over the hill-climbing method there.

we predict the average subject-object position congruence of grammars at this point with a normalized Gaussian kernel as

$$f(x,y) := \sum_{i=1}^{N} w_i z_i \tag{12}$$

where

$$w_i \propto L_1(x_i - x)^2 + L_2(y_i - y)^2 \tag{13}$$

and $\sum_i w_i = 1$, and $L_1, L_2 > 0$ are chosen to minimize the regularized leave-one-out objective:

$$\frac{1}{N}\sum_{i=1}^{N} |f_i(x,y) - z_i|^2 + \lambda \cdot (L_1^2 + L_2^2) \tag{14}$$

where f_i arises by leaving out ξ_i from the dataset in the definition of f. We determined a small regularization weight $\lambda = 0.00001$ to prevent smoothing artifacts arising due to excessively large weights L_i. Optimization uses 5K iterations of random search over $L_1, L_2 \in [0,100] \times [0,100]$.

Pareto Frontier We fit the approximate Pareto frontier as a spline covering the convex hull of all grammar samples. The definition is very similar to standard cubic splines [29], except that we constrained the spline to be convex and monotonic. More precisely, we selected all sampled grammar points $\xi_i = (x_i, y_i)$ that were not Pareto-dominated by any other point in the convex hull. For each segment between adjacent points x_i, x_{i+1}, we defined a cubic polynomial $g_i(x)$, and determined the coefficients of these cubic polynomials to maximize the area under the curve $\sum_{i=1}^{N} \int_{x_i}^{x_{i+1}} g_i$ (thus, making the spline fit as closely to the convex hull as possible), suject to the constraints of (i) lower-bounding the convex hull: $g_i(x_i) \leq y_i$, (ii) continuity: $g_i(x_{i+1}) = g_{i+1}(x_{i+1})$, (iii) continuity of the slope $g'_i(x_{i+1}) = g'_{i+1}(x_{i+1})$, (iv) convexity: $g''_i \geq 0$, (v) monotonicity: $g'_i \leq 0$. This is a standard linear program, which we solved using cvxpy [30].

Part II
Languages and Datasets

S4 Corpora and Corpus Sizes

As described in Methods, we included all UD 2.8 languages with at least 10,000 available words, plus Xibe (new in UD 2.9, published after the other experiments were finished). We however excluded corpora of code-switched text (Hindi English and Turkish German). Table S1 shows the corpus sizes for the included UD languages. Table S3 shows excluded treebanks from languages otherwise included.

The hillclimbing algorithm is computationally very costly when corpora are very large. We thus had to focus on subcorpora for three languages: we focused on German-GSD (292K words) for German, Japanese-GSD (193K words) for Japanese, and Czech-PDT (1,509 words) for Czech. We used all available corpora for the gradient descent method.

S5 Historical Languages

Table S5 shows the historical languages in our dataset, with approximate dating assigned.

Language	Number of Sentences	Nunber of Words
Afrikaans	1,934	49,260
Akkadian	2,008	25,434
Amharic	1,074	10,010
Ancient Greek	30,999	416,988
Arabic	28,402	1,042,024
Armenian	2,502	52,630
Bambara	1,026	13,823
Basque	8,993	121,443
Belarusian	25,231	305,099
Breton	888	10,054
Bulgarian	11,138	156,149
Buryat	927	10,185
Cantonese	1,004	13,918
Catalan	16,678	546,638
Chinese	11,998	277,871
Classical Chinese	55,514	269,002
Coptic	1,873	48,632
Croatian	9,010	199,409
Czech	127,507	2,223,222
Danish	5,512	100,733
Dutch	20,944	306,720
English	33,251	570,631
Erzya	1,690	17,147
Estonian	36,508	506,637
Faroese	2,829	50,486
Finnish	36,981	397,001
French	42,832	1,132,460
Galician	4,993	164,385
German	208,440	3,753,947
Gothic	5,401	55,336
Greek	2,521	63,441
Hebrew	6,216	161,411
Hindi	17,647	375,533
Hungarian	1,800	42,032
Icelandic	51,957	1,162,040
Indonesian	7,623	168,286
Irish	5,776	131,423
Italian	35,879	818,562
Japanese	67,031	1,490,840
Kazakh	1,078	10,536
Kiche	1,435	10,013
Komi Zyrian	872	10,321
Korean	34,702	446,996
Kurmanji	754	10,260
Latin	22,405	284,794
Latvian	15,351	252,334
Lithuanian	3,905	75,403
Maltese	2,074	44,162
Manx	2,319	20,630
Mbya Guarani	1,144	13,089
Naija	9,242	140,859
North Sami	3,122	26,845
Norwegian	42,869	666,984
Old Church Slavonic	6,338	57,563
Old East Slavic	17,901	180,110
Old French	17,678	170,740
Persian	35,104	654,696
Polish	40,398	499,392
Portuguese	22,442	571,085
Romanian	40,480	937,540
Russian	85,789	1,420,647
Sanskrit	4,227	28,960
Scottish Gaelic	3,798	72,422
Serbian	4,384	97,673
Slovak	10,604	106,097
Slovenian	11,188	170,158
Spanish	34,693	1,015,119
Swedish	12,269	206,856
Tamil	1,134	12,165
Thai	1,000	22,322
Turkish	72,151	628,938
Ukrainian	7,060	122,091
Upper Sorbian	646	11,196
Urdu	5,130	138,077
Uyghur	3,456	40,236
Vietnamese	3,000	43,754
Welsh	1,833	36,837
Western Armenian	1,780	35,926
Wolof	2,107	44,258
Xibe (UD 2.9)	810	15,401

Table S1: Corpus sizes of the included UD languages. Experiments used UD 2.8, except in Xibe (UD 2.9), which was published after the other experiments were finished.

Treebank	Rationale
Chinese-CFL	Text written by non-native speakers
English-ESL	Text written by non-native speakers
English-Pronouns	Specifically targets pronouns
French-FQB	Consists entirely of questions
Latin-ITTB	Consists of Medieval Latin text
Latin-LLCT	Consists of Medieval Latin text

Table S3: UD corpora excluded, from languages otherwise included.

Language	Time	Rationale
Classical Chinese	300 BC	Life of Mengzi (died around 300 BC); the treebank contains his teachings as collected by his followers.
Ancient Greek	400 BC	Approximate mean age of texts used
Coptic	400 AD	Dating of the Apophthegmata Patrum texts used in the UD treebank
Gothic	350 AD	Life of bible translator Ulfilas (311–383)
Latin	0 AD	Approximate mean age of texts used
Medieval Spanish	1400 AD	Approximate mean age of texts used (not from Universal Dependencies, see Section S25).
Medieval Portuguese	1400 AD	Approximate mean age of texts used (not from Universal Dependencies, see Section S25).
Old Church Slavonic	850 AD	Bible translation after invention of Glagolitic alphabet around 850 AD.
Old English	900 AD	Approximate mean age of texts used (not from Universal Dependencies, see Section S25).
Old East Slavic	1200 AD	Approximate mean age of texts used
Old French	1200 AD	Approximate mean age of texts used
Sanskrit	900 BC	Approximate mean age of texts used

Table S5: Historical languages in our dataset.

S6 Phylogenetic Tree

S6.1 Tree Topology

We obtained tree topologies from Glottolog [31]. We only retained interior nodes when more than one of their daughter nodes had languages in our dataset. The resulting tree topology is displayed in Figure S2.[3]

S6.2 Dating Inner Nodes

We labeled interior nodes for the time at which they split into descendants, using estimates based on historical evidence and the linguistic literature:

Group	Split	Source or Rationale
Afroasiatic	10,000 BC	Diakonoff [32]
Arabic	1,100 AD	Calibration from Holman et al. [33] based on end of Arabic domination of Malta.
Armenian	1,750 AD	Separate development of Eastern and Western standards [34, p. 1]
Balto-Slavic	1,400 BC	Gray and Atkinson [35]
Brythonic	500 AD	Migrations from Britain to Brittany [33]
Central-Semitic	2,450 BC	Kitchen et al. [36]
Common Turkic	700AD	Savelyev and Robbeets [37, p. 49] estimate Common Turkic to have split around 474 AD. However, in their model, Old Turkic split off around 650 AD, earlier than the languages in our dataset, with uncertainty about the time of split of the remaining Common Turkic languages. It should predate the earliest documentation of Karluk Middle Turkic after 900AD. We thus put the divergence of the other Common Turkic languages at 700AD.
Eastern Baltic	600 AD	Split between Latvian and Lithuanian [38, p. 209]
Finnic	800 AD	Maurits et al. [39, Section 4.1]
Germanic	250AD	[35]
Global Dutch	1,600 AD	Dutch colony in South Africa
Goidelic	950 AD	Migrations from Ireland to Scotland. Holman et al. [33], citing Jackson [40], calibrates the divergence between Irish and Scottish Gaelic to 950 AD.
Hindustani	1,800 AD	Standardization of Hindi and Urdu
Iberian Romance	1,000 AD	Expansion of Christian kingdoms in Iberia, earliest Iberian Romance texts
Icelandic-Faroese	1,400 AD	Sound shifts specific to Faroese
Indo-European	5,300 BC	Gray and Atkinson [35] (excluding Hittite and Tocharian, for which we have no corpus data).
Indo-Iranian	2,500 BC	Parpola [41, p. 138]
Insular Celtic	900BC	Gray and Atkinson [35] estimate 900BC.
Iranian	500 BC	Gray and Atkinson [35].
Italo-Western-Romance	500 AD	End of the Western Roman empire [33].
Macro-English	1900AD	In our dataset, this is the common ancestor of contemporary English and Naija (Nigerian Pidgin).

[3] Tree obtained with https://icytree.org/.

Niger-Congo	5000BC	Holman et al. [33] estimate an age of 6227 years, but the family has to be older than Atlantic-Congo, which they estimate at 6525 years. We thus place Niger-Congo at 5000BC.
North-Germanic	650 AD	Split of Old Norse into regional variants, such as assimilation of nasals to following stops in Western Norse in the 7th century [42, p. 1856, 1859]. Similarly Holman et al. [33] calibrates this to 900 AD.
Semitic	3,750 BC	Kitchen et al. [36]
Serbo-Croatian	1,900 AD	Standardization of Serbian and Croatian
Slavic	700AD	Gray and Atkinson [35]. Novotná and Blazek [38, p. 209] date the split of East Slavic to the 6th century, Holman et al. [33] calibrates it to 550AD.
South-Slavic	750 BC	Expansion of Slavic into Balkan. Postdates Slavic and antedates Old Church Slavonic (attested after 800AD)
Uralic	3,000 BC	Maurits et al. [39, Section 4.7], cf [41, p. 144] for references
West Iberian	1,100AD	Independence of Portugal
West-Germanic	500 AD	Migrations into Britain and southern central Europe
West-Scandinavian	1,100 AD	Sound shifts specific to Norwegian
West-Semitic	3,400 BC	Kitchen et al. [36]
West-Slavic	750 BC	Expansion of Slavic.
Western Romance	800 AD	Expansion of Christian kingdoms into Iberia
Western South Slavic	1,000 AD	Antedates earliest Slovenian and Serbo-Croatian texts

Part III
Phylogenetic Analyses

S7 Details for Phylogenetic Models

S7.1 Calculating the Likelihood

For completeness, we describe how to calculate the likelihood of a multidimensional Ornstein-Uhlenbeck model on phylogenetic trees [43, 44, 45]. As described in the Methods section, it is described by the following stochastic differential equation for the instantaneous change of the state $\xi_{L,t} \in \mathbb{R}^4$ of a language L at a given time t:

$$\mathrm{d}\xi_{L,t} = \Gamma \cdot (\xi_{L,t} - \mu)\,\mathrm{d}t + \sqrt{\Sigma}\,\mathrm{d}B_t \tag{15}$$

where $\mu \in \mathbb{R}^4$, Γ is non-degenerate, and $\Sigma \in \mathbb{R}^{4\times 4}$ is a covariance matrix, and B_t is multidimensional Brownian motion. In our model, Γ is diagonal with positive entries.

The conditional distribution of a future observation at time $t+\Delta$ given an earlier one at time t is given by the following equation [46, Theorem 3.3], [47], [48, p. 156, eq. 6.124]:

$$\xi_{L,t+\Delta}|\xi_{L,t} \sim N\left(\mu + e^{-\Delta\Gamma}(\xi_{L,t} - \mu),\ \Omega - e^{-\Delta\Gamma}\Omega e^{-\Delta\Gamma^T}\right) \tag{16}$$

where the matrix $\Omega \in \mathbb{R}^{4\times 4}$ is obtained as the solution of the equation [47, p. 110, eq. 4.4.51] [48, p. 156, eq. 6.126]:

$$\Gamma\Omega + \Omega\Gamma^T = \Sigma \tag{17}$$

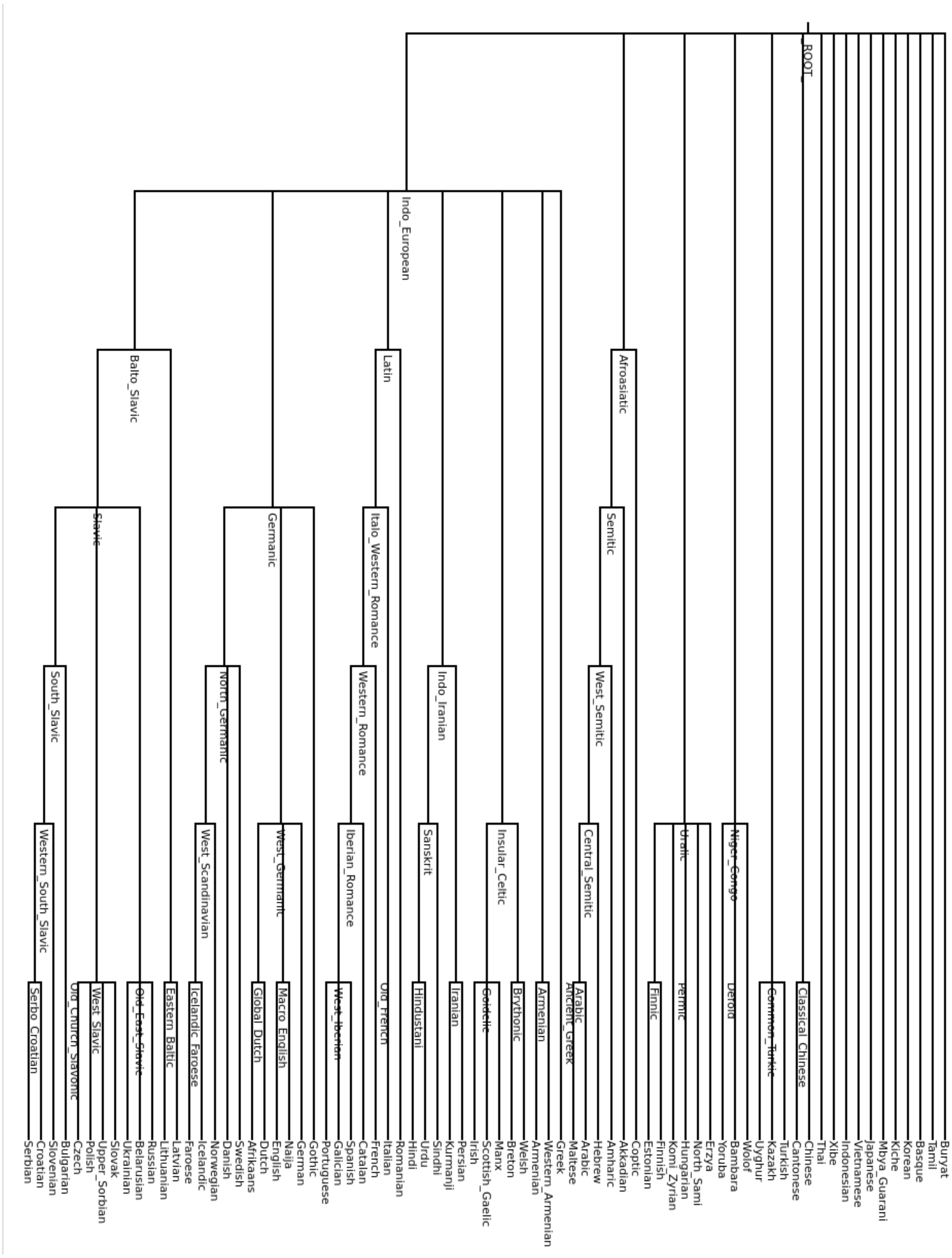


Figure S2: Phylogenetic tree topology of the languages in our sample. Compare Figure S3 for a version indicating the time depth of different families.

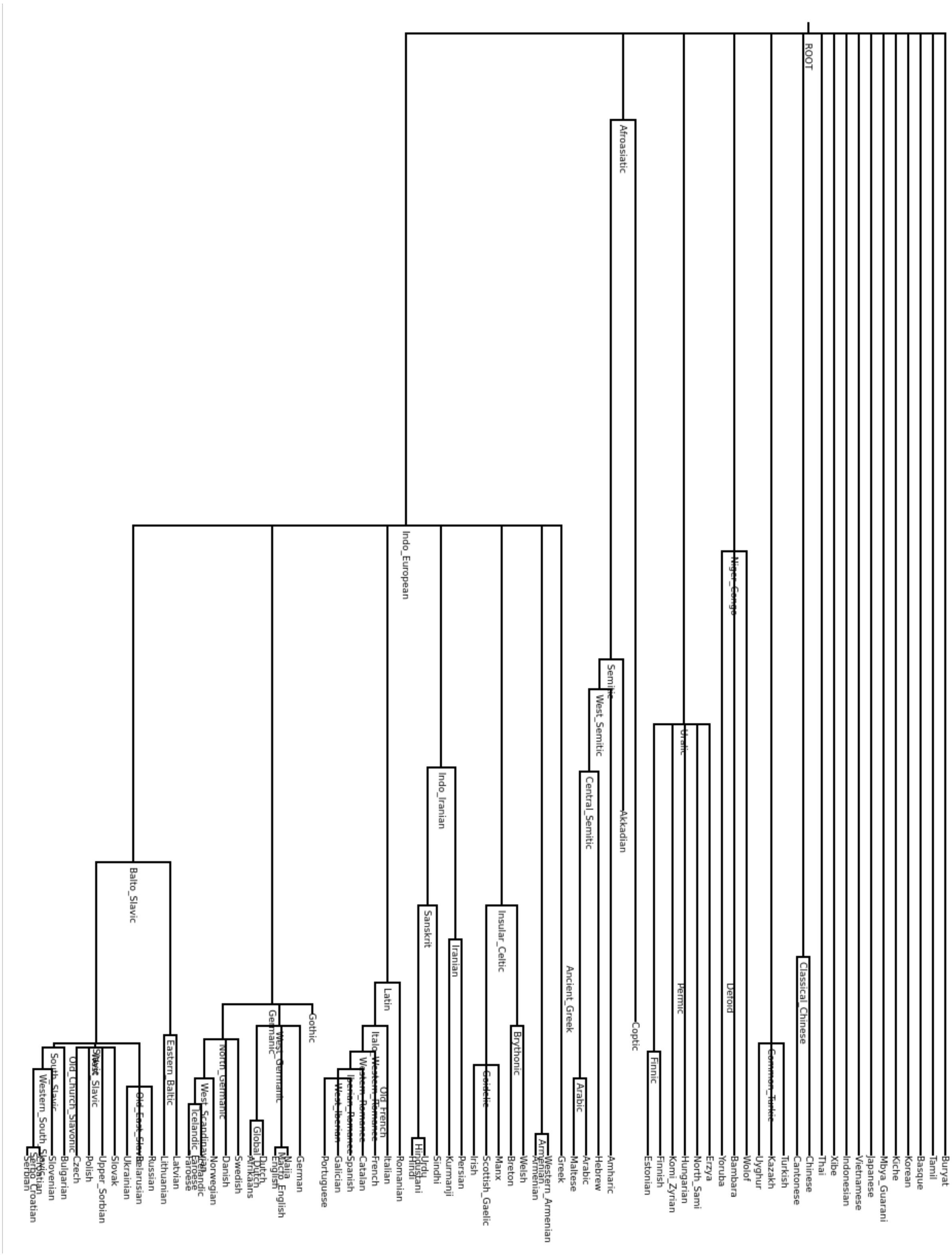


Figure S3: Phylogenetic tree of the languages in our sample. The length of branches reflects distance in time. Compare Figure S2 for a version indicating the raw topology without time depths.

This can be solved as follows (recall that Γ is diagonal in our model): [4]

$$\Omega_{ij} = \frac{\Sigma_{ij}}{\Gamma_{ii} + \Gamma_{jj}} \tag{19}$$

One can compute the stationary distribution that solves the differential equation as follows. The stationary distribution of an individual observation is

$$\xi_t \sim N(\mu, \Omega) \tag{20}$$

The stationary cross-covariance between the states of two languages L_1, L_2, possibly on different branches of the phylogenetic tree, is given by

$$Cov(\xi_{L_1}, \xi_{L_2}) = e^{-\Delta_1 \Gamma} \Omega e^{-\Delta_2 \Gamma^T} \tag{21}$$

where Δ_1, Δ_2 are the times of evolution from their last common ancestor to L_1 and L_2, respectively. [5] If L_1, L_2 do not share a common ancestor (the root in Figure S2 does not count as an ancestor), the covariance is zero.[6]

Since any Ornstein-Uhlenbeck process is Gaussian [46], the joint distribution of any set of observations $\xi_{L,t}$ is determined by (20-21).

S7.2 Implementation

We defined the following priors on the parameters. We parameterized Σ as the combination of a correlation matrix and a vector of standard deviations [49]. Stated differently, we parameterized $\Lambda := \sqrt{\Sigma}$ as DU, where U is a lower-diagonal matrix, and D is a diagonal matrix. We directly parameterized Γ using its diagonal entries.

To define a prior over Σ, we modeled U as the lower Cholesky factor of a correlation matrix subject to an LKJ(1) prior (Lewandowski et al. [50], i.e., the uniform distribution over 4×4-correlation matrices). We placed a standard normal prior $N(0,1)$ on the entries of μ and on the non-zero entries of D and Γ.

We rescaled times so that 1000 years corresponded to one unit. We further rescaled the four components to range from -1 to 1 (instead of [-1,0] for IL/DL, and [0,1] for position congruence).

We implemented the models in Stan [51] and obtained posterior samples using the No-U-Turn sampler. We ran four chains with 10,000 iterations each, of which the first half each were discarded as warmup samples.

The model can be implemented either using the analytical formula for the cross-covariance (21), or by explicitly modeling the inner nodes of the tree using (16). The first approach makes posterior inference more efficient, but we specifically used the second approach in the analysis where μ, Γ depended

[4] If Γ is not diagonal, and ξ has two dimensions:

$$\begin{pmatrix} \Omega_{11} \\ \Omega_{12} \\ \Omega_{22} \end{pmatrix} = \begin{pmatrix} 2\Gamma_{11} & 2\Gamma_{12} & 0 \\ \Gamma_{21} & \Gamma_{11} + \Gamma_{22} & \Gamma_{12} \\ 0 & 2\Gamma_{21} & 2\Gamma_{22} \end{pmatrix}^{-1} \begin{pmatrix} \Sigma_{11} \\ \Sigma_{12} \\ \Sigma_{22} \end{pmatrix} \tag{18}$$

[5] This can be shown as follows: If ξ_A is the last common ancestor, then (we set $\mu = 0$ without loss of generality, as it does not affect the covariance):

$$\begin{aligned} Cov(\xi_{L_1}, \xi_{L_2}) &= \mathbb{E}\left[\xi_{L_1} \xi_{L_2}^T\right] - \mathbb{E}\xi_{L_1} \mathbb{E}\xi_{L_2}^T \quad = \mathbb{E}\left[\mathbb{E}\left[\xi_{L_1} \xi_{L_2}^T | \xi_A\right]\right] - 0 \cdot 0 = \mathbb{E}\left[\mathbb{E}\left[\xi_{L_1} | \xi_A\right] \mathbb{E}\left[\xi_{L_2}^T | \xi_A\right]\right] \quad = \mathbb{E}\left[e^{-\Delta_1 \Gamma} \xi_A \xi_A^T e^{-\Delta_2 \Gamma^T}\right] \\ &= e^{-\Delta_1 \Gamma} \Omega e^{-\Delta_2 \Gamma^T} \end{aligned}$$

[6] As $\lim_{\Delta \to \infty} e^{-\Delta B} = 0$, this is practically equivalent to assuming a very large time-depth of the last common ancestor, which would be the case under the assumption of macrofamilies with very large time depth.

on geography (Section S11) or case marking (Section S12), as the cross-covariance is hard to compute explicitly when parameters vary.

We computed marginal likelihoods using Stepping Stone Sampling [52] with $K = 10$ stones. We verified stability of the estimates by running the procedure ten times for each model, and averaging the obtained marginal likelihoods.

S7.3 Correlation Component of Σ

In the main analysis, we reported the correlation between two dimensions in the stationary distribution Ω. In some analyses, there are multiple stationary distributions (depending on geography in Section S11 and case marking in the main paper and Section S12). In these cases, we therefore report correlations for the instantaneous changes at any point in time: The matrix Σ indicates the variance-covariance structure of the instantaneous changes at any time t [53, 54]. The main quantity of interest is the correlation between changes in two dimensions (e.g., attested and average optimized subject-object position congruence) [cf. 53, 54], which is given by

$$R_{ij} := \frac{\Sigma_{i,j}}{\sqrt{\Sigma_{i,i}\Sigma_{j,j}}} \tag{22}$$

A positive value indicates that changes in both directions are positively correlated.

S8 Detailed Results for Phylogenetic Model

Figures S4 and S5 visualize the stationary distribution, both for all languages and when excluding the Indo-European phylum.

Further Model Versions We also considered a version of the model where we explicitly accounted for imprecise measurements due to limitations in corpus data by assuming Gaussian observation noise, i.e., observations are modeled as $\widehat{\xi_L} = \xi_L + \varepsilon$, with $\varepsilon_i \sim \mathcal{N}(0, \sigma_i)$. [7] While the added model complexity did not improve model fit[8], it did not alter the conclusions: Indeed, with this model, the correlations were estimated to be even somewhat larger than without assuming observation noise ($R = -0.54$, 95% CrI $[-0.79, -0.30]$, $P(R > 0) < 0.0001$ for the correlation between DL and congruence; $R = 0.61$, 95% CrI $[0.32, 0.87]$, $P(R < 0) < 0.0001$ for the correlation between attested and average congruence). We also conducted a version of the model where the noise in different dimensions was allowed to be correlated, $\varepsilon \sim N(0, T)$ where T has the same prior as the instantaneous covariance matrix Σ (see Section S7.2). This is a particularly conservative model, because it allows the correlation in the noise to potentially explain some of the observed correlations. Nonetheless, correlations continued to be estimated similarly to before ($R = 0.42$, 95% CrI $[0.10, 0.72]$, $P(R < 0) = 0.00815$ for attested and optimized subject-object position congruence; $R = -0.52$, 95% CrI $[-0.73, -0.26]$, $P(R > 0) = 0.00005$ for DL and attested subject-object position congruence).

S9 Details for Model and Random Mutations

We sampled 40 random grammars and 40 approximately optimized grammars, each from one of the 80 languages. For each grammar, we ran 30 chains, either of 200 random mutations, or of $\approx$ 200 years of

[7] In terms of implementation, this leads to the addition of a diagonal matrix $diag([\sigma_1, \ldots, \sigma_4])$ to $Cov(\xi_L, \xi_L)$, and does not affect the other terms of the covariance.

[8] The marginal log-likelihood for a model applied to optimized and attested subject-object position congruence without noise is -94; the analogous model with noise has a less favorable marginal likelihood of -96.

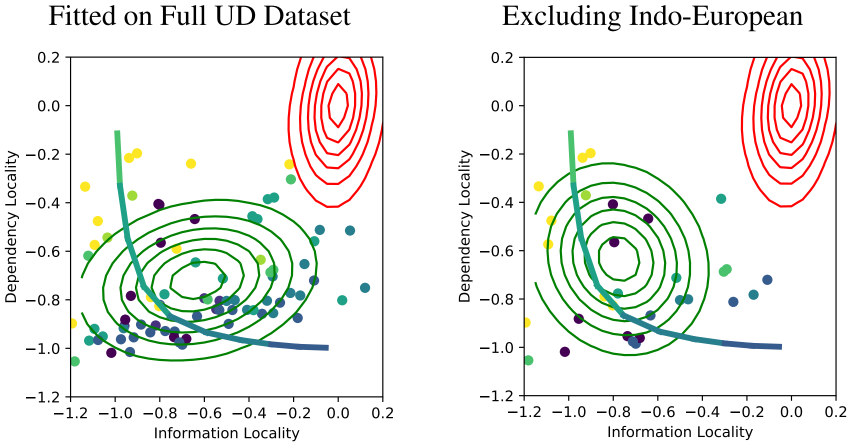


Figure S4: Stationary distribution in the plane spanned by optimized and attested subject-object position congruence; this indicates the region in which languages tend to move over the course of long-term evolution. The left column shows results on the entire dataset, the right column shows results excluding the Indo-European family.

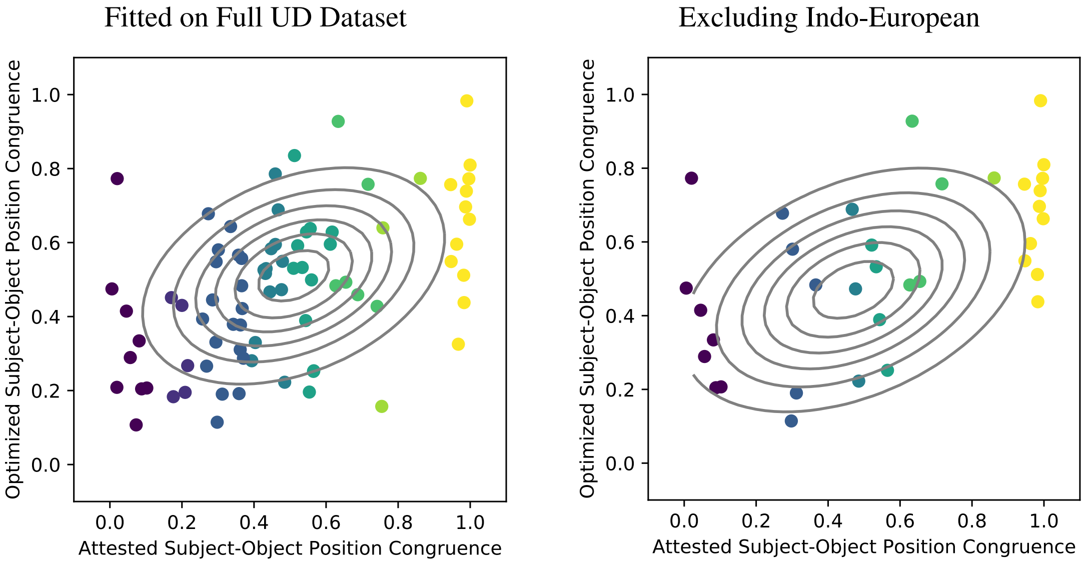


Figure S5: Stationary distribution in the plane spanned by optimized and attested subject-object position congruence.

evolution under the fitted model. In Main Paper, Figure 5, we show for each sampled grammar, an arrow from the original point to the mean position at the end of the 30 chains.

For an optimized grammar, random mutations usually deteriorated efficiency: chains of 200 mutations had a 12% chance of improving IL, and 17% of improving DL. In contrast, under the fitted model, change was neutral: 55% of chains improved IL, 51% of chains improved DL. For a baseline grammar, the pattern was in the opposite directions, random mutations were mostly neutral (39 % chance of improving IL, and 50% chance of improving DL). In contrast, under the fitted model, 71% of chains improved IL, 81% of chains improved DL.

S10 Comparison with Simple Brownian Model

The simple Brownian model leaves out the drift term, leading to the stochastic differential equation:

$$d\xi_t = \sqrt{\Sigma} dB_t$$

This is known as the Independent Contrasts model [53, 54], and underlies standard phylogenetic regression models [55].

Brownian motion differs from the Ornstein-Uhlenbeck process in that it does not have a long-term stationary solution. Instead, trajectories ξ_t tend to move arbitrarily far away from the origin over time t. This is clearly unrealistic in our setting, as subject-object position congruence is bounded between 0 and 1. As there is no stationary solution, there is no straightforward way to jointly apply the model to data from languages that do not share a common ancestor. For modelling purposes, we assumed that all families had a common ancestor at some large time T_0 in the past. This modelling assumption corresponds to the assumption of macro-families of very large time-depth. We considered T_0 to be 15,000 BC, 20,000 BC, 50,000 BC, and measured the instantaneous correlation of changes R for each fit.[9] To evaluate model fit, we compared marginal likelihood of the Brownian model with the Ornstein-Uhlenbeck model, computed using using Stepping Stone Sampling [52] with $K = 10$ stones. Note that an assumption of a specific time depth is not necessary for the Ornstein-Uhlenbeck model, as unrelated languages can be modeled as draws from the stationary distribution for that model.

In the absence of a stationary distribution, the Brownian model cannot make statements about whether languages evolve to maintain efficiency. We therefore only applied this to the attested and optimized subject-object position congruence, not to IL/DL.

Results Model fit as measured by marginal likelihood is much weaker than in the Ornstein-Uhlenbeck model, across choices of T_0 (Table S7), corresponding to a Bayes factor of about 10^{24} in favor of the Ornstein-Uhlenbeck model. Nonetheless, across different choices of T_0, the Brownian model strongly supports a positive correlation R between attested and optimized subject-object position congruence, very similar to the Ornstein-Uhlenbeck analysis; the posterior probability of $R \leq 0$ is 0.00425 at the best-fitting setup, $T_0 = -15,000$.

S11 Accounting for Areal Convergence

The model of random walks on phylogenetic trees assume that languages evolve independently once they have split [e.g. 56, 57]. However, linguistic evolution can include borrowing between geographically neighboring languages [e.g. 58, 59, 60, 61, 62]. Fully integrating such borrowing within phylogenetic modeling is an open problem for computational modeling. Here, we describe a possible modeling approach that explicitly models convergence in linguistic areas, geographic regions in which languages

[9] The maximum possible T_0 can be no later than Proto-Afroasiatic, which we calibrated at 10,000 BC (see Section S6.2).

Model	Log-Likelihood
Ornstein-Uhlenbeck	-94
Lesioned Ornstein-Uhlenbeck (No Coadaptation)	-119
Brownian ($T_0 = -100,000$)	-140
Brownian ($T_0 = -50,000$)	-129
Brownian ($T_0 = -20,000$)	-117
Brownian ($T_0 = -15,000$)	-114

Table S7: Marginal log-likelihoods for Ornstein-Uhlenbeck and simple Brownian models. Values closer to 0 indicate better model fit. We ran the Brownian model at different time depths, because it does not have a stationary distribution, necessitating the assumption of a common root node. The lesioned Ornstein-Uhlenbeck model without coadaptation is obtained by constraining the matrix Σ (Equation 15) to be diagonal.

tend to show convergent evolution due to borrowing [e.g. 63, 64, 65, 66], with a proof-of-concept implementation. We note that there may be other possible approaches, and have to leave a complete investigation of models fully integrating both phylogeny and borrowing to future research.

We propose to model linguistic areas as latent variables defining time- and location-dependent values $\mu(x,t)$ (where x is a point on the surface of the earth and t is a point in time) that languages at time t and place x drift towards. These values are inferred from the data together with the other parameters of the Ornstein-Uhlenbeck process. By placing a suitable Gaussian process prior on $\mu(x,t)$, we encourage parameters that smoothly vary over space and time, reflecting the idea that areal convergence between languages depends on their geographic distance. This approach is related to the model described by [67], who propose to model convergence between species by assuming correlations between the means μ of different species. For other approaches to model interactions between species from the bioinformatics literature, see [68, 69, 70, 71].

Model We model the grammar and usage components of μ as depending on the language's geographic position and the time a language was spoken. This models the impact of linguistic areas, and allows this impact to change over time.

We assume that a language L observed at time $t+\Delta$ (e.g. French) developed from a prior state at time t (e.g., Old French) during time $[t, t+\Delta]$ according to the Ornstein-Uhlenbeck SDE

$$\mathrm{d}\xi_{L,t} = \Gamma \cdot (\xi_{L,t} - \mu_L)\,\mathrm{d}t + \sqrt{\Sigma}\,\mathrm{d}B_t \tag{23}$$

where μ_L is defined by the temporal and geographical location of the language L.

We placed a Gaussian process prior with a Laplace kernel on μ. That is, the covariance between μ at points x, y on the surface of the earth at times T_1, T_2 is taken to be

$$Cov(\mu_x, \mu_y) = \alpha \cdot \exp\left(-\frac{1}{\rho_1^2} d(x,y) - \frac{1}{\rho_2^2}|T_1 - T_2|\right) \tag{24}$$

where $d(x,y)$ is the great circle (geodesic) distance between points x, y, and $\alpha, \rho > 0$ are hyperparameters. The Laplace kernel is positive-definite with the great-circle distance $d(\cdot,\cdot)$ [72] and thus provides a valid covariance for this distance; many other popular kernels like the RBF kernel are not valid for this distance [72]. This prior favors values of μ_L that vary smoothly over space and time, encoding the idea of linguistic areas. We placed Gaussian priors with mean 0 and variance 1, truncated to positive values, on the hyperparameters α, $\frac{1}{\rho^2}$ of the kernel (24).

We extracted locations of languages from the World Atlas of Linguistic Structures [73]. For ancestors, we recursively defined their location as the mean of the locations of their immediate children.

Due to substantial computational cost of this model, we applied it only to the main correlation of interest, i.e., the correlation between attested and average optimized subject-object position congruence. As convergence is slow compared to our other models, we ran MCMC for 40,000 iterations, again discarding the first half as warmup samples. We used the $\widehat{R}$ statistic and visual inspection of chains to assess model convergence.

Results As the mean μ depends on the geographic position, there is no single stationary distribution. As described in Section S7.3, we thus instead consider the correlation component of Σ, the covariance matrix of instantaneous changes. The correlation between changes in attested and average optimized subject-object position congruence was estimated at $R = 0.44$, 95% CrI $[0.2, 0.65]$, $P(R < 0) = 0.0004$, suggesting that coadaptation is found even when accounting for areal convergence in addition to phylogenetic relations.

S12 The Role of Case Marking

Here, we report details on the analysis of coadaptation when controlling for the presence of case marking. We do this by fitting an extension of the model that can model different directions of change in languages with and without case marking, and checking whether the analysis continues to provide evidence for coevolution between word order and usage *even beyond* what is captured by correlations of usage and word order with the presence of case marking.

Coding Languages for Case Marking We coded languages from our sample for the presence or absence of case marking on the basis of Iggesen [74], supplemented with information from the grammatical literature where no information was provided. We amended the annotation from Iggesen [74] to include only case marking that distinguishes between subjects and objects; this concerns several modern Celtic and Germanic languages, which have some nominal case marking but do not distinguish subjects and objects (e.g., Swedish and English use *-s* to mark possessives, but do not distinguish nominal subjects and objects.).

We furthermore coded all interior nodes of the phylogenetic tree for case marking based on the linguistic literature. In some cases, this annotation was unambiguous due to available historical documentation even though no treebank data was available (e.g., Proto-West-Scandinavian was a late form of Old Norse and had case markers). In many other cases, cognate case markers are unambiguously attested both within and without a group, showing that they were present in the protolanguage (e.g., Proto-Germanic, Proto-Indo-Iranian). Furthermore, in many protolanguages, case markers are commonly reconstructed based on their presence in different descendant branches (e.g., Proto-Indo-European, Proto-Afroasiatic, Proto-Common-Turkic, Proto-Uralic and Proto-Ugric). Case is not unambiguously reconstructed for Proto-Niger-Congo; we verified that both possible parameter settings lead to qualitatively equivalent results (we report results under the assumption that it did not have case, with essentially indistinguishable results for the other cases).

Model of Change conditioned on Case Marking Based on the prior literature, we expect that languages without case marking will be biased towards low subject-object position congruence [75]. To take this into account, we modified the model by conditioning the mean vector μ on the presence or absence of case in the language L.

$$\mathrm{d}\xi_{L,t} = \Gamma_{C(L)} \cdot (\xi_{L,t} - \mu_{C(L)})\,\mathrm{d}t + \sqrt{\Sigma}\,\mathrm{d}B_t$$

where $C(L)$ is 1 if L has case and 0 else. We set priors $\mu_{C(L)} \sim N(0,1)$ for both $C(L) = 0$, and 1.

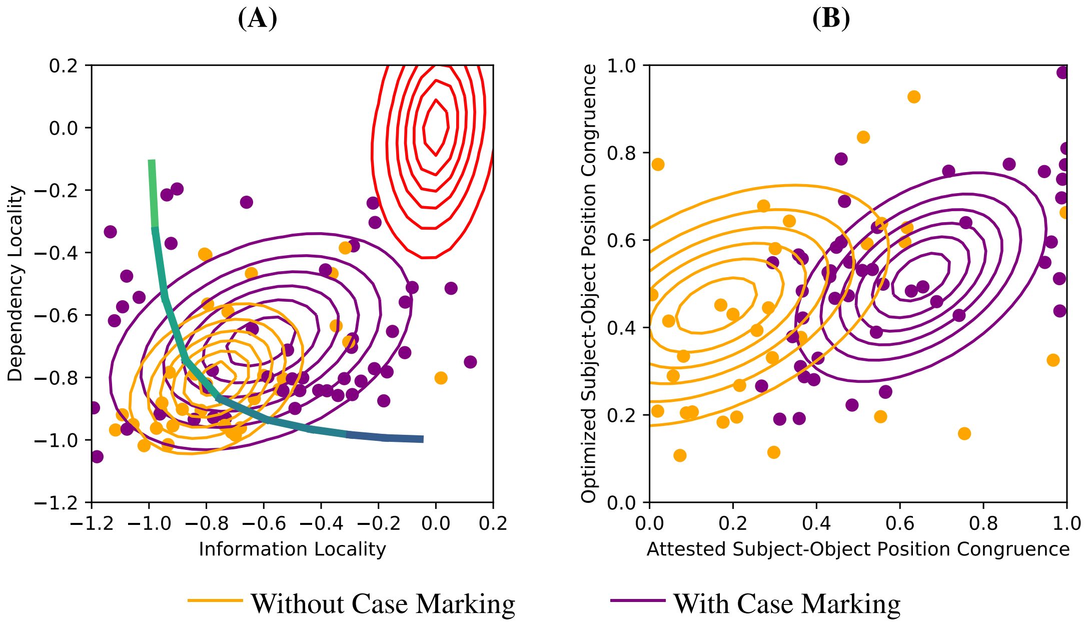


Figure S6: Fitted stationary distribution, conditioned on case marking. (A) Languages with and without case marking similarly concentrate in the region between the baseline distribution and the Pareto frontier. The difference between the mean values of IL and DL of the two stationary distributions is not statistically meaningful. (B) While the presence of case marking is associated with higher subject-object position congruence ($\mu = 0.65$, 95% CrI $[0.56, 0.76]$ with case, $\mu = 0.16$, 95% CrI $[-0.07, 0.36]$ without case), coadaptation is predicted even beyond this association, as evidenced by the shape of the two stationary distributions.

Results We plot the distribution of languages and the fitted stationary distributions, conditioned on $C(L)$, in Figure S6. In accordance with the prior literature, the model indicated that languages without case marking favor regions with low observed subject-object position congruence. For languages with case marking, there was evidence for a bias towards higher subject-object position congruence. We quantified correlations using the correlation component of the instantaneous changes Σ (see Section S7.3), i.e., the correlation between short-term stochastic changes in different dimensions. Similarly to the results in the main analysis, there was a negative correlation between DL and subject-object position congruence ($R = -0.39$, 95% CrI $[-0.60, -0.18]$, $P(R \geq 0) < 0.0001$), and a positive correlation between attested congruence and average optimized congruence along the Pareto frontier ($R = 0.49$, 95% CrI $[0.29, 0.67]$, $P(R \leq 0) < 0.0001$). This shows that languages show coadaptation between usage and grammar in basic word order, even beyond an association with case marking.

Conclusion We found that, while case marking has a robust impact on subject-object position congruence, coadaptation continues to hold when controlling for this.

Part IV
Additional Analyses

S13 Optimized and Baseline Grammars

In Figure S7, we show the position of optimized (orange) and baseline (blue) grammars for each of the 80 languages. Optimized grammars inhabit the area between the baseline grammars and the Pareto frontier. Compare Section S22 for results by subject-object position congruence.

S14 Within-Language Correlates of Basic Word Order

Here, we show that basic word order reflects optimization for Dependency Length Minimization (DLM) not only on the level of languages, but also on the level of individual sentences.

In many SVO languages, certain intransitive subjects can appear after the verb ("along came a dog"). This kind of "intransitive inversion" has been documented in many SVO languages, including English, Romance languages, and Chinese [76, Chapter 17.2]. There are also languages whose basic word order is different in transitive and in intransitive clauses [77]; the World Atlas of Language Structures lists 13 languages with transitive SVO and intransitive VS basic word order [78, 77], while it lists no languages with transitive VSO and intransitive SV order. This observation has been formalized as the following language universal: *If VS is dominant with transitives, it is also dominant with intransitives* (Plank and Filimonova [79, No 344], citing Kozinsky [80]). DLM provides an explanation for this universal.

We conjectured that, more generally, the rate of VS order is higher when no object is present than when an object is present. For each language in our dataset, we collected statistics for all verbs with a subject and conducted the following logistic analysis:

$$\text{SV Order} \sim \text{Object is present} \qquad (25)$$

A positive effect indicates that presence of an object makes SV order more likely, compared to VS order. Results are shown in the table below. As predicted, in most languages where there is variation between SV and VS order, a significant positive effect was observed.

Coefficients in logistic analysis regressing SV/VS Order based on the presence of an object. 'SV Frequency' indicates the overall rate of SV order (as opposed to VS) in the language. A positive coefficient ($\beta > 0$) indicates that SV is more common in the presence of an object than when there is no object.

Language	SV Frequency	β	p
Afrikaans	0.989	-0.12	0.6568
Akkadian	0.98	1.56	0.0449
Amharic	0.665	-0.38	0.0011
Ancient Greek	0.786	0.31	< 0.00001
Arabic	0.492	0.55	< 0.00001
Armenian	0.89	0.83	< 0.00001
Bambara	0.999	16.68	0.9948
Basque	0.872	-0.11	0.0811
Belarusian	0.773	1.24	< 0.00001
Breton	0.541	0.1	0.6465
Bulgarian	0.813	1.29	< 0.00001

Buryat	0.996	-0.42	0.731
Cantonese	0.994	1.12	0.3061
Catalan	0.932	0.07	0.0502
Chinese	0.999	17.7	0.985
Classical Chinese	0.999	17.51	0.9852
Coptic	0.922	18.18	0.9581
Croatian	0.827	0.99	< 0.00001
Czech	0.733	0.47	< 0.00001
Danish	0.865	0.62	< 0.00001
Dutch	0.813	0.42	< 0.00001
English	0.962	3.49	< 0.00001
Erzya	0.677	0.94	< 0.00001
Estonian	0.737	0.49	< 0.00001
Faroese	0.854	-0.16	0.0559
Finnish	0.867	1.23	< 0.00001
French	0.957	1.17	< 0.00001
Galician	0.877	1.09	< 0.00001
German	0.843	0.63	< 0.00001
Gothic	0.733	0.6	< 0.00001
Greek	0.839	0.65	< 0.00001
Hebrew	0.692	0.93	< 0.00001
Hindi	0.996	2.03	< 0.00001
Hungarian	0.81	0.77	< 0.00001
Icelandic	0.75	0.2	< 0.00001
Indonesian	0.943	4.54	< 0.00001
Irish	0.162	0.45	< 0.00001
Italian	0.821	1.76	< 0.00001
Japanese	1	15.4	0.9952
Kazakh	0.992	-0.36	0.6167
Kiche	0.474	0.97	< 0.00001
Komi Zyrian	0.762	1	3e-04
Korean	1	15.18	0.9953
Kurmanji	0.997	16.57	0.9948
Latin	0.833	0.56	< 0.00001
Latvian	0.79	0.76	< 0.00001
Lithuanian	0.785	0.33	9e-04
Maltese	0.731	2.34	< 0.00001
Manx	0.001	-16.59	0.9969
Mbya Guarani	0.866	1.73	0.0929
Naija	0.982	17.81	0.959
North Sami	0.799	1.91	< 0.00001
Norwegian	0.837	0.85	< 0.00001
Old Church Slavonic	0.686	0.76	< 0.00001
Old East Slavic	0.661	0.38	< 0.00001
Old French	0.861	0.78	< 0.00001
Persian	0.999	0.22	0.462
Polish	0.756	0.83	< 0.00001
Portuguese	0.909	2.14	< 0.00001
Romanian	0.74	0.58	< 0.00001

Russian	0.772	1.09	< 0.00001
Sanskrit	0.893	0.35	0.0091
Scottish Gaelic	0.013	-0.55	0.2019
Serbian	0.801	1.36	< 0.00001
Slovak	0.724	0.71	< 0.00001
Slovenian	0.778	0.37	< 0.00001
Spanish	0.849	1.28	< 0.00001
Swedish	0.865	0.72	< 0.00001
Tamil	0.987	1.71	0.1
Thai	0.999	17.45	0.9967
Turkish	0.972	-0.25	< 0.0001
Ukrainian	0.806	1.05	< 0.00001
Upper Sorbian	0.799	0.84	5e-04
Urdu	0.996	1.24	0.0038
Uyghur	0.961	2.69	< 0.00001
Vietnamese	0.989	0.76	0.0209
Welsh	0.053	1.21	0.0011
Western Armenian	0.915	0.81	< 0.0001
Wolof	0.999	16.7	0.9914

In some predominant VSO languages, SVO is an alternative word order in unembedded clauses, whereas embedded clauses tend to only allow VSO. This is in accordance with the predictions of DLM, which favors high subject-object position congruence in embedded clauses (see Figure 1B in the main paper). Examples include relative clauses in Afroasiatic and Celtic (Standard Arabic [81], Breton [82, p. 80], Ancient Egyptian [83], Tuareg [84, Chapter 12.1.2]). Conversely, in some SVO languages, embedded clauses show VSO order (Bantu, Demuth and Harford [85]); Miza (Chadic) has SVO/VOS in main clauses and VOS in embedded clauses [78]. However, it is not generally true that VS order is more common in embedded clauses across all languages that have variation in basic word order. For instance, German and Dutch can have VS in main clauses, but are almost always SV in subordinate clauses; the same holds for Quileute (Chimakuan) [78].

S15 Coexpression of Subjects and Objects

In Figure S8, we show attested subject-object congruence together with the fraction of verbs that simultaneously express a subject and an object among those verbs expressing at least one, for each language. In Figure S9, we compare this fraction with the average subject-object position congruence along the Pareto frontier.

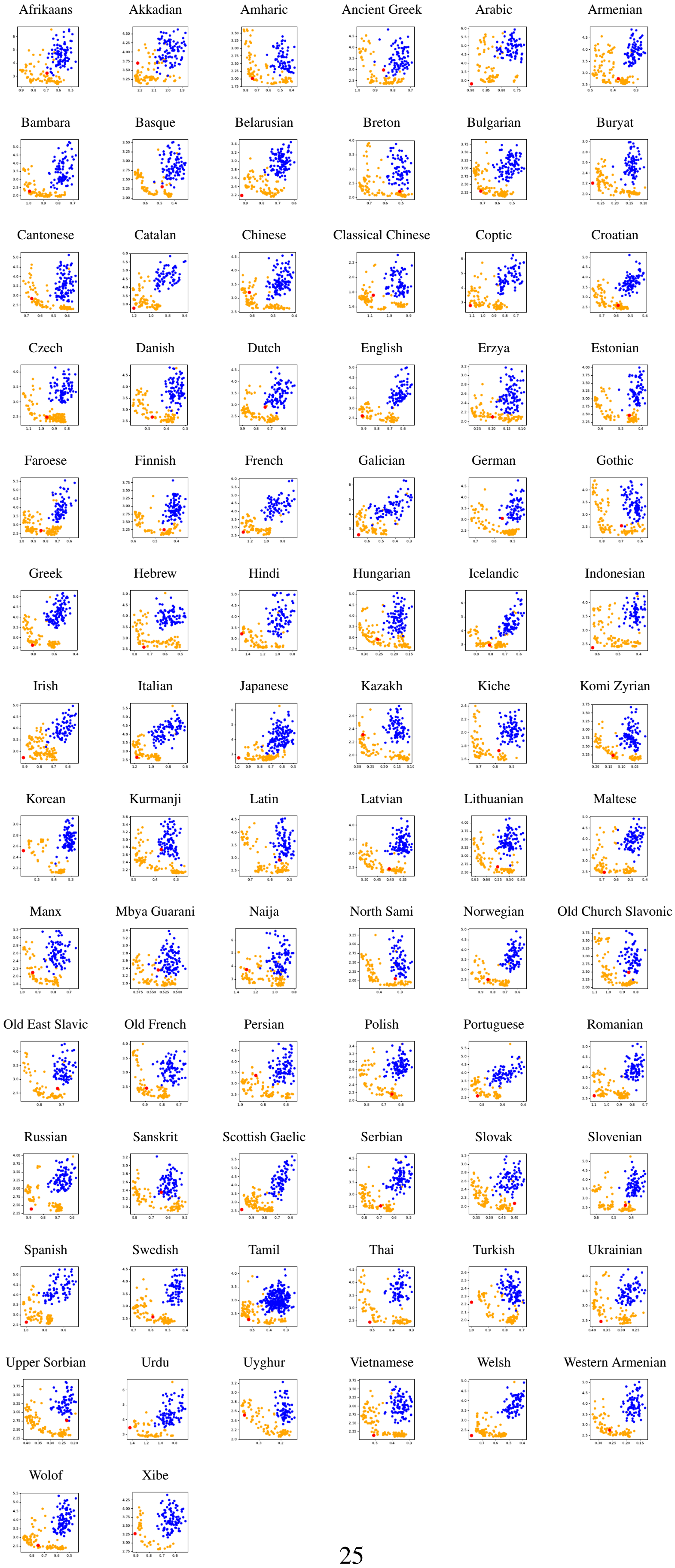
S16 Details for Mixed-Effects Analyses

Priors We conducted standard Bayesian linear mixed-effects regressions [86] where the response y_i belonging to language i is given by

$$y_i = (\alpha + \alpha_{f_i}) + (\beta + \beta_{f_i})x_i + \varepsilon_i \tag{26}$$

where x_i is the predictor (e.g., attested subject-object position congruence), f_i is the family of language i, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, and α_{f_i} and β_{f_i} are per-family adjustments to the intercept α and the slope β respectively.

As described in the main paper, we assumed the prior $N(0,1)$ for the fixed effects slopes, $N(0.5,1)$ for the intercepts, weakly informative Student's t priors ($\nu = 3$ degrees of freedom, location 0, and scale



Afrikaans, Akkadian, Amharic, Ancient Greek, Arabic, Armenian

Bambara, Basque, Belarusian, Breton, Bulgarian, Buryat

Cantonese, Catalan, Chinese, Classical Chinese, Coptic, Croatian

Czech, Danish, Dutch, English, Erzya, Estonian

Faroese, Finnish, French, Galician, German, Gothic

Greek, Hebrew, Hindi, Hungarian, Icelandic, Indonesian

Irish, Italian, Japanese, Kazakh, Kiche, Komi Zyrian

Korean, Kurmanji, Latin, Latvian, Lithuanian, Maltese

Manx, Mbya Guarani, Naija, North Sami, Norwegian, Old Church Slavonic

Old East Slavic, Old French, Persian, Polish, Portuguese, Romanian

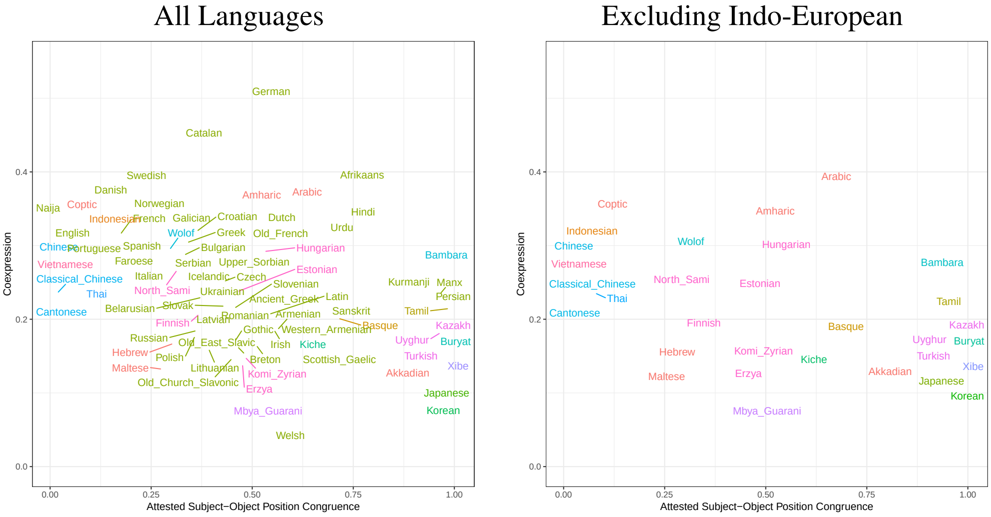
Russian, Sanskrit, Scottish Gaelic, Serbian, Slovak, Slovenian

Spanish, Swedish, Tamil, Thai, Turkish, Ukrainian

Upper Sorbian, Urdu, Uyghur, Vietnamese, Welsh, Western Armenian

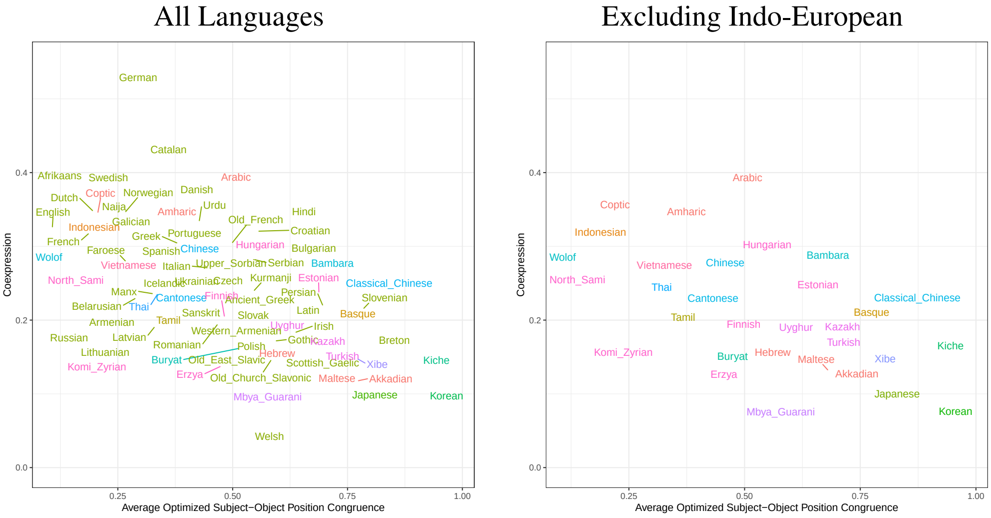
Wolof, Xibe

Figure S7: Position of optimized (orange) and baseline (blue) grammars for each of the 80 languages. Red dots indicate the attested ordering.



β = −0.11, 95% CrI [−0.20, −0.03]   β = −0.1, 95% CrI [−0.2, 0]

Figure S8: Comparison of attested subject-object position congruence (x-axis) and the fraction of verbs that simultaneously express a subject and an object, among those verbs expressing at least one (“coexpression”, y-axis). Attested subject-object position congruence predicts coexpression in a linear mixed-effects regression with per-family intercept and slope ($\beta = -0.11$, $SE = 0.04$, 95% CrI $[-0.20, -0.03]$, $P(\beta > 0) = 0.006$).



β = −0.23, 95% CrI [−0.33, −0.13]   β = −0.2, 95% CrI [−0.33, −0.07]

Figure S9: Comparison of average subject-object position congruence along the Pareto frontier (x-axis) and the fraction of verbs that simultaneously express a subject and an object, among those verbs expressing at least one (“coexpression”, y-axis). We show coefficients and Bayesian coefficients of determination in a linear mixed-effects regression with per-family intercept and slope.

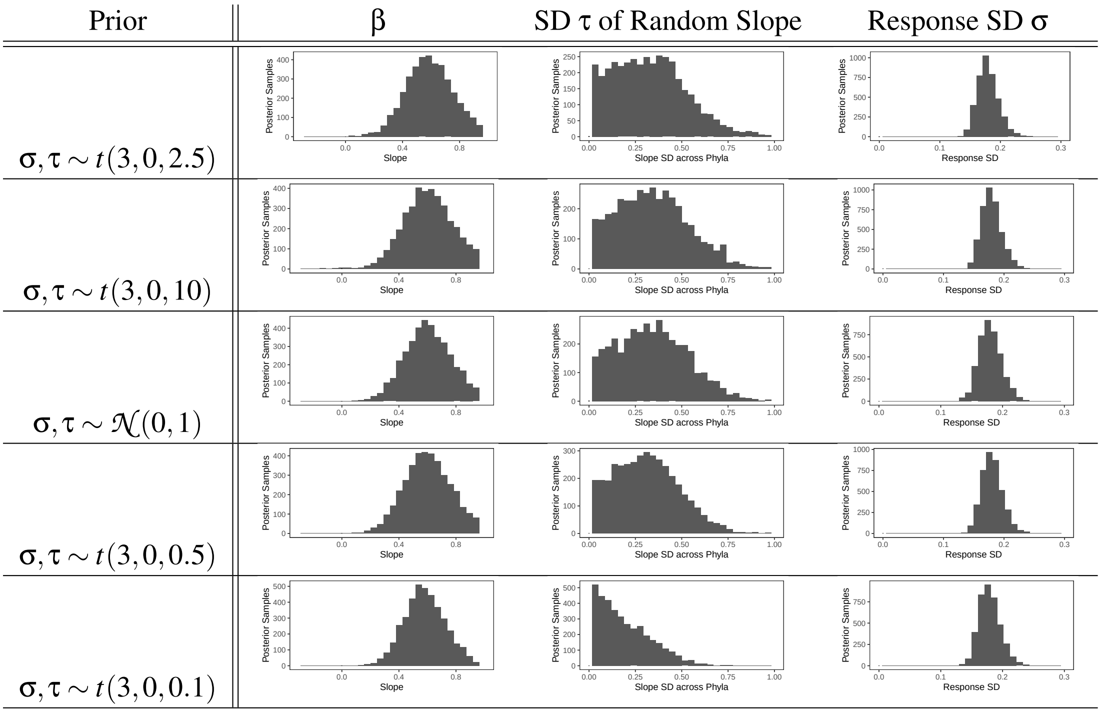


Figure S10: Impact of the prior for the variance terms on the posterior in the Bayesian mixed-effects analysis predicting attested subject-object position congruence from average congruence along the Pareto frontier. The first line corresponds to the prior used in our analysis; the other priors differ in the degree to which they regularize towards 0, from mild regularization (top) to very strong regularization (bottom). We write $t(\nu, \mu, \sigma)$ for the Student's t distribution with ν degrees of freedom, location μ, and scale σ. For each prior, we show the posterior of the coefficient β (the quantity of interest), the standard deviation of the slope across families, and the standard deviation of the Gaussian response. While changing the prior affects the estimated posterior of the slope variance across families, it has little effect on the estimate of β. This shows that the estimate of β is not impacted by a possible inflation of the variance components linked to the large number of isolated languages.

$\sigma = 2.5$) for the standard deviations of the residuals and the random effects, and an LKJ(1) prior [50] for the correlation matrix of random effects.

S16.1 Insensitivity to Priors

A potential concern is that, because our dataset includes many families represented by only one or a few languages, the mixed-effects model might suffer from inflated estimates of the variance components, as the slopes cannot be individually estimated for those families.

We repeated the analysis predicting attested subject-object position congruence from optimized subject-object position congruence with several more strongly regularizing priors on the variance components.

In Figure S10, we plot how the posteriors for β, the standard deviation τ of the per-family adjustments β_f, and the response standard deviation σ vary as a function of the prior. The priors for the fixed effect coefficient β ($N(0,1)$) and the intercept ($N(0.5,1)$) are as in the main analysis. Results show that, while more strongly regularizing priors shrink the estimated range of τ, they have limited impact on the posterior of the key quantity, β. Even an unrealistic extremely regularizing prior $t(3,0,0.1)$ does not change the posterior of β much.

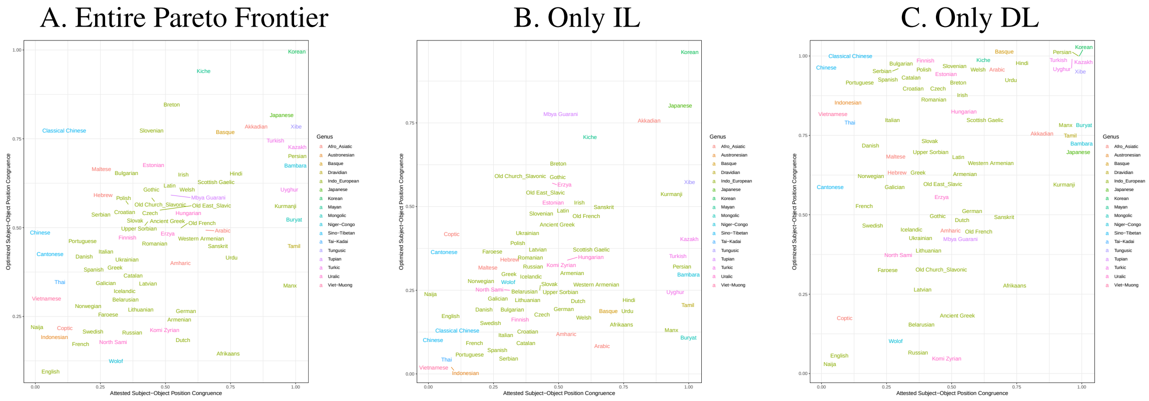


Figure S11: Attested and optimized subject-object position congruence (compare Figure 3 in the main paper), with language names, colored by the 17 families represented in the dataset. Compare Figure S12 for results from joint analysis in higher resolution.

S17 Further Visualizations for Coadaptation

See Figures S11–S13 for versions of Figure 4 in the main paper with language names.

We further investigated the robustness of the correlation between attested and average optimized subject-object position congruence to possible outliers. Correlations, in particular Pearson correlations, are vulnerable to outliers and points of high leverage. In order the evaluate whether this impacted the results, we considered all subsets of ≤ 3 languages, and recomputed the correlation when excluding this subset. The correlation was in the range $[0.42, 0.61]$ for all such subsets. This suggests that the correlation is not inflated due to individual points of high leverage.

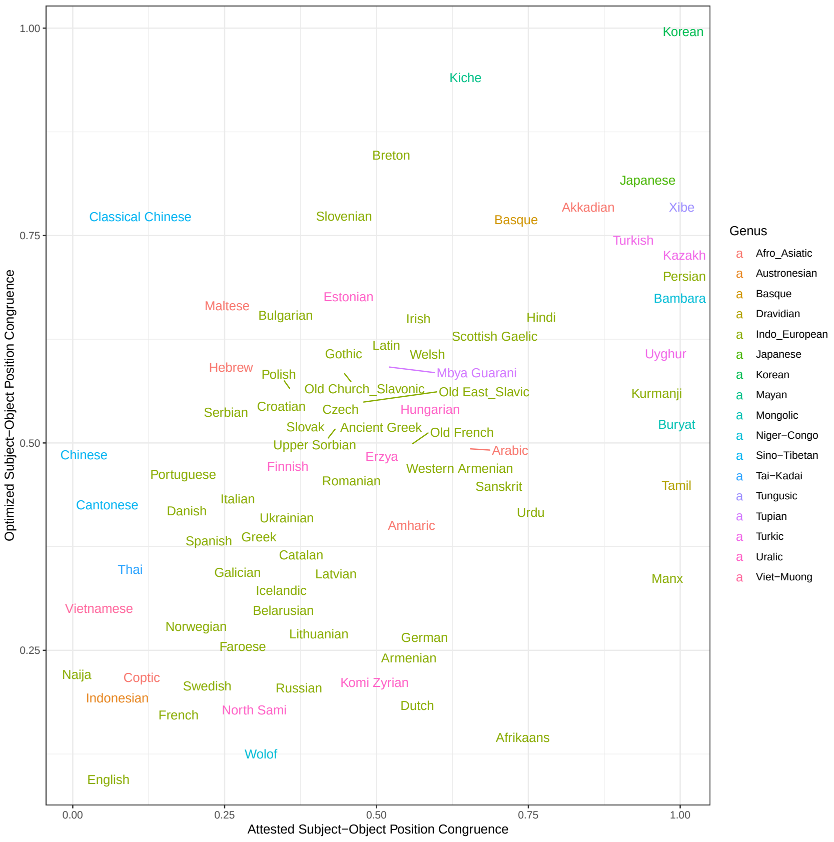


Figure S12: Attested and optimized subject-object position congruence (compare Figure 3 in the main paper), with language names, colored by the 17 families represented in the dataset. Compare Figure S11 for results optimizing only for DL or IL.

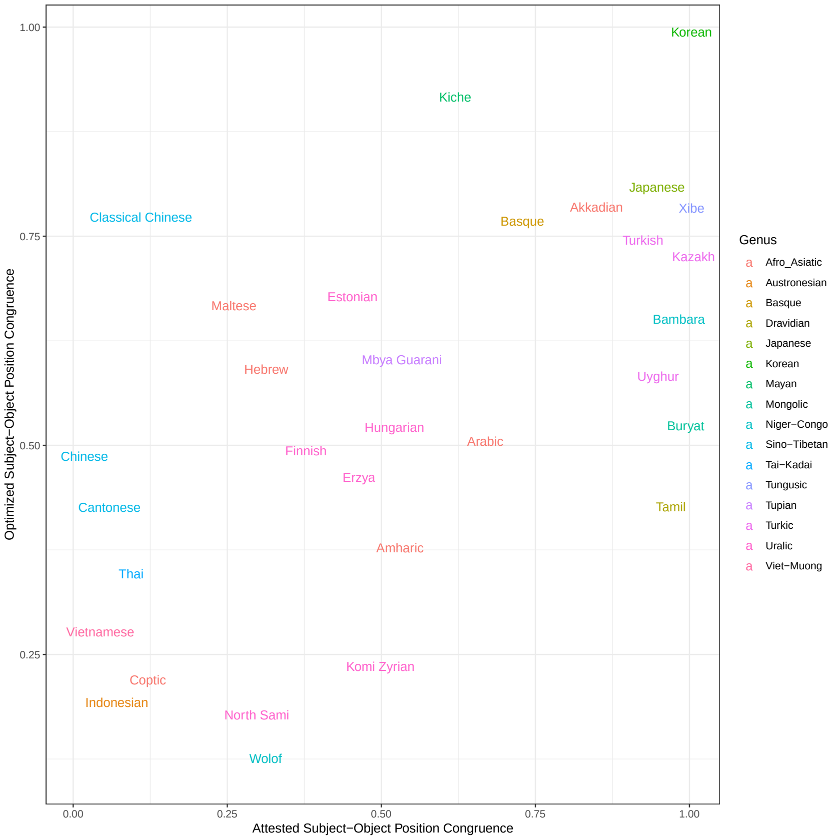
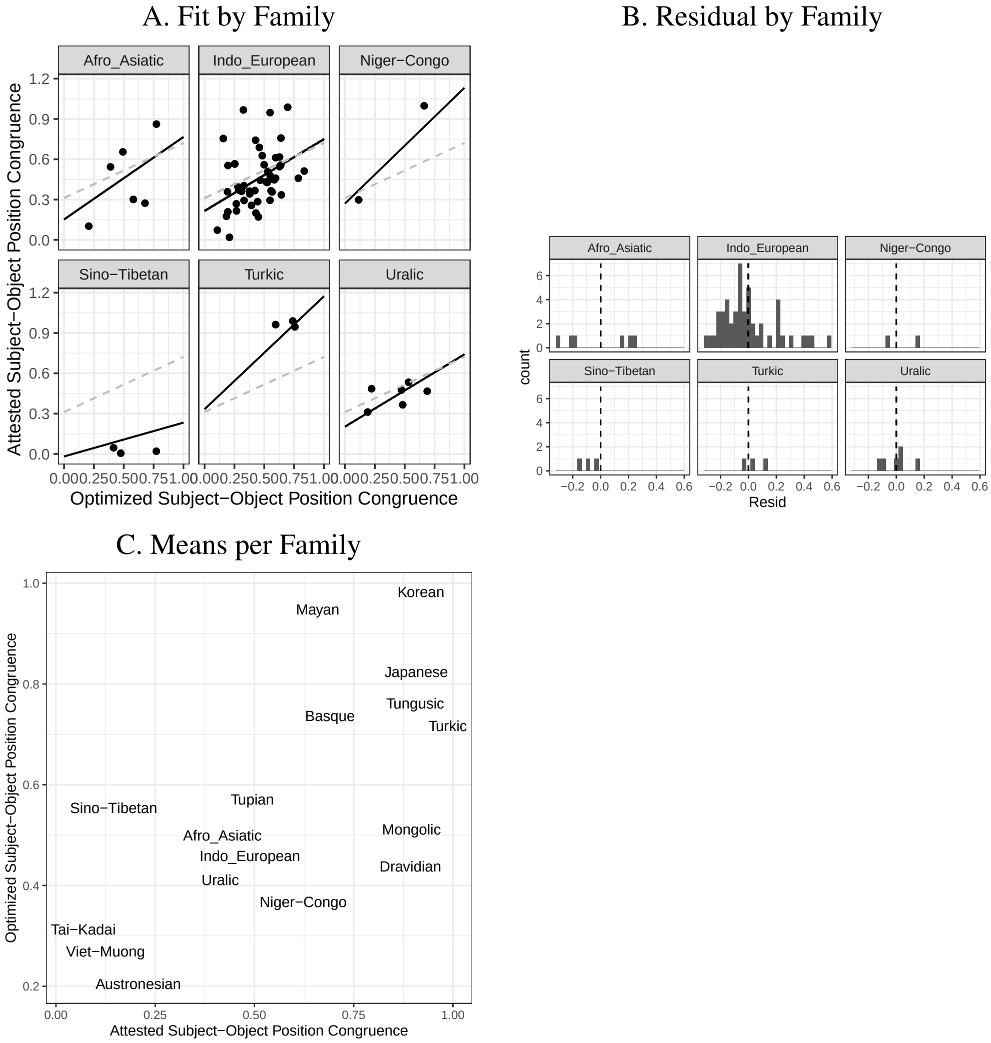


Figure S13: Attested and optimized subject-object position congruence (compare Figure 3 in the main paper), excluding the Indo-European languages. Compare Figure S12 for results on all 80 langiages.

S18 Per-Family Results and Fitted Slopes

Figure S14 shows results across the 17 families, including the six ones represented by at least two languages, for the analysis of optimized and attested subject-object position congruence. In Figure S15, we show the fitted slope $\beta + \beta_f$ (fixed effects slope β plus per-family adjustment β_f) for each family that has at least two languages.

We note that, while smaller families do not provide sufficient evidence for a positive relationship on their own, estimating the overall slope in a mixed-effects analysis does not require independent estimates of the slopes in each family. Instead, the mixed-effects regression obtains its slope estimate by combining (i) the data across isolates and smaller families, and (ii) the slope within the well-represented Indo-European family. Thus, for the purposes of the mixed-effects analyses, the presence of many families, even isolates and sparsely represented ones, can provide an advantage, because it increases the amount of statistical independence in the dataset.



A. Fit by Family

B. Residual by Family

C. Means per Family

Figure S14: A: Fit of the mixed-effects model across the 6 families represented by at least two languages in the Universal Dependencies dataset. We show the overall slope fitted by the mixed-effects analysis across the 80 languages as a dashed line, and the per-family adjusted slope as a solid line. In both cases, we use the posterior mean of intercepts and slopes. Note that, for less well represented families, the model has substantial uncertainty about the slope, not well represented by the point estimates, and even in families with seemingly divergent slope, the data are statistically consistent with the slope being in fact the same across families, see Figure S15. B: Residuals by family tend to be centered around zero. C: Means across all languages within each family. This illustrates that the per-family means also exhibit a positive correlation: That is, a positive correlation is supported both across families, and within the larger families individually.

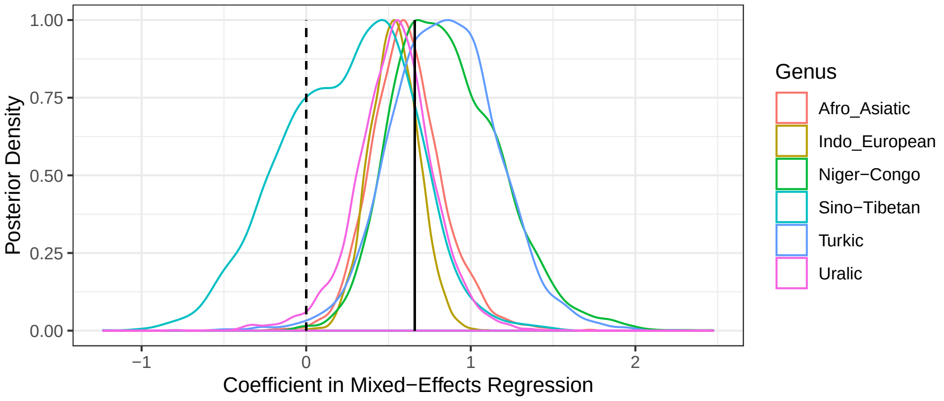


Figure S15: Posterior Densities (scaled so all are bounded by 1) for the slope in the linear mixed-effects regression in the six families with at least two languages. For poorly represented families, the posterior is wider. Nonetheless, across families, the model assigns almost all of the posterior probability mass to a positive sign, except in Sino-Tibetan, where the dependent variable has almost no variance. While the posterior mode differs between the families, the posterior is always well compatible with the overall estimated β, indicated by a solid vertical line.

S19 Detailed Results for Figure 3 in Main Paper

See Figure S17 for results per word order category, including less frequent categories “VSO” and “No dominant order”.

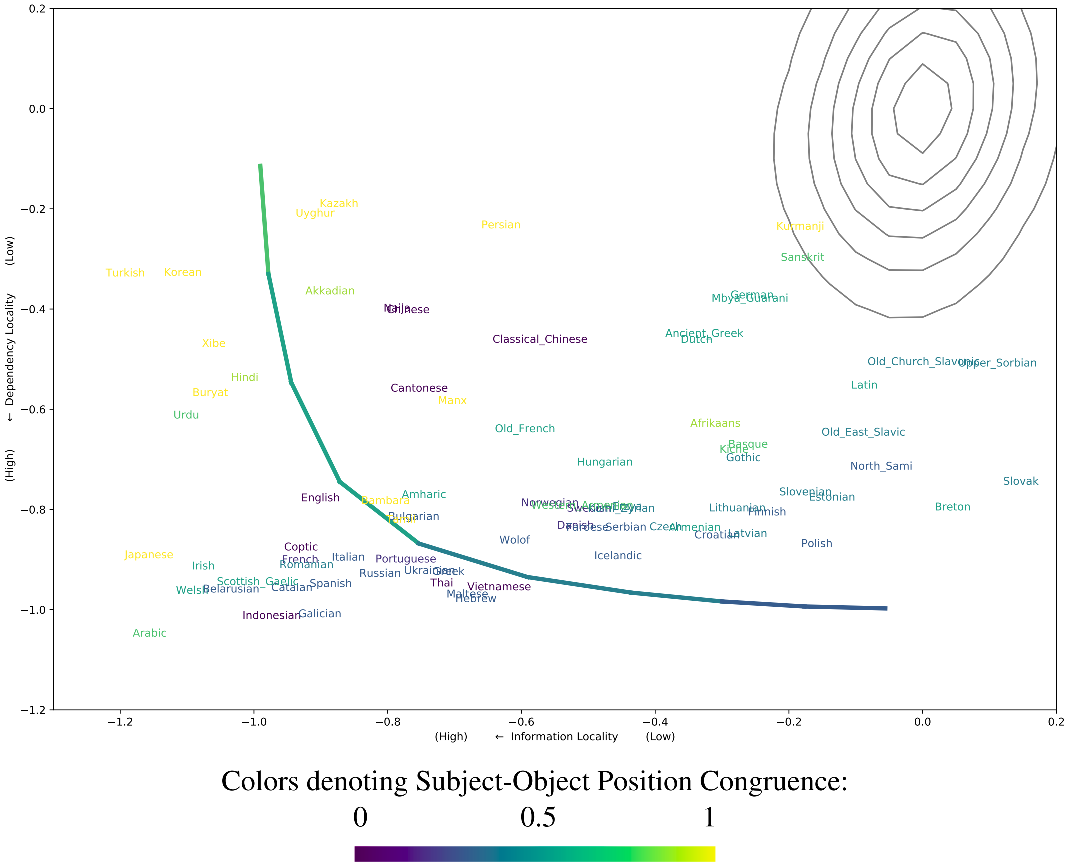


Figure S16: Position of the 80 languages in the efficiency plane with all language names. Compare Figure 3 in the main paper.

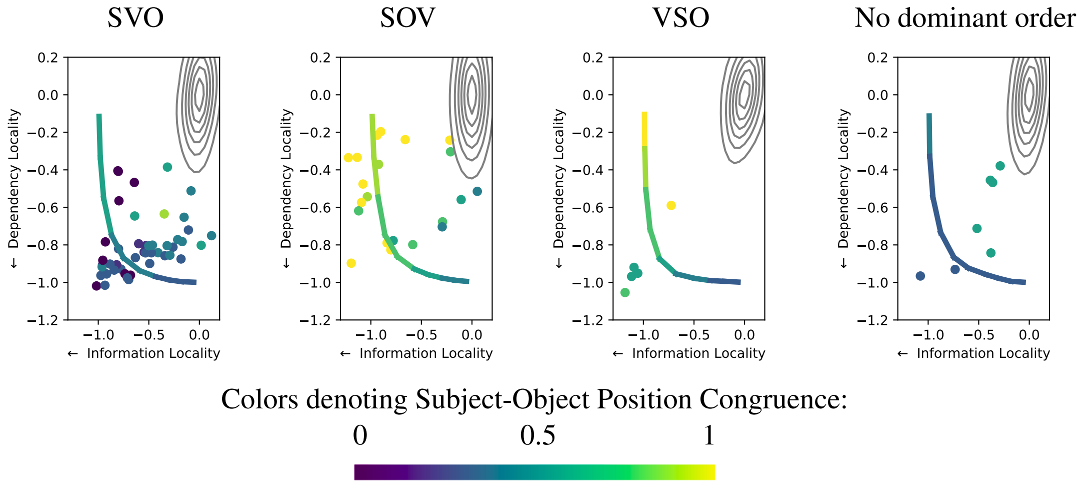


Figure S17: Position of languages in the efficiency plane spanned by IL and DL, per word order category. Compare Figure 3 in the main paper.

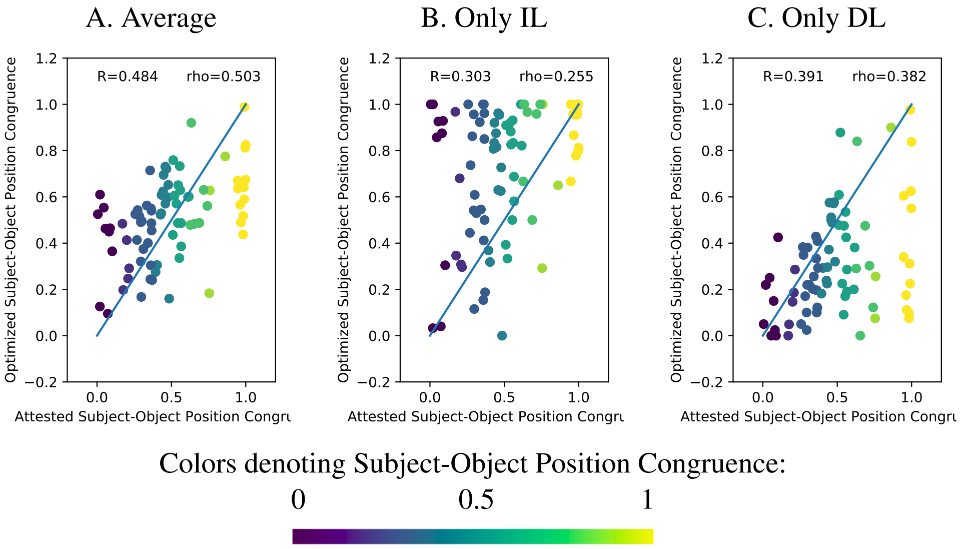


Figure S18: Results using raw counts for grammars optimized only for IL (center), only DL (right), and the average of the two counts (left). Results are very similar to those obtained using smoothed counts along the interpolated Pareto frontier, but do not depend on the method used to interpolate along the frontier.

S20 Results using Raw Counts

Here, we show that results concerning co-adaptation do not depend on the choice of a specific method for interpolating the Pareto frontier, or for interpolating position congruence along it. Figure S18 shows results corresponding to Figure 4 in the main paper, but representing the average optimized subject-object position congruence directly in terms of the average over optimized grammars, instead of smoothed values along the interpolated frontier. Results closely mirror those reported in the main paper.

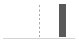
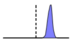
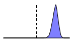
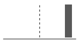
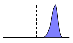
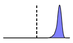
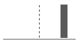
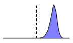
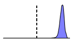
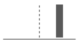
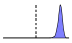
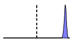
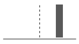
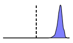
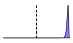
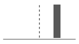
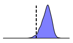
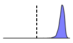
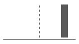
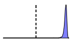
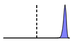
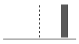
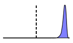
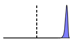
	Correlates with...		Real	Optimized for IL	Optimized for IL+DL
	verb *wrote*	object *letters*			
①	adposition *to*	noun phrase *a friend*			
②	copula *is*	noun phrase *a friend*			
③	auxiliary *has*	verb phrase *written*			
④	noun *friend*	genitive *of John*			
⑤	noun *books*	relative clause *that you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

Figure S19: Optimizing for Information Locality predicts the Greenberg correlations. Following Dryer [87], each correlation defines a pair of syntactic elements whose ordering is correlated with the relative order of object and verb; for instance, languages that place the object after the verb ("wrote letters") tend to place adpositions before the noun phrase ("to a friend"); languages that place the object after the verb (letters – wrote, Japanese) tend to place adpositions after the noun phrase (friend – to). For each correlation, we provide its prevalence (between 0% and 100%) among actual grammars of languages represented in Universal Dependencies (left, from Hahn et al. [27]), and the posterior distribution of the prevalence among grammars optimized for IL and DL, obtained from a mixed-effects analysis with by-language and by-family random effects (as in the analysis of Hahn et al. [27], but using the 80 languages from our sample used here). Optimization predicts all eight correlations to hold in the majority of grammars, matching the distribution observed in real languages.

S21 Comparison to Greenberg's Correlations

Here, we show that Greenberg's correlation universals [88, 87] arise from both IL and DL individually. Prior work has argued, using theoretical arguments, that these universals arise from optimizing DL [89, 90, 91]. This was confirmed by Hahn et al. [27, SI Appendix, Table S15] using word order grammars optimized for DL on 51 UD languages.[10] Here, we show that IL (and IL+DL) also predict these universals. Figure S19 shows the eight correlations as formalized in the Universal Dependencies format by Hahn et al. [27]. Results show that optimization for IL and IL+DL each predicts all of the correlations to hold in the majority of optimized grammars. This shows that, unlike in basic word order, the predictions of IL and DL converge on the Greenberg correlation universals, and explains why these

[10] We note that the predictions of DL for three of the correlations (1, 2, 6) are affected by specific properties of the Universal Dependencies format that deviate from the psycholinguistic theories underlying DL [1, 3] and from some other syntactic theories [92, 93]. Hahn et al. [27] followed Futrell et al. [94] in measuring dependency length in terms of a converted representation closer to those other theories; such a representation format is necessary to derive correlations 1, 2, 6 from DL [89, 90, 91]. In contrast, IL predicts Greenberg's correlations irrespective of these modeling assumptions, as it does not directly refer to syntactic structures.

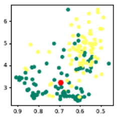
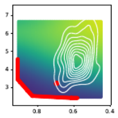
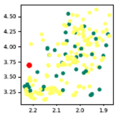
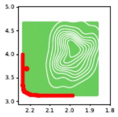
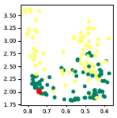
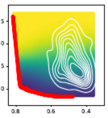
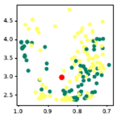
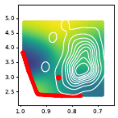
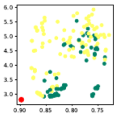
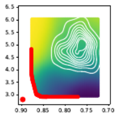
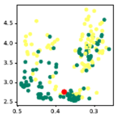
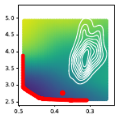
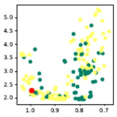
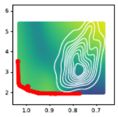
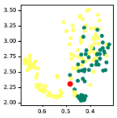
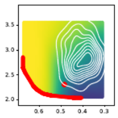
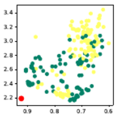
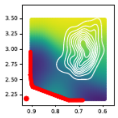
tend to hold across languages, whereas basic word order is much more variable.

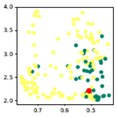
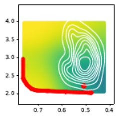
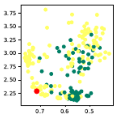
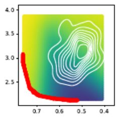
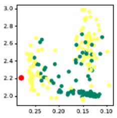
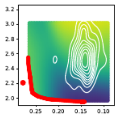
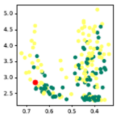
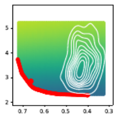
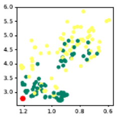
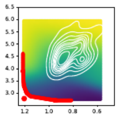
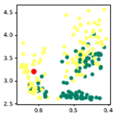
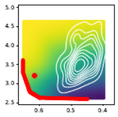
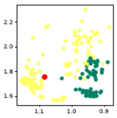
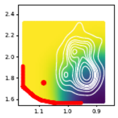
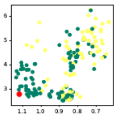
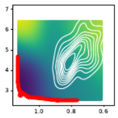
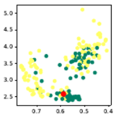
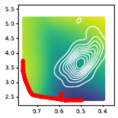
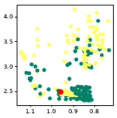
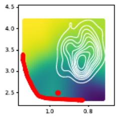
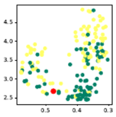
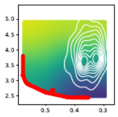
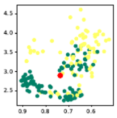
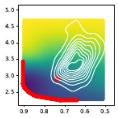
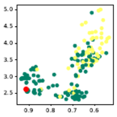
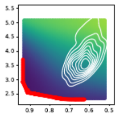
Hahn et al. [27] further argued that the Greenberg correlation universals can be derived from a principle of communicative efficiency closely related to efficiency principles that have found success in other domains of language [e.g. 95, 96, 97, 98, 99], balancing predictability with parseability, noting that optimizing communicative effiency also leads to efficiency in DL. We believe that communicative efficiency might be seen best as a possible justification of DL rather than being an orthogonal pressure. Evaluating the grammars optimized by Hahn et al. [27] for communicative efficiency on 51 languages, we found that they exhibit evidence for coadaptation, but overpredict SVO in a way very similar to grammars optimized solely for DL.

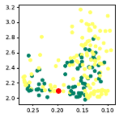
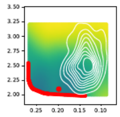
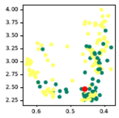
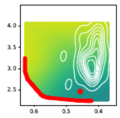
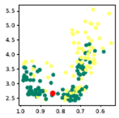
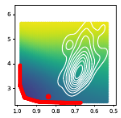
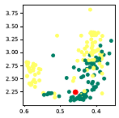
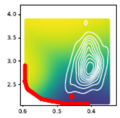
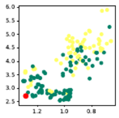
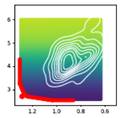
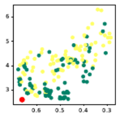
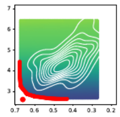
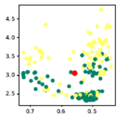
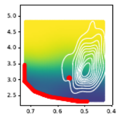
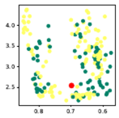
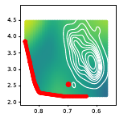
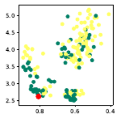
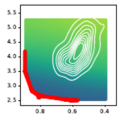
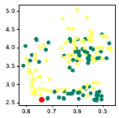
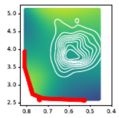
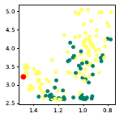
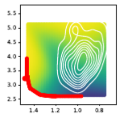
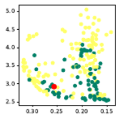
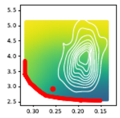
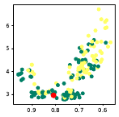
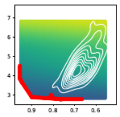
S22 Raw and Interpolated Efficiency Plane per Language

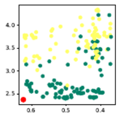
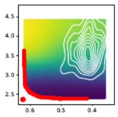
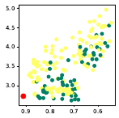
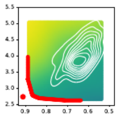
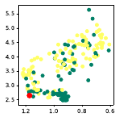
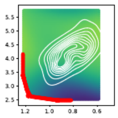
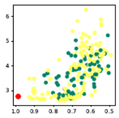
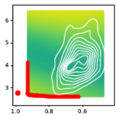
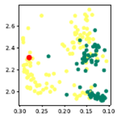
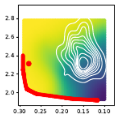
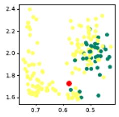
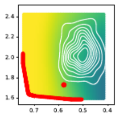
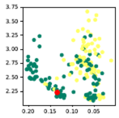
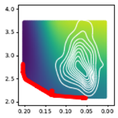
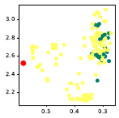
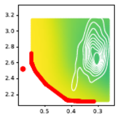
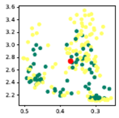
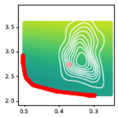
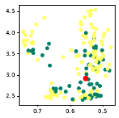
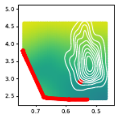
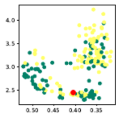
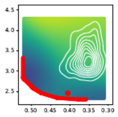
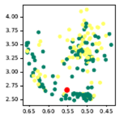
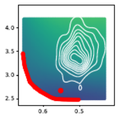
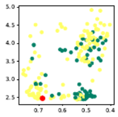
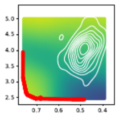
Here, we report per-language results for the efficiency planes. For each language, we first plot both the set of grammar samples, including both randomly constructed baseline grammars, and approximately optimized grammars inhabiting the area close to the Pareto frontier. These are colored depending on their subject-object position congruence, which is either 0 (green) or 1 (yellow). The red dot denotes the position of the real observed orderings. Second, we plot the interpolated average subject-object position congruence throughout the entire efficiency plane, the interpolated approximate Pareto frontier, and the distribution of randomly generated baseline grammars.

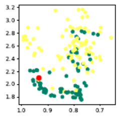
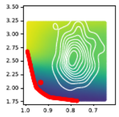
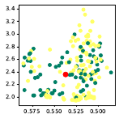
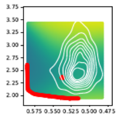
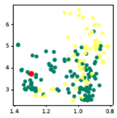
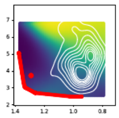
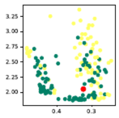
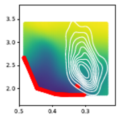
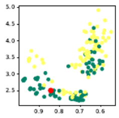
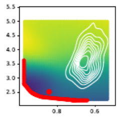
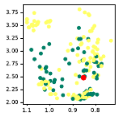
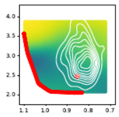
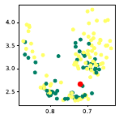
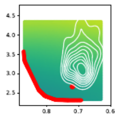
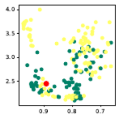
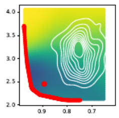
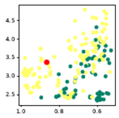
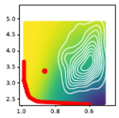
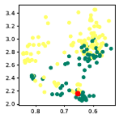
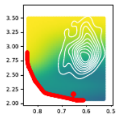
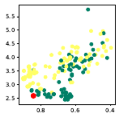
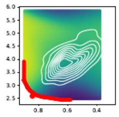
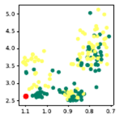
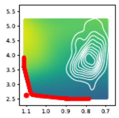
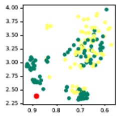
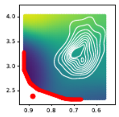
Note that some languages are beyond the approximate Pareto frontier; this can happen both because the optimization algorithm is approximate, and because real orderings are not subject to the same representational constraints as the grammars, enabling them to potentially be more efficient than is possible in the grammar formalism (see Section S28).

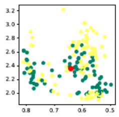
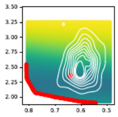
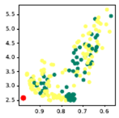
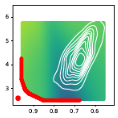
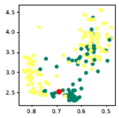
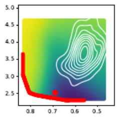
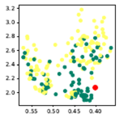
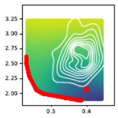
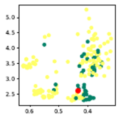
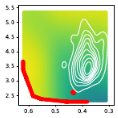
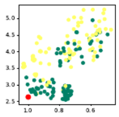
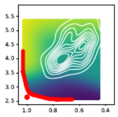
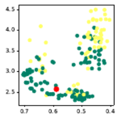
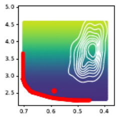
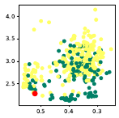
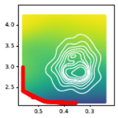
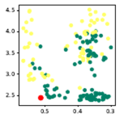
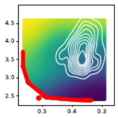
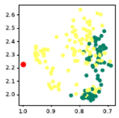
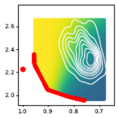
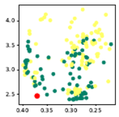
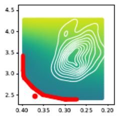
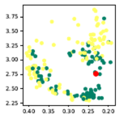
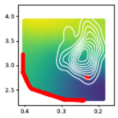
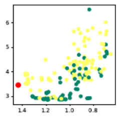
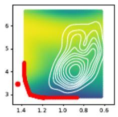
Language	ISO Code	Samples	Interpolated
Afrikaans	afr		
Akkadian	akk		
Amharic	amh		
Ancient Greek	grc		
Arabic	arb		
Armenian	hye		
Bambara	bam		
Basque	eus		
Belarusian	bel		

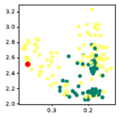
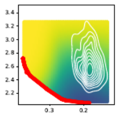
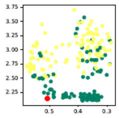
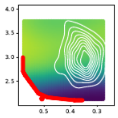
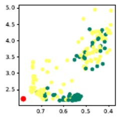
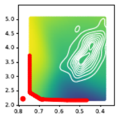
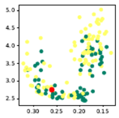
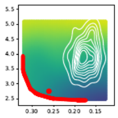
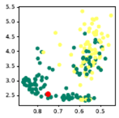
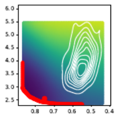
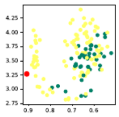
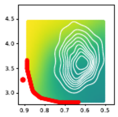
Breton	bre		
Bulgarian	bul		
Buryat	bua		
Cantonese	yue		
Catalan	cat		
Chinese	cmn		
Classical Chinese	lzh		
Coptic	cop		
Croatian	hrv		
Czech	ces		
Danish	dan		
Dutch	nld		
English	eng		

Erzya	myv		
Estonian	est		
Faroese	fao		
Finnish	fin		
French	fra		
Galician	glg		
German	deu		
Gothic	got		
Greek	ell		
Hebrew	heb		
Hindi	hin		
Hungarian	hun		
Icelandic	isl		

Indonesian	ind		
Irish	gle		
Italian	ita		
Japanese	jpn		
Kazakh	kaz		
Kiche	quc		
Komi Zyrian	kpv		
Korean	kor		
Kurmanji	kmr		
Latin	lat		
Latvian	lav		
Lithuanian	lit		
Maltese	mlt		

Manx	glv		
Mbya Guarani	gun		
Naija	pcm		
North Sami	sme		
Norwegian	nob		
Old Church Slavonic	chu		
Old East Slavic	orv		
Old French	fro		
Persian	pes		
Polish	pol		
Portuguese	por		
Romanian	ron		
Russian	rus		

Sanskrit	san		
Scottish Gaelic	gla		
Serbian	srp		
Slovak	slk		
Slovenian	slv		
Spanish	spa		
Swedish	swe		
Tamil	tam		
Thai	tha		
Turkish	tur		
Ukrainian	ukr		
Upper Sorbian	hsb		
Urdu	urd		

Uyghur	uig		
Vietnamese	vie		
Welsh	cym		
Western Armenian	hye2		
Wolof	wol		
Xibe	sjo		

S23 Neural Network Estimates of Information Locality

Here, we compare our formalization and estimation method for information locality to the method used by Hahn et al. [9], which is based on neural language models, estimating the next-word predictive distribution using LSTM recurrent neural networks [100]. Compared to the method used here, the use of recurrent neural networks can potentially result in better modeling of longer-range statistical relationships, and we asked whether this impacts our estimates of IL. Running the estimation method used by Hahn et al. [9] on all 80 languages was not feasible due to the high computational cost of neural network estimators.[11] We thus selected twelve languages representing typological, genetic, and geographic diversity within the bounds afforded by the UD data, and particularly where the Pareto frontier shows variability in subject-object position congruence:

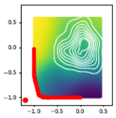
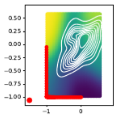
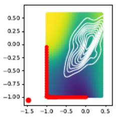
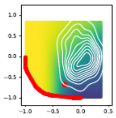
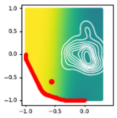
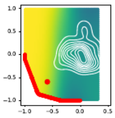
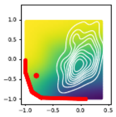
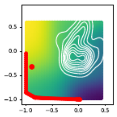
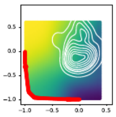
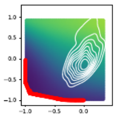
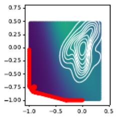
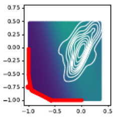
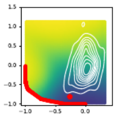
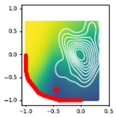
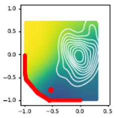
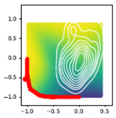
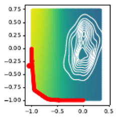
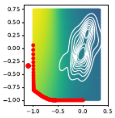
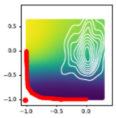
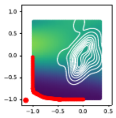
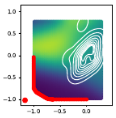
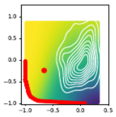
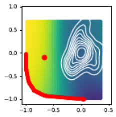
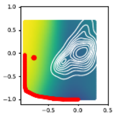
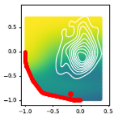
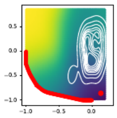
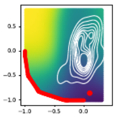
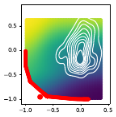
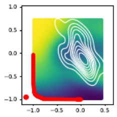
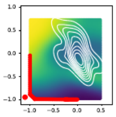
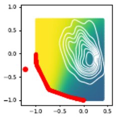
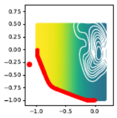
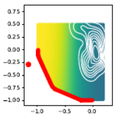
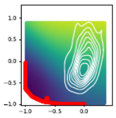
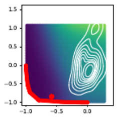
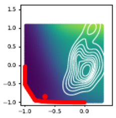
1. Arabic (VSO, Afro-Asiatic, Asia/Africa)
2. Basque (SOV, isolate, European)
3. Chinese (SVO, Sino-Tibetan, Asia)
4. English (SVO, Indo-European, European)
5. Finnish (SVO, Uralic, European)
6. Hindi (SOV, Indo-European, Asia)
7. Indonesian (SVO, Austronesian, Asia)
8. Persian (SOV, Indo-European, Asia)
9. Polish (SVO, Indo-European, Europe)
10. Thai (SVO, Tai-Kadai, Asia)
11. Turkish (SOV, Turkic, Asia)
12. Wolof (SVO, Niger-Congo, Africa)

We used the neural network-based method of Hahn et al. [9] to compute both I_1 and the AUC measure they used to quantify IL (see Section S1), for the approximately optimized grammars, the real orderings, and for ≥ 30 randomly constructed baseline grammars per language.[12] For the LSTM network, we used the hyperparameters that they had determined for each of the 12 languages to minimize surprisal on random baseline grammars.

Here, we show for each of the twelve languages the efficiency plane, with IL represented by (i) I_1 as computed by our method (Section S1.2), (ii) I_1 as computed using neural language models, and (iii) the AUC measure also computed using neural language models. For comparability across the three methods, we normalize DL and IL as in the main paper. Results are very similar, in the shape of the Pareto frontier, in the position of the real ordering, and in the distribution of subject-object position congruence throughout the efficiency plane.

[11] Hahn et al. [9] estimated information locality for 10–20 grammars in 54 languages. In contrast, we have ≈ 150 approximately optimized grammars for each of 80 languages.

[12] The baselines are different from those used in the other studies, as we had not recorded the baseline grammars, only their IL/DL values.

Language	I_1 (Our)	I_1 (Neural)	AUC (Neural)
Arabic			
Basque			
Chinese			
English			
Finnish			
Hindi			
Indonesian			
Persian			
Polish			
Thai			
Turkish			
Wolof			

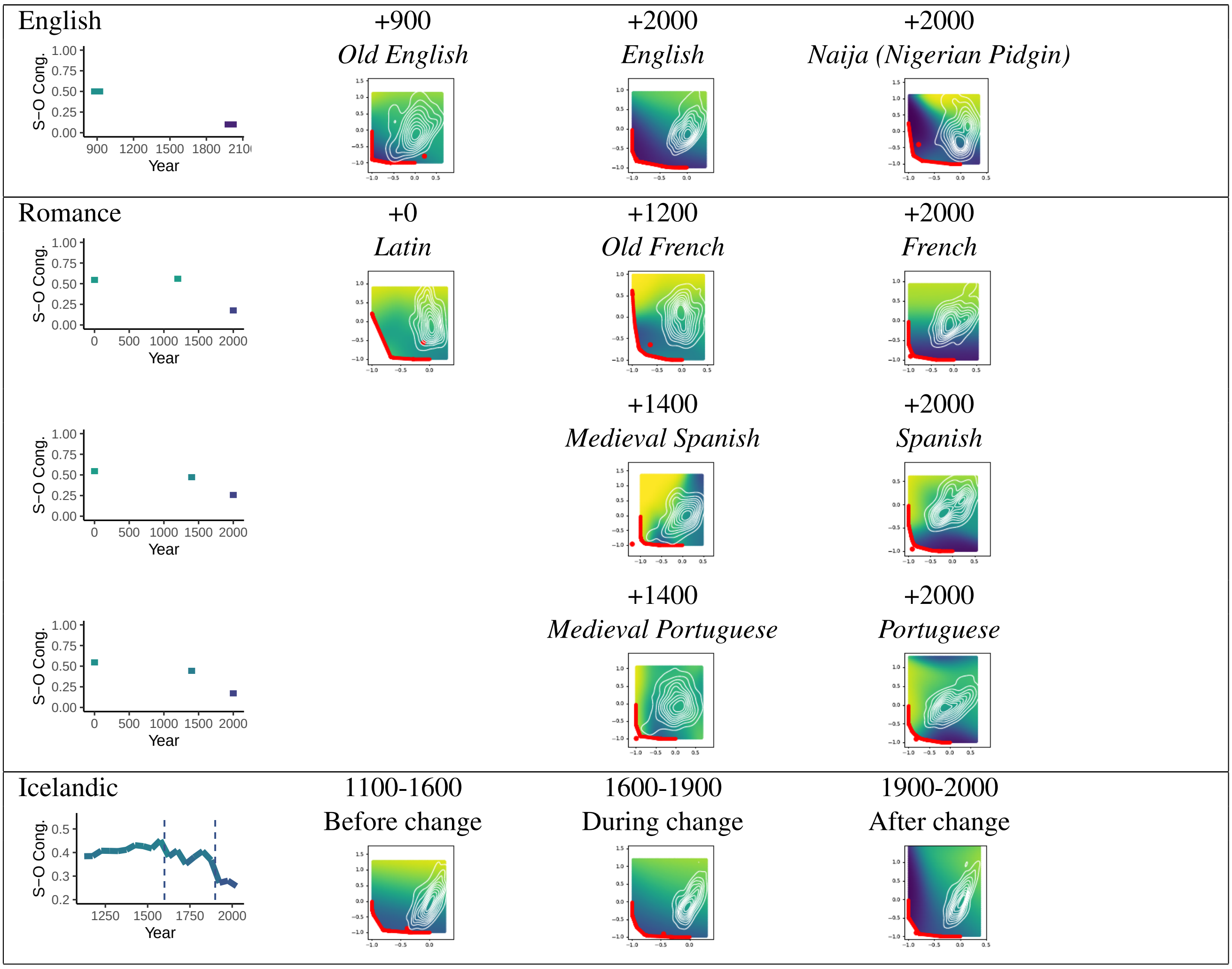
S24 Historical Changes

Below, we show efficiency planes for all languages that are attested in our dataset at multiple points in time.[13] For comparability, we normalize DL and IL as in the main paper.

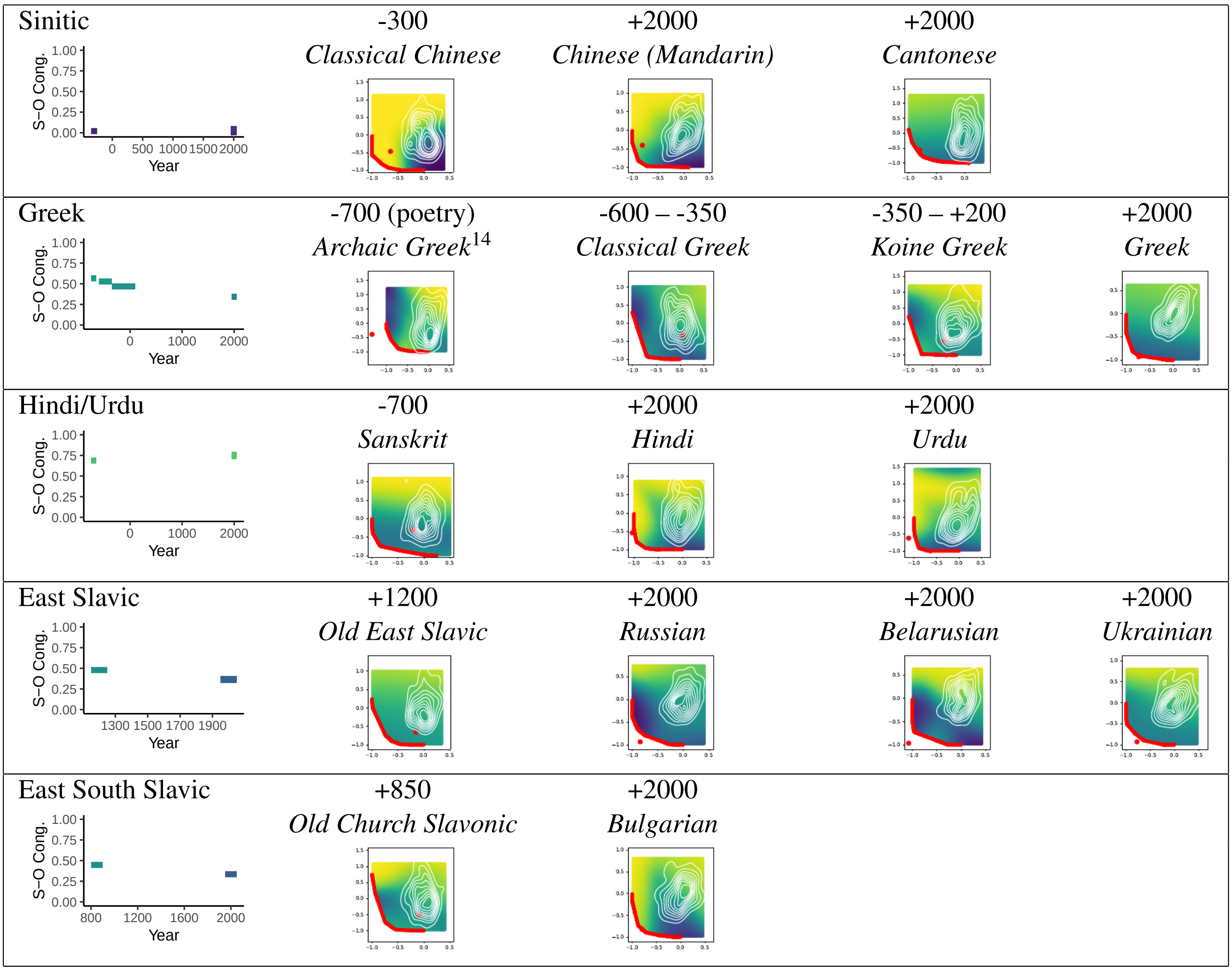
For each such language, we show the trajectory of subject-object position congruence (left), and the efficiency planes over time (right). In three cases of closely related languages with essentially identical subject-object position congrunece values, we plot those together (Sinitic, Hindi/Urdu, East Slavic).

In one case (Icelandic), data is available continuously across a word order change (compare Section S25); for the others, it is available at two or more points in time.

In terms of word order changes, there are multiple cases of changes towards SVO (lower subject-object position congruence; English, Romance, Icelandic), and also cases with only limited change (Sinitic, Greek, Hindi/Urdu, East and East South Slavic). In terms of efficiency, languages have moved closely along the frontier (e.g. English, Icelandic) or towards the frontier (e.g., Romance). Languages with change towards SVO (English, Romance, Icelandic) correspondingly exhibit an increase of SVO-like orderings (dark blue) along the Pareto frontier (compare Figure S20).



[13] For space and clarity, we only those descendants of Latin with an intermediate attested stage (French, Spanish, Portuguese). Trajectories for languages without data for intermediate stages (Italian, Catalan, Galician, Romanian) are similar; compare Figure S20.



[14] This corpus consists of poetry (Homer and Hesiod), potentially explaining the high efficiency in IL compared both to later stages, and to typologically similar ancient Indo-European languages like Latin and Sanskrit. See also caption of Figure S21.

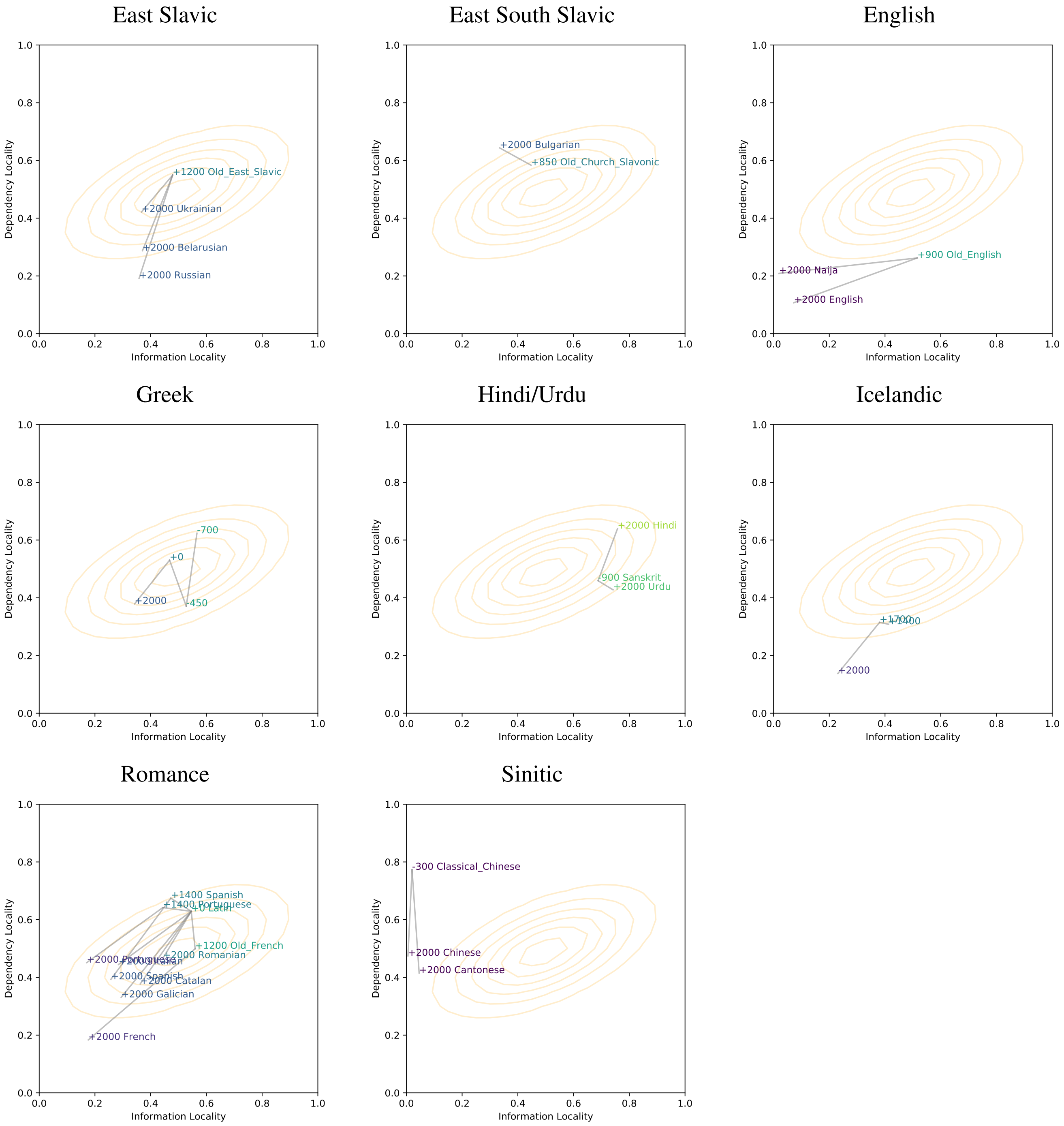


Figure S20: Historical trajectories of attested and average optimized subject-object position congruence. Faint contours describe the stationary distribution identified by the phylogenetic model. See Section S24.

S25 Other Historical Datasets

In Figure S21, we provide details on the historical datasets obtained from outside the Universal Dependencies Project, and by splitting UD corpora into multiple epochs. We plot per-dataset results as in Section S22.

Periodization of Ancient Greek As shown in Figure S21, we split Ancient Greek into three conventional phases: Archaic Greek (covering data from Homer and Hesiod, about 700BC), Classical Greek ($\approx$ 600-350 BC, represented e.g. by Herodotus and Sophocles, both $\approx$ 450BC), and Koine Greek ($\approx$ 350 BC–200 AD, represented e.g. by the New Testament and Diodorus Siculus).

Periodization of Icelandic In Icelandic, continuous corpus data is available from the 12th century onwards [103]. While the grammatical structure of Icelandic remained largely constant during this time, Icelandic witnessed a word order change where previously common OV order became much rarer. Hróarsdóttir [102, p. 59] states "*OV word order seems to have been as frequent as VO word order in texts until the seventeenth century, but the frequency of OV-orders drops 30–40% in texts dating from the seventeenth and eighteenth centuries. In the nineteenth century texts studied, the frequency of OV word order has dropped to an average of 24.8%.*" Figure S22 shows the trajectory of attested subject-object position congruence, binning all texts in the dataset by half-centuries (e.g., 1200–1250, 1250–1300, etc.). Subject-object position congruence appears to drop in the 17th century, and only reaches its current low level in the 20th century. We thus grouped the data into three periods, 1100–1600, 1600–1900, 1900–today. The first period largely coincides with the conventional period of Old Icelandic, which is conventionally taken to have ended about 1540 with the translation of the New Testament [104].

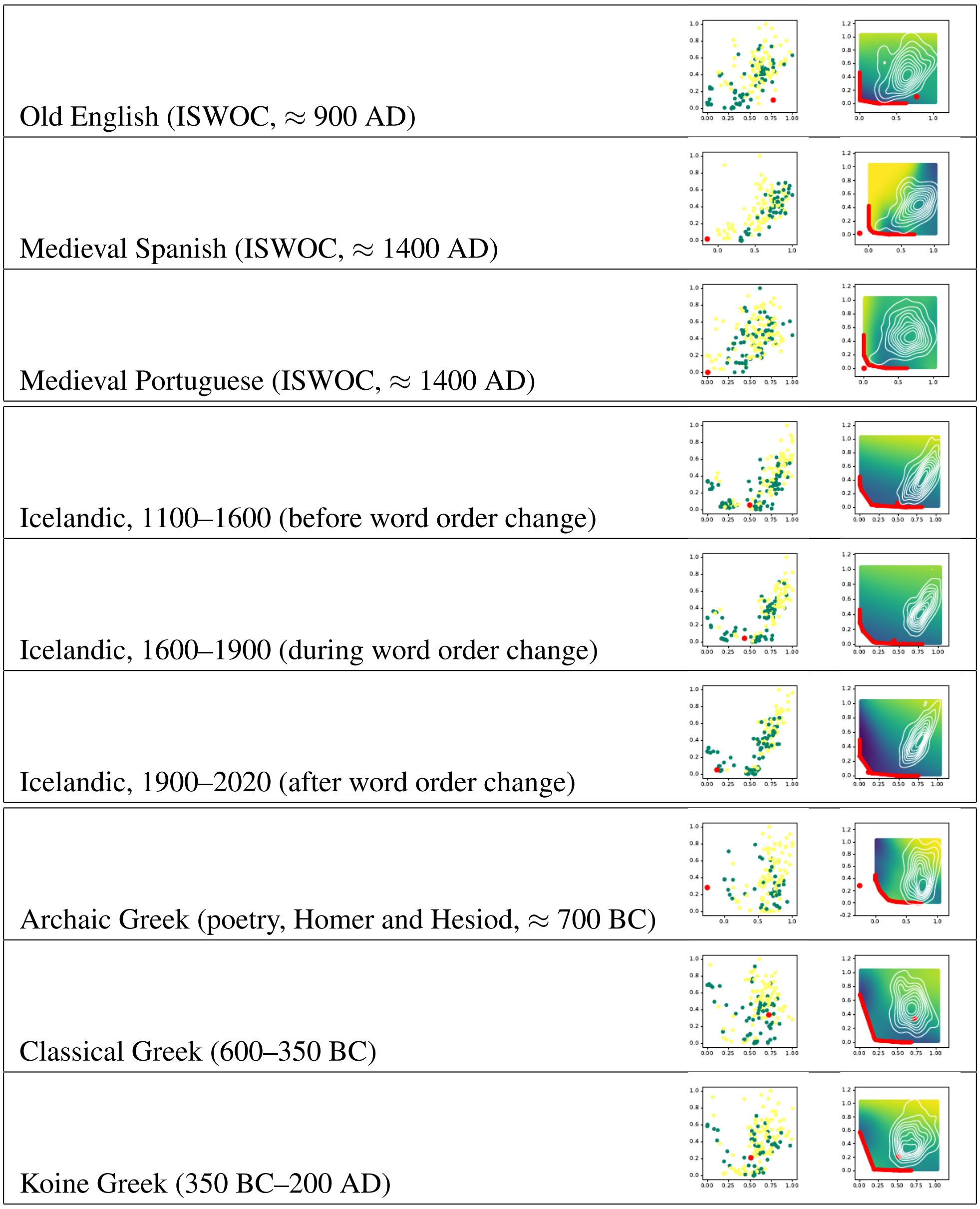


Figure S21: Additional historical corpora, in other dependency grammar formalisms, or obtained by splitting UD treebanks into distinct epochs. First, we considered the treebanks in the **ISWOC collection** [101], covering Old English, Medieval Spanish, and Medieval Portuguese. These corpora are annotated in a dependendency grammar format, though with differences from the Universal Dependencies formalism. Second, we split the **Icelandic** data, which spans almost a millenium, into three phases. We conducted the split based on a documented word-order change, whereby SOV was partly replaced by SVO order, between the 16th and 19th centuries [102] (see text for details). Finally, we split the **Ancient Greek** data, based on three conventional phases. What stands out is that Archaic Greek appears highly efficient on IL, in contrast both to later forms of Ancient Greek and related early Indo-European languages. We attribute this to the fact that the Archaic Greek subset consists entirely of poetry. The presence of meter might increase local predictability, though we leave an investigation of possible interactions of information locality and meter to future research.

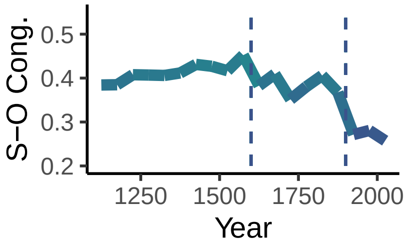


Figure S22: Attested subject-object position congruence in Icelandic. The vertical bars denote the boundaries between the three periods for which we compute the Pareto frontier. See text (Section S25) for details.

Language	Number of Sentences
English SWBD	110,504
French-Spoken	2,789
Norwegian-NynorskLIA	5,250
Slovenian-SST	3,188
TuebaJS (Japanese)	17,753

Table S14: Corpus sizes for spoken corpora.

S26 Role of Modality

Here, we consider the effect of text modality on usage patterns. Most corpora in the Universal Dependencies collection reflect written text. We identified six datasets of spoken languages in the Universal Dependencies format or other dependency grammar formalisms. The Naija corpus consists entirely of spoken text. For Slovenian, French, and Norwegian, there are sub-corpora reflecting spoken text. We further considered the Tueba J/S corpus of spontenous dialogue in Japanese [105] For English, we used an automated conversion [106] of the Switchboard section of the Penn Treebank [107] to Universal Dependencies. Corpus sizes are shown in Table S14.

We compare attested and average optimized subject-object position congruence on these six datasets in Figure S23. Results confirm that usage patterns as observed in spoken corpora also support the proposal of coadaptation between usage patterns and word order.

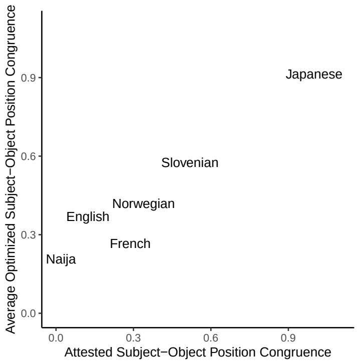


Figure S23: Attested and average optimized subject-object position congruence for six datasets of spoken text. Attested and average optimized congruence are correlated ($R = 0.96$, $p = 0.002$; $\rho = 0.94$, $p = 0.02$).

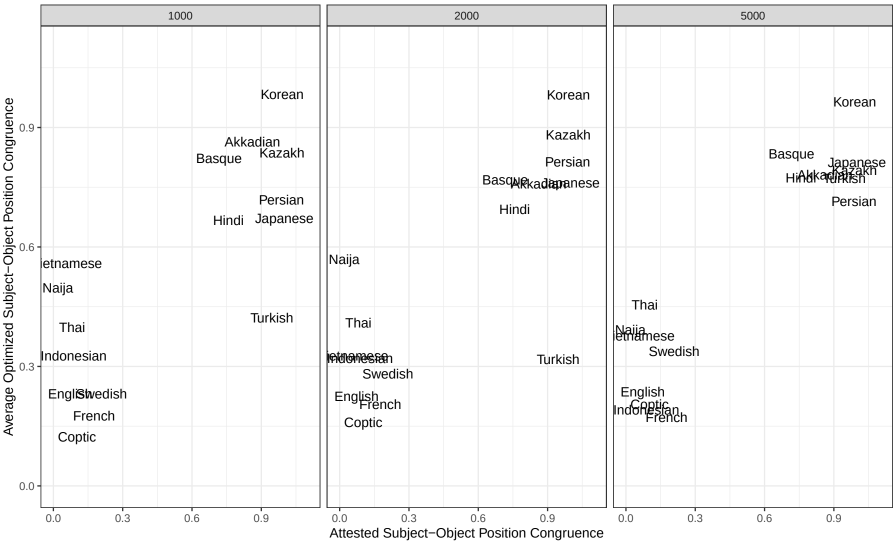


Figure S24: Comparing attested and average optimized subject-object position congruence in subsampled subsets (1K, 2K, 5K sentences), in 16 languages.

S27 Effects of Corpus Size

The datasets available for different langusages differ substantially in their sizes. While some languages have tens of thousands of sentences available, corpora for other languages are substantially smaller. This raises the question whether estimates of the efficiency plane provide sufficiently reliable signal when corpora are small. To evaluate this, we selected 8 langages with very high subject-object position congruence, and 8 languages with very low subject-object position congruence. For each language, we randomly selected subsets of 1,000, 2,000, and of 5,000 sentences, and estimated average optimized subject-object position congruence along the Pareto frontier. Results are shown in Figure S24. Results suggest that estimates at 1,000 sentences may be noisier, but they are nonetheless highly correlated with estimates at 5,000 sentences ($R = 0.89$, $p < 0.00001$). In order to account for errors due to finiteness of corpora, we also considered a version of the Ornstein-Uhlenbeck process incorporating measurement noise, finding that it continues to support our conclusions, and in fact estimates stronger correlations than our main analysis (Section S8).

S28 Comparison to Fitted Grammars

Here, we compare to results obtained when representing the grammar of real languages using ordering grammars subject to the same representational constraints as the baseline and approximately optimized grammars, in order to assess the role of word order flexibility beyond the constraints of the ordering grammar formalism in efficiency optimization. For each language, we used the hill-climbing method also used for optimizing grammars for efficiency to find a grammar which fits the observed orderings, in the sense that it maximizes the fraction of pairs of dependents of the same head that are ordered in the same order as in the attested order. We then evaluated these for DL and IL. Results are shown in Figure S25.

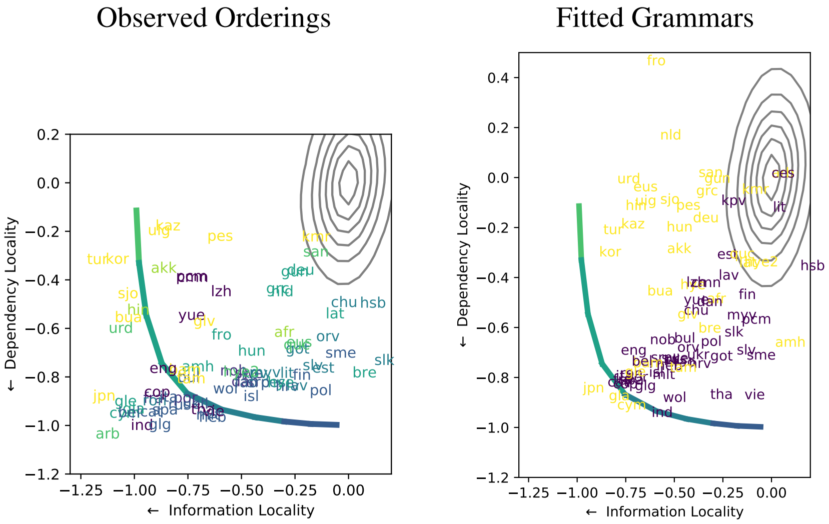


Figure S25: Comparing efficiency of the attested orderings with ordering grammars fitted to the attested orderings for each language. The fitted grammars are subject to exactly the same representational constraints as the baseline and approximately optimized grammars; in particular, they are a deterministic fucntion of the sentences and the syntactic relations between them. Similar to the observed orderings, fitted orderings tend to be more efficient than the baseline orderings, inhabiting the region between the baselines and the Pareto frontier. Note that, due to their design, the subject-object position congruence of fitted grammars is either 0 or 1. Observed orderings tend to be even more efficient than fitted grammars, suggesting that human languages use word order flexibility to achieve higher efficiency.

References

[1] E. Gibson. Linguistic complexity: locality of syntactic dependencies. *Cognition*, 68(1):1–76, August 1998. ISSN 0010-0277.

[2] Brian McElree, Stephani Foraker, and Lisbeth Dyer. Memory structures that subserve sentence comprehension. *Journal of Memory and Language*, 48(1):67–91, January 2003. ISSN 0749596X. doi: 10.1016/S0749-596X(02)00515-6. URL `http://linkinghub.elsevier.com/retrieve/pii/S0749596X02005156`.

[3] Richard L. Lewis and Shravan Vasishth. An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29(3):375–419, May 2005. ISSN 0364-0213. doi: 10.1207/s15516709cog0000_25.

[4] John T. Hale. 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*, pages 1–8, 2001.

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

[6] Vera Demberg and Frank Keller. A computational model of prediction in human parsing: Unifying locality and surprisal effects. In *Proceedings of the 29th meeting of the Cognitive Science Society (CogSci-09)*, 2009. URL `http://www.coli.uni-saarland.de/~vera/demberg_keller_cogsci_2009.pdf`.

[7] Marisa Ferrara Boston, John T Hale, Shravan Vasishth, and Reinhold Kliegl. Parallel processing and sentence comprehension difficulty. *Language and Cognitive Processes*, 26(3):301–349, 2011.

[8] Richard Futrell, Edward Gibson, and Roger P. Levy. Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, 44, 2020.

[9] Michael Hahn, Judith Degen, and Richard Futrell. Modeling word and morpheme order in natural language as an efficient tradeoff of memory and surprisal. *Psychological Review*, 2021.

[10] Maryellen C MacDonald and Morten H Christiansen. Reassessing working memory: Comment on just and carpenter (1992) and waters and caplan (1996). 2002.

[11] Stefan L. Frank, Thijs Trompenaars, and Shravan Vasishth. Cross-linguistic differences in processing double-embedded relative clauses: Working-memory constraints or language statistics? *Cognitive Science*, 40(3):554–578, 2016.

[12] Nathan E Rasmussen and William Schuler. Left-corner parsing with distributed associative memory produces surprisal and locality effects. *Cognitive science*, 42:1009–1042, 2018.

[13] Ting Qian and T Florian Jaeger. Cue effectiveness in communicatively efficient discourse production. *Cognitive science*, 36(7):1312–1336, 2012.

[14] Jennifer Culbertson, Marieke Schouwstra, and Simon Kirby. From the world to word order: Deriving biases in noun phrase order from statistical properties of the world. *Language*, 2020.

[15] W Ebeling and T Pöschel. Entropy and Long-Range Correlations in Literary English. *Europhysics Letters (EPL)*, 26(4):241–246, May 1994. ISSN 0295-5075, 1286-4854. doi: 10.1209/0295-5075/26/4/001. URL http://stacks.iop.org/0295-5075/26/i=4/a=001?key=crossref.4a8da3b3f5e80b828e6995b6bfc3e5be.

[16] George A. Miller and Jennifer A. Selfridge. Verbal context and the recall of meaningful material. *The American journal of psychology*, 63 2:176–85, 1950.

[17] Murray Aborn, Herbert Rubenstein, and Theodor D. Sterling. Sources of contextual constraint upon words in sentences. *Journal of experimental psychology*, 57 3:171–80, 1959.

[18] Claude E Shannon. Prediction and entropy of printed english. *Bell system technical journal*, 30 (1):50–64, 1951.

[19] Christian Bentz, Dimitrios Alikaniotis, Michael Cysouw, and Ramon Ferrer-i Cancho. The entropy of words – learnability and expressivity across more than 1000 languages. *Entropy*, 19(6): 275, 2017.

[20] Shuntaro Takahashi and Kumiko Tanaka-Ishii. Cross entropy of neural language models at infinity – a new bound of the entropy rate. *Entropy*, 20(11):839, 2018.

[21] Daniel Gildea and T. Florian Jaeger. Human languages order information efficiently. *arXiv:1510.02823 [cs]*, October 2015. URL http://arxiv.org/abs/1510.02823. arXiv: 1510.02823.

[22] David McAllester and Karl Stratos. Formal limitations on the measurement of mutual information. In Silvia Chiappa and Roberto Calandra, editors, *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pages 875–884. PMLR, 2020.

[23] Reinhard Kneser and Hermann Ney. Improved backing-off for m-gram language modeling. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, volume 1, pages 181–184. IEEE, 1995.

[24] Yee Whye Teh. A bayesian interpretation of interpolated Kneser-Ney. 2006.

[25] Daniel Gildea and David Temperley. Optimizing Grammars for Minimum Dependency Length. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 184–191, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P07-1024`.

[26] Daniel Gildea and David Temperley. Do Grammars Minimize Dependency Length? *Cognitive Science*, 34(2):286–310, March 2010. ISSN 1551-6709. doi: 10.1111/j.1551-6709.2009.01073.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1551-6709.2009.01073.x`.

[27] Michael Hahn, Dan Jurafsky, and Richard Futrell. Universals of word order reflect optimization of grammars for efficient communication. *Proceedings of the National Academy of Sciences of the United States of America*, 117(5):2347–2353, 2020.

[28] Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992. URL `http://link.springer.com/article/10.1007/BF00992696`.

[29] Josef Stoer and R. Bulirsch. Introduction to numerical analysis. 2002.

[30] Steven Diamond and Stephen P. Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of machine learning research : JMLR*, 17, 2016.

[31] Sebastian Nordhoff and Harald Hammarström. Glottolog/langdoc: defining dialects, languages, and language families as collections of resources. In *LISC'11 Proceedings of the First International Conference on Linked Science - Volume 783*, pages 53–58, 2011.

[32] Igor Diakonoff. The earliest semitic society linguistic data. *Journal of Semitic Studies*, 43(2): 209–219, 1998.

[33] Eric W. Holman, Cecil H. Brown, Søren Wichmann, André Müller, Viveka Velupillai, Harald Hammarström, Sebastian Sauppe, Hagen Jung, Dik Bakker, Pamela Brown, Oleg Belyaev, Matthias Urban, Robert Mailhammer, Johann-Mattis List, and Dmitry Egorov. Automated dating of the world's language families based on lexical similarity. *Current Anthropology*, 52(6): 841–875, 2011. doi: 10.1086/662127. URL `https://doi.org/10.1086/662127`.

[34] Jasmine Dum-Tragut. Armenian: Modern eastern armenian. 2009.

[35] Russell D. Gray and Quentin D. Atkinson. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, 426(6965):435–439, 2003.

[36] Andrew Kitchen, Christopher Ehret, Shiferaw Assefa, and Connie J. Mulligan. Bayesian phylogenetic analysis of semitic languages identifies an early bronze age origin of semitic in the near east. *Proceedings of The Royal Society B: Biological Sciences*, 276(1668):2703–2710, 2009.

[37] Alexander Savelyev and Martine Robbeets. Bayesian phylolinguistics infers the internal structure and the time-depth of the turkic language family. *Journal of Language Evolution*, 5(1):39–53, 2020.

[38] Petra Novotná and Václav Blazek. Glottochronology and its application on the balto-slavic languages. *Baltistica*, 42(2):185–210, 2011.

[39] L Maurits, M de Heer, T Honkola, M Dunn, and O Vesakoski. Best practices in justifying calibrations for dating language families. *Journal of Language Evolution*, 5(1):17–38, 2020.

[40] K. H. Jackson. Common gaelic: The evolution of the goidelic languages. *Proceedings of the British Academcy*, 37:71–97, 1951.

[41] Asko Parpola. Formation of the Indo-European and Uralic language families in the light of archaeology: Revised and integrated 'total' correlations. pages 119–184, 2013.

[42] Helge Sandøy, Oskar Bandle, Kurt Braunmüller, Ernst Hakon Jahr, Allan Karker, Hans-Peter Naumann, Ulf Teleman, Lennart Elmevik, and Gun Widmark. The typological development of the Nordic languages i: Phonology. 2017.

[43] Joseph Felsenstein. Phylogenies from molecular sequences: Inference and reliability. *Annual Review of Genetics*, 22(1):521–565, 1988.

[44] Thomas F. Hansen. Stabilizing selection and the comparative analysis of adaptation. *Evolution*, 51(5):1341–1351, 1997.

[45] P. G. Blackwell. Bayesian inference for markov processes with diffusion and discrete components. *Biometrika*, 90(3):613–627, 2003.

[46] Siegfried Schach. Weak convergence results for a class of multivariate markov processes. *The Annals of Mathematical Statistics*, 42(2):451–465, 1971.

[47] Crispin W. Gardiner. *Handbook of Stochastic Methods*. Springer, 1983.

[48] Hannes Risken. *The Fokker-Planck equation*. Springer, 2 edition, 1989.

[49] John Barnard, Robert McCulloch, and Xiao Li Meng. Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, 10(4): 1281–1311, 2000.

[50] Daniel Lewandowski, Dorota Kurowicka, and Harry Joe. Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, 100(9):1989–2001, 2009.

[51] Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus A. Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1):1–32, 2017.

[52] Wangang Xie, Paul O. Lewis, Yu Fan, Lynn Kuo, and Ming Hui Chen. Improving marginal likelihood estimation for bayesian phylogenetic model selection. *Systematic Biology*, 60(2):150–160, 2011.

[53] Joseph Felsenstein. Maximum-likelihood estimation of evolutionary trees from continuous characters. *American Journal of Human Genetics*, 25(5):471–492, 1973.

[54] Robert P. Freckleton. Fast likelihood calculations for comparative analyses. *Methods in Ecology and Evolution*, 3(5):940–947, 2012.

[55] Emmanuel Paradis, Julien Claude, and Korbinian Strimmer. Ape: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20 2:289–90, 2004.

[56] Michael Dunn, Simon J. Greenhill, Stephen C. Levinson, and Russell D. Gray. Evolved structure of language shows lineage-specific trends in word-order universals. *Nature*, 473(7345):79–82, May 2011. ISSN 0028-0836. doi: 10.1038/nature09923. URL `http://www.nature.com/nature/journal/v473/n7345/full/nature09923.html`.

[57] Luke Maurits and Thomas L. Griffiths. Tracing the roots of syntax with bayesian phylogenetics. *Proceedings of the National Academy of Sciences of the United States of America*, 111(37): 13576–13581, 2014.

[58] Matthew S. Dryer. Large linguistic areas and language sampling. *Studies in Language*, 13(2): 257–292, 1989.

[59] Walter Bisang. Areal typology and grammaticalization: Processes of grammaticalization based on nouns and verbs in East and Mainland South East Asian languages. *Studies in Language*, 20 (3):519–597, 1996.

[60] Bernd Heine and Tania Kuteva. On contact-induced grammaticalization. *Studies in Language*, 27 (3):529–572, 2003.

[61] Alexandra Y. Aikhenvald. Grammars in contact: a cross-linguistic perspective. 2007.

[62] Siva Kalyan, Alexandre François, and Harald Hammarström. Problems with, and alternatives to, the tree model in historical linguistics. *Journal of Historical Linguistics*, 9(1):1–8, 2019.

[63] Lyle Campbell, Terrence Kaufman, and Thomas C. Smith-Stark. Meso-America as a linguistic area. *Language*, 62(3):530–570, 1986.

[64] Johanna Nichols. *Linguistic Diversity in Space and Time*. 1992.

[65] Martin Haspelmath. The european linguistic area: Standard average european. pages 1492–1510, 2001.

[66] Rik van Gijn, Harald Hammarström, Simon van de Kerke, Olga Krasnoukhova, and Pieter Muysken. Linguistic areas, linguistic convergence, and river systems in South America. pages 964–996. 2017.

[67] Scott L. Nuismer and Luke J. Harmon. Predicting rates of interspecific interaction from phylogenetic trees. *Ecology Letters*, 18(1):17–27, 2015.

[68] Marc Manceau, Amaury Lambert, and Heléne Morlon. A unifying comparative phylogenetic framework including traits coevolving across interacting lineages. *Systematic Biology*, 66(4): 551–568, 2016.

[69] Jonathan Drury, Julien Clavel, Marc Manceau, and Hélène Morlon. Estimating the effect of competition on trait evolution using maximum likelihood inference. *Systematic Biology*, 65(4): 700–710, 2016.

[70] Krzysztof Bartoszek, Sylvain Glémin, Ingemar Kaj, and Martin Lascoux. Using the ornstein-uhlenbeck process to model the evolution of interacting populations. *Journal of Theoretical Biology*, 429:35–45, 2017.

[71] Jonathan P. Drury, Gregory F. Grether, Theodore Garland Jr., and Hélène Morlon. An assessment of phylogenetic tools for analyzing the interplay between interspecific interactions and phenotypic evolution. *Systematic Biology*, 67(3):413–427, 2018.

[72] Aasa Feragen, Francois Lauze, and Soren Hauberg. Geodesic exponential kernels: When curvature and linearity conflict. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3032–3042, 2015.

[73] Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie. *The World Atlas of Language Structures*. 2005.

[74] Oliver A. Iggesen. Number of cases. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013. URL `https://wals.info/chapter/49`.

[75] Theo Vennemann. Explanation in syntax. In John Kimball, editor, *Syntax and Semantics*, volume 2, pages 1–50. New York, 1974.

[76] Charles N. Li and Sandra A. Thompson. *Mandarin Chinese, a functional reference grammar*. University of California Press, Berkeley/Los Angeles/London, 1981.

[77] Matthew S. Dryer. Order of subject and verb. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013. URL `https://wals.info/chapter/82`.

[78] Matthew S. Dryer. Order of subject, object and verb. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013. URL `https://wals.info/chapter/81`.

[79] Frans Plank and Elena Filimonova. The Universals Archive: A brief introduction for prospective users. *STUF - Language Typology and Universals*, 53(1), 2000.

[80] Isaak Kozinsky. *Nekotorye grammaticeskie universalii v podsistemax vyrazenija subjektno-objektnyx otnnosenij. [Some grammatical universals in subsystems of expression of subject-object relations.]*. 1981. Doctoral dissertation, Moskovskij gosudarstvennyj universitet.

[81] Abdulrahman Alqurashi. An HPSG approach to free relatives in arabic. In Stefan Müller, editor, *Proceedings of the 19th International Conference on Head-Driven Phrase Structure Grammar, Chungnam National University Daejeon*, pages 6–25, Stanford, CA, 2012. CSLI Publications. URL `http://cslipublications.stanford.edu/HPSG/2012/alqurashi.pdf`.

[82] Lenora A. Timm. Relative clause formation in breton. *WORD*, 39(2):79–107, 1988.

[83] Alan Gardiner. *Egyptian Grammar, Being an Introduction to the Study of Hieroglyphs*. 1957.

[84] Jeffrey Heath. *A Grammar of Tamashek (Tuareg of Mali)*. 2005.

[85] Katherine Demuth and Carolyn Harford. Verb raising and subject inversion in Bantu relatives. *Journal of African Languages and Linguistics*, 20(1):41–61, 1999.

[86] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. 1995.

[87] Matthew S Dryer. The Greenbergian word order correlations. *Language*, 68(1):81–138, 1992.

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

[89] Lyn Frazier. Syntactic complexity. *Natural language parsing: Psychological, computational, and theoretical perspectives*, pages 129–189, 1985.

[90] Jan Rijkhoff. Word order universals revisited: The principle of head proximity. *Belgian Journal of Linguistics*, 1(1):95–125, 1986.

[91] John A. Hawkins. *A performance theory of order and constituency*. Cambridge University Press, Cambridge, 1994.

[92] Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. Sud or surface-syntactic universal dependencies: An annotation scheme near-isomorphic to ud. In *UDW@EMNLP*, 2018.

[93] Timothy Osborne and Kim Gerdes. The status of function words in dependency grammar: A critique of universal dependencies (ud). *Glossa: a journal of general linguistics*, 2019.

[94] Richard Futrell, Kyle Mahowald, and Edward Gibson. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341, August 2015. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1502134112. URL `http://www.pnas.org/lookup/doi/10.1073/pnas.1502134112`.

[95] Ramon Ferrer i Cancho and Ricard V Solé. Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences*, 100(3):788–791, 2003.

[96] Yang Xu, Terry Regier, and Barbara C. Malt. Historical semantic chaining and efficient communication: The case of container names. *Cognitive Science*, 40(8):2081–2094, 2016.

[97] Noga Zaslavsky, Charles Kemp, Terry Regier, and Naftali Tishby. Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, 115(31):7937–7942, 2018.

[98] Noga Zaslavsky, Mora Maldonado, and Jennifer Culbertson. Let's talk (efficiently) about us: Person systems achieve near-optimal compression. In *CogSci*, 2021.

[99] Francis Mollica, Geoff Bacon, Noga Zaslavsky, Yang Xu, Terry Regier, and Charles Kemp. The forms and meanings of grammatical markers support efficient communication. *Proceedings of the National Academy of Sciences*, 118(49), 2021. ISSN 0027-8424. doi: 10.1073/pnas.2025993118. URL `https://www.pnas.org/content/118/49/e2025993118`.

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

[101] Kristin Bech and Kristine Eide. The ISWOC corpus, 2014. http://iswoc.github.io.

[102] Thorbjörg Hróarsdóttir. *Word Order Change in Icelandic: From OV to VO*. J Benjamins, Amsterdam, Philadelphia, 2000.

[103] Þórunn Arnardóttir, Hinrik Hafsteinsson, Einar Freyr Sigurðsson, Kristín Bjarnadóttir, Anton Karl Ingason, Hildur Jónsdóttir, and Steinþór Steingrímsson. A Universal Dependencies conversion pipeline for a Penn-format constituency treebank. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 16–25, Barcelona, Spain (Online), December 2020. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/2020.udw-1.3`.

[104] Endre Mørck. Mellomnorsk sprak. In Odd Einar Haugen, editor, *Handbok i norrøn filologi*, pages 407–450. Bergen, 2004.

[105] Sabine Buchholz and Erwin Marsi. Conll-x shared task on multilingual dependency parsing. pages 149–164, 2006.

[106] Sebastian Schuster, Joakim Nivre, and Christopher D. Manning. Sentences with gapping: Parsing and reconstructing elided predicates. In *NAACL HLT 2018: 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 1156–1168, 2018.

[107] Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics*, 19(2):313–330, 1993.