Tracing the roots of syntax with Bayesian phylogenetics

Luke Maurits[1] and Thomas L. Griffiths

Department of Psychology, University of California, Berkeley, CA 94704

Edited by Richard N. Aslin, University of Rochester, Rochester, NY, and approved July 25, 2014 (received for review October 8, 2013)

The ordering of subject, verb, and object is one of the fundamental components of the syntax of natural languages. The distribution of basic word orders across the world's languages is highly nonuniform, with the majority of languages being either subject-object-verb (SOV) or subject-verb-object (SVO). Explaining this fact using psychological accounts of language acquisition or processing requires understanding how the present distribution has resulted from ancestral distributions and the rates of change between orders. We show that Bayesian phylogenetics can provide quantitative answers to three important questions: how word orders are likely to change over time, which word orders were dominant historically, and whether strong inferences about the origins of syntax can be drawn from modern languages. We find that SOV to SVO change is more common than the reverse and VSO to SVO change is more common than VSO to SOV, and that if the seven language families we consider share a common ancestor then that common ancestor likely had SOV word order, but also that there are limits on how confidently we can make inferences about ancestral word order based on modern-day observations. These results shed new light on old questions from historical linguistics and provide clear targets for psychological explanations of word-order distributions.

language evolution | computational historical linguistics | psycholinguistics

The sentence "dog bites man" is easily understood by a speaker of English, but switching the order of the words either renders it incomprehensible ("dog man bites") or changes its meaning ("man bites dog"). Only one of the six possible orders of these three words—the subject (S), verb (V), and object (O)—is commonly used in English sentences. However, the ordering of these three elements is a fundamental component of syntax, and varies significantly across languages. English is an SVO language, but among the world's languages, the six possible basic word orders are, from most to least common: SOV (48%), SVO (41%), VSO (8%), VOS (2%), OVS (1%), and OSV (0%) (1, 2).

Several attempts have been made to explain the cross-linguistic word-order distribution in psychological terms (3–6). All of these efforts have proceeded under the assumption that SOV is somehow optimal (e.g., requiring minimum working memory for parsing), since it is the most common word order. This is a "functionalist" approach, where some word orders are claimed to better facilitate the transmission of information. However, several linguists have noted that change from SOV to SVO appears to have been more common historically than change from SVO to SOV (7–10). If we suppose that, on the whole, languages tend to change from less to more functional word orders, then we should seek theories by which SVO is psychologically superior to SOV, not the other way around. Under this view, SOV is predominant not due to greater functionality but as a vestige of even greater dominance in the past.

Traditional functionalist explanations also assume that VSO is less functional than both SOV and SVO, again based on current frequency. Interestingly, the linguistics literature contains inconsistent hypotheses about change between these orders. Li (8) and Givón (9) consider VSO to SVO change to be most typical, while Gell-Mann and Ruhlen (11) claim SVO to VSO is most common with "occasional reversion" from VSO to SVO. Vennemann (7) does not appear to propose a preference, permitting change back and forth between the two orders. Without a definitive answer to this question, we cannot completely assess the explanatory adequacy of any putative theory of word-order functionality.

Some authors have suggested that all present languages are descended from a common ancestor with SOV word order (10, 11). This hypothesis is interesting in light of results suggesting a preference for SOV in improvised gestural communication (12–15), although the "monogenesis" idea of one common ancestor for all modern languages is controversial (16). The idea that SOV to SVO change is preferred over the opposite change plays a central role in these claims about ancestral SOV word order; however, it usually takes the form of a bare assertion, not backed up by quantitative data.

Psychology and linguistics thus leave us with three important questions: how word orders are likely to change over time, which word orders were dominant historically, and whether strong inferences about the origins of syntax can be drawn from modern languages. Insight into questions like the first two is increasingly being sought using (iterated) artificial language learning experiments, which can help to identify learning biases (17–19). We instead address all three questions using statistical analysis of language data, while agreeing that these two approaches can complement one another (20). Specifically, we show how methods from Bayesian phylogenetics can shed light on these questions. Recent work has explored how similar methods might be used to determine the time depth and homeland of proto-Indo-European (21, 22) (the most recent common ancestor of all Indo-European languages), as well as a range of other problems in historical linguistics (23–28). This previous work has most often, but not always, focused on the lexicon of ancient languages, examining how words tend to change over time. Only two previous papers have dealt specifically with word order. Pagel (24) reconstructs the ancestral word order of Indo-European to be SOV and examines the frequent directions of change but does not consider any other language families. Dunn et al. (25) examine word-order change in four different language families, but they consider SV vs. VS and VO vs. OV as separate parameters, and do not attempt to infer ancestral values of either parameter. Our work is, to our knowledge, the first attempt to infer ancestral

Significance

Word order varies in a systematic way across languages. Psychological explanations have been proposed for this, but they are informed only by data on modern word-order frequencies and are unable to explain proposed patterns in historical word order change. We use novel computational techniques to make inferences about this historical change and the probable ancestral word orders of major language families. Our results simultaneously establish firm criteria for psychological explanation of word-order change and provide insight into the history of word order, informing theories of language evolution.

Author contributions: L.M. designed research; L.M. performed research; L.M. and T.L.G. contributed new reagents/analytic tools; L.M. and T.L.G. analyzed data; and L.M. and T.L.G. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1] To whom correspondence should be addressed. Email: luke@maurits.id.au.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1319042111/-/DCSupplemental.

word orders and frequent directions of word-order change for multiple language families simultaneously.

Model

We considered a total of 671 languages from the following families: Afro-Asiatic, Austronesian, Indo-European, Niger-Congo, Nilo-Saharan, Sino-Tibetan, and Trans-New Guinea. Family classifications and basic word-order data for each language were taken from the World Atlas of Language Structures (1). The number of languages with each basic word order in each family is given in *Supporting Information*. The distribution of word orders in this sample does not accurately reflect the global distribution: SVO is overrepresented, comprising 55% of the languages in our samples but 41% of all languages. This limitation and its implications are discussed in *Discussion*. For each family, we made inferences about word-order ancestry and change using two probabilistic generative models: one for phylogenetic trees and one for word orders at the leaves of trees. Further details are provided in *Materials and Methods*.

Generative Model for Trees. Trees were generated using neighbor joining (29), specifically the NINJA implementation (30). Neighbor joining constructs trees with n leaves from $n \times n$ matrices of pairwise distances, such that languages that are close together according to the matrix become closely related in the tree. We specified four procedures for generating distance matrices. Each method constructs a base matrix based on either geographical distance between language locations, expert genetic classification of languages (e.g., Swedish is not just Indo-European but also Germanic, North Germanic, and East Scandinavian), comparison of linguistic features, or a combination of these. Individual matrices were obtained by adding random noise to the base matrix. This process defines a probability distribution over trees, $P(T)$. Sampling from this distribution results in trees that feature some random variation but are still constrained (by the structure of the base matrix) to not vary too wildly from what data and expert opinion suggests is probable.

Generative Model for Word-Order Data. Our generative model for leaf word orders was based on an ensemble of continuous-time Markov chains (CTMCs). The state space of each chain consists of six values, corresponding to the six basic word orders. Transitions between states are probabilistic in two senses. First, the state that the chain will be in after a transition is a random variable that depends only on the state of the chain at the time of the transition, as per a regular Markov chain. Second, transitions do not occur at regular time steps but rather at randomly distributed times along a continuum. For example, if a CTMC begins at time $t = 0$, the first state transition may happen at time $t = 0.76$, the next at time $t = 2.45$, the next at $t = 3.88$, etc. CTMCs are commonly used in biology for estimating phylogenetic trees from present-day genetic data, or for inferring ancestral genetic traits using data and known trees; see ref. 31 for a summary.

The behavior of one of our CTMCs is characterized by 36 parameters: 25 probabilities $t_{ij}, i, j = 1, \ldots, 6, i \neq j$, denoting the probability of transitioning between any two states and six rate parameters $\lambda_1, \ldots, \lambda_6$, which control the waiting times between transitions. Each waiting time is sampled from an exponential probability distribution with rate parameter λ_i, where i is the state of the chain at the beginning of the wait time. These parameters can be represented as a single 6×6 "rate matrix" $Q = [q_{ij}]$, where:

$$q_{ij} = \begin{cases} -\lambda_i & if \quad i=j \\ \lambda_{it_{ij}} & otherwise \end{cases}. \qquad [1]$$

For any time $t \geq 0$, the matrix $P = \exp\,(tQ)$ gives the probability p_{ij} that a chain starting in state i will be in state j after a time delay of t.

We model the generation of word-order data for the leaf nodes of trees as follows. The word order of the tree's root language is selected uniformly at random. Then word orders are sampled for each descendent node of the root, from the probability distribution over the state of a CTMC l units of time after transitioning to order k, where l is the length of the branch connecting the node to the root and k is the root word order. This process continues downward to the leaves of the tree. At each branching point in the tree, the Markov chain splits into two chains that are in the same instantaneous state at the branching point but thereafter evolve independently of one another, according to a single parameter matrix Q. This process induces a probability distribution over word-order data for a given tree and rate matrix, $P(D|T, Q)$.

Inference. Given known word orders D for languages at the leaves of a tree T, we used Bayes' theorem to invert the generative model for leaf data and compute $P(Q|T, D) \propto P(D|T, Q)P(Q)$. We used the Metropolis–Hastings algorithm (32, 33) to sample from the posterior distribution over rate matrices. The posterior mean matrix tells us about the dynamics of word-order change. The sampled matrices can also be used with belief propagation (34) to assess the evidence that present word-order data provides about each family's ancestral word order.

For each language family that we analyze, we apply the same inference algorithm to 100 different trees, sampled randomly from the distribution $P(T)$ induced by one of our generative models for trees. We draw 1,000 samples of Q for each tree. These are then pooled to give 100,000 samples total, and we use these to estimate posterior means, etc. This is an approximation to integrating out our uncertainty about the trees using $P(T)$ as a prior:

$$P(Q|D) = \int_T P(Q|T,D)P(T)dT \simeq \frac{1}{100}\sum_{i=1}^{100} P(Q|T_i,D), \qquad [2]$$

where the T_i are our sampled trees.

Because of this approach, our results do not depend critically on any one putative phylogenetic tree for any language family. We wish to emphasize the theoretical point that uncertainty about a family's phylogenetic tree need not translate into an equivalent amount of uncertainty about anything that may be inferred from that tree, such as historical change dynamics or the ancestral word order. It is possible that the observed data D and the high-level structure that is common to most or all plausible candidate trees can together provide enough constraints for the inference process to return some findings with high confidence. The approach of marginalizing over unknown details of tree topology, branch length, and change dynamics has been previously applied to estimating ancestral genetic traits in biology (35).

Results

Two decisions must be made before applying this inference scheme to our data. The first is which of the four tree-generating methods should be used. The second is whether to fit a separate Q matrix to each of the seven language families or use a common Q, assuming that the word-order change dynamics have been the same for all seven families over their collective lifetime. In *Supporting Information*, we perform model selection using the Bayesian information criterion (36), which prefers the "combination" trees with a common Q parameter. We present the results for this analysis here, and note in passing the extent to which each result is robust across the different approaches. *Supporting Information* includes further discussion of the relative merits of the different models, the results for other analyses, and various work intended to establish the reliability of our methods.

Word-Order Change Dynamics. Each sample of Q contains the rate parameters λ_i for each word order, from which we can compute the mean time that a language will remain in any given word order. Taken together, all of the samples define a posterior distribution over these mean times, and Fig. 1 shows these distributions. We find that SOV is the most diachronically stable word order, with posterior probability 0.63, followed by SVO with probability 0.37. VSO is the third most stable word order, and VOS, OVS, and OSV are all substantially less stable than all other orders. A similar picture holds for the "geographic" trees (SOV most stable with probability 0.55) and "genetic" trees (SOV probability 0.67). The probability of SOV being more stable than SVO is in all cases fairly low, so we cannot make this claim with certainty; however, we can claim with a high degree of confidence that SOV and SVO are both more stable than VSO and that VSO is more stable than the remaining three orders.

Each Q sample also provides information about the different transition probabilities, and in this regard, our results are much clearer. We find that SOV languages prefer changing to SVO over VSO, with posterior probability 0.84, and that changing to SVO is on average roughly 3.5 times more likely than changing to VSO. This is more consistent with the characterization of Gell-Mann and Ruhlen (11) (SOV → SVO → VSO) than with those of Li (8) and Givón (9) (SOV → VSO → SVO). However, we find Gell-Mann and Ruhlen's SVO → VSO transition to be only roughly half as probable as an SVO → SOV "reversion," with the reversion being more probable, with posterior probability 0.74. We find that VSO → SVO is preferred over VSO → SOV, with posterior probability 0.91, and that VSO changing to SVO is on average more than 4 times as likely as changing to SOV. This is consistent with Li and Givón's proposals as well as some comments by Gell-Mann and Ruhlen. Overall, we find that word-order change is best characterized as being dominated by slow cycles between SOV and SVO and faster cycles between SVO and VSO (these cycles are faster due to VSO's lower stability). In *Supporting Information*, we show that this characterization generally holds for the geographic, genetic, and feature trees as well.

Ancestral Word Orders. Using our recovered Q matrix, we can compute posterior distributions over the ancestral word orders for each language family, according to $P(W|D) \propto P(D|W)P(W)$, where $P(W)$ is some prior over ancestral orders (dependencies on T and Q omitted for clarity). If we use a uniform prior distribution over ancestors, such that the maximum posterior (MAP) ancestor is the one that assigns the maximum likelihood to the data, the MAP ancestors for Indo-European, Niger-Congo, Nilo-Saharan, and Sino-Tibetan are all OSV, while Afro-Asiatic is OVS, Austronesian is VOS, and Trans-New Guinea is SOV.

These findings are surprising, as OVS and OSV combined account for barely more than a single percent of today's languages. This result is due to the extreme instability of these orders (a similar phenomenon gives VOS unexpectedly high posterior probability). If a language tree that is thousands of years old has, say, OVS at its root, that OVS language is almost certain to have changed to an SOV or SVO language well before the time of the leaf languages, and these more stable orders can persist until the leaves. In this way, an ancestral OVS or OSV word order is no less able to explain the leaf data than an SOV or SVO order. Because of this, modern-day language data cannot reliably give us strong evidence either for or against ancestral VOS, OVS, or OSV word order.

Despite this finding, we should not believe that most of the families considered likely to have had ancestral OVS or OSV word order. Because these word orders were found to be highly unstable, we should in fact consider them to be very improbable ancestors for most families. We can incorporate this expectation into our analysis through the use of a nonuniform prior. We define a parameterized family of priors of the form $P_t(w_i) \propto \exp(-t\lambda_i)$.

For any particular value of t, P_t assigns to any word order a prior probability proportional to the probability that a language with that word order has not changed after t years. P_0 is the uniform distribution, and as $t \to \infty$, the prior probability of unstable word orders such as VOS, OVS, and OSV quickly tends toward zero. Fig. 2 shows how the posterior probability of different ancestral word orders for each family changes as we sweep through the family of priors, beginning at the uniform prior P_0 and stopping at $t = 10{,}000$ y. Observe that the posterior probability of the unstable orders drops rapidly for all families. For the majority of priors in the family, Afro-Asiatic, Indo-European, Sino-Tibetan, and Trans-New Guinea all have SOV as their most probable ancestral word orders. Niger-Congo's most probable ancestral word order is SVO, while Austronesian switches from being most probably VOS to VSO after $t \simeq 3{,}750$. Nilo-Saharan is the only family that does not develop a clearly preferred ancestral word order at any point. VSO has the highest posterior probability for most t values, but its probability never exceeds SOV's by more than 0.2, and the difference between these two orders shrinks to almost zero by $t = 10{,}000$. In *Supporting Information*, we consider a number of alternative priors to the family outlined above, and we also present a separate analysis of Indo-European using extinct language data.

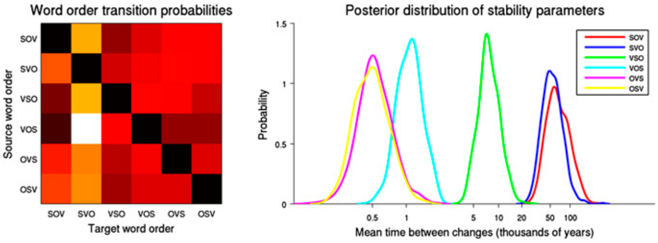

Fig. 1. Results of inferring a single mutation matrix Q for all six language families. (*Left*) Heat map showing the transition probabilities between word orders. Higher intensity (white, yellow) indicates more-probable transitions compared with lower intensity (red, brown), so SOV is most likely to transition to SVO and SVO to SOV. VSO is much more likely to transition to SVO than to SOV. (*Right*) Inferred posterior distributions of stability parameters for each word order. The horizontal axis shows the stability parameter, expressed as the mean time between transitions; i.e., higher values indicate a more stable word order.

At this point, SOV is a most probable ancestor for four of our seven families. This makes it the most widespread ancestral word order in our sample, but SOV ancestry is not ubiquitous. The evidence for ancestral VOS or VSO word order in Austronesian and SVO order in Niger-Congo is strong, while the evidence for SOV ancestry in Nilo-Saharan is not especially strong. However, an important theoretical point is that we have investigated the ancestors of our seven families independently so far, associating no penalty with different language families having widely differing ancestral word orders. If we take a polygenetic view, where each family's protolanguge was created de novo, this is fine. However, under a monogenetic view, the protolanguages of these seven families are siblings, descended from a common ancestor that could only have had a single word order. We should therefore not treat each family independently, and instead expect a degree of consistency across the families. This perspective is pursued in *Looking into the Past and Future*.

Looking into the Past and Future. In addition to answering questions about the ancestral word orders of language families, knowledge of Q lets us answer questions about the long-term behavior of word-order dynamics, looking into both the past and the future.

Previous researchers have made the (controversial) claim that SOV was the word order of a language that is a common ancestor for all present-day languages (10, 11), and our model allows us to perform quantitative investigation of this claim. We can join the roots of the seven family trees to a single parent node, creating a mongenetic "supertree." We can then compute a distribution over word orders at the common ancestor. Of course, this method is not able to establish that such a monogenetic relationship actually exists, but it lets us establish what could be said, and with how much confidence, about the common ancestor language if it does exist.

Fig. 3*A* shows how the posterior probability of different word orders at the root changes with the assumed age of the common ancestor. For young common ancestors (younger than about 48,000 y), VSO is the most likely ancestral word order. This is due to the high posterior probability of VSO for Afro-Asiatic, which is the oldest language family we consider and therefore has its root nearest to the common ancestor. For common ancestors any older than 48,000 y, the influence from Afro-Asiatic is decreased and SOV is the most probable word order. So, under a monogenetic view of language, "proto-human" was most probably SOV. However, this claim cannot be made with great confidence: The probability of an SOV common ancestor is never much higher than 0.25; put differently, the probability that the common ancestor had some word order other than SOV is never far below 0.75. This suggests that our ability to make strong claims about the word order of the earliest human languages is quite limited. SOV may be the safest bet for a common ancestral word order, but it is not an especially safe bet to take.

If the seven families had a common ancestor with SOV word order, this fact constrains the word order of each family's root language, just like the known word orders of the leaf languages do. If the common ancestor of all families is SOV, then SOV is also the MAP common ancestor of every family except Niger-Congo (which remains most probably SVO) for any common ancestor age less than 100,000 y. Thus, a monogenetic perspective, which introduces a bias for consistency among the ancestors of each family, suggests that SOV is the most probable ancestor of six of our seven families, rather than of four under a polygenetic perspective.

Regarding the future, we can explore how the present-day cross-linguistic distribution may be expected to evolve (see Fig. 3*B*). SOV and SVO are predicted to swap places as first and second most frequent word order after around a little less than 10,000 y. After around 35,000 y, there is relatively little change, as the Markov chain represented by our inferred Q slowly settles into a "stationary distribution," an unchanging distribution that depends only on the evolutionary dynamics represented by Q. The mean stationary distribution has SVO as its mode, at 52% (compare to ~41% today), followed by SOV at 41% (compare to ~48% today). Note that this is the only finding we present in which SVO is in any way privileged over SOV. Since SVO is the most frequent word order in our database, it is possible that this finding may be due to that bias. Of course, these predictions cannot and do not take into account possible, e.g., demographic, social, or political influences.

Discussion

Our primary motivation was to clarify the explanatory target for psychological explanations of word-order change. The traditional view has been that SOV must be more functional than SVO due to it being more common, and several explanations have been formulated based on this view. Our results suggest that a more nuanced view must be taken of cross-linguistic word-order distributions and how they come to be. We find only quite weak evidence that SOV is more stable than SVO, and find that cycling back and forth between these two orders is a common form of word-order change. We also find strong evidence that VSO languages appear to prefer changing to SVO over SOV. The overall picture does not seem consistent with SOV being more functional than SVO. We find that SOV is the most probable ancestral word order for four of our seven families, and that SOV and VSO are roughly equally probable for a fifth. We also find that a hypothetical common ancestor of all seven families is most likely to have had SOV word order (unless the ancestor was not much older than Afro-Asiatic). However, this is only slightly more probable than other ancestral orders. On the whole, it seems that the higher frequency of SOV in present languages is perhaps best attributed to widespread descent from ancestral SOV languages, while the high frequency of SVO seems to be due to preferred directions of word-order change and variation in word-order stability, both of which are presumably affected by functionality.

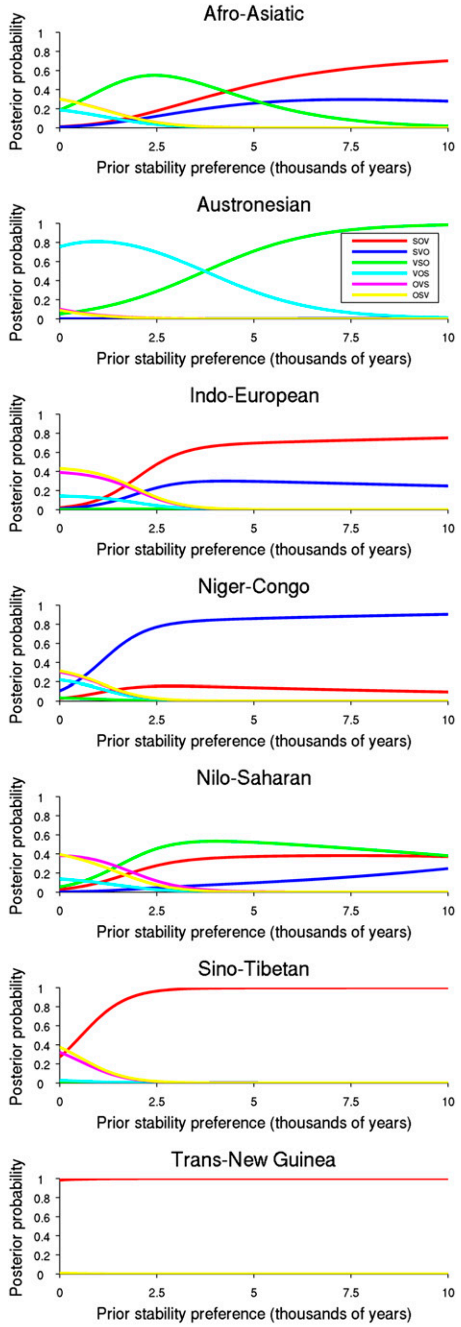

Fig. 2. Posterior probabilities of different ancestral word orders for each language family, under a parameterized family of priors. Each prior assigns word orders probabilities proportional to the probability of a language starting with that word order not changing after a certain length of time. The longer the timespan, the stronger the preference for stable ancestral word orders.

The implications of this work for psychological explanations of the distribution of word orders are that two independent lines of research should be pursued. One asks whether there is any psychological reason that SOV should be such a frequent ancestral word order. In light of other results showing a preference for SOV in improvised gestural communication (12–15), it seems likely that there is something special about SOV (this suggests

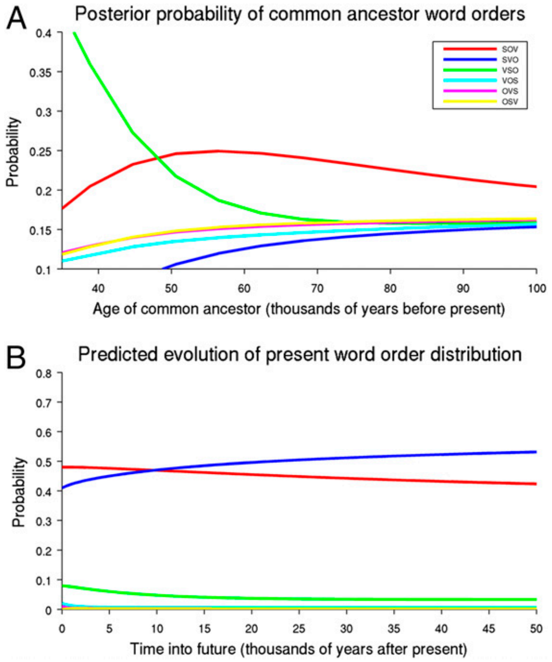

Fig. 3. Inferences about the distant past and future of the cross-linguistic word-order distribution. (*A*) How the posterior probability distribution over the word order of a hypothetical common ancestor for all six families changes as we increase the age of that ancestor (the prior distribution is uniform). (*B*) How the probability of a randomly selected language having a particular word order is expected to evolve from the present-day distribution.

that perhaps we should consider a nonuniform prior for ancestral word orders that favors SOV, which would make an SOV common ancestor for all seven families a safer bet). We should endeavor to learn what is special about SOV and why. The other line of research seeks to identify functional concerns that might drive the most frequent changes (SOV to SVO, SVO to both SOV and VSO, and VSO to SVO) as well as inhibit rare changes (SOV to VSO and VSO to SOV). Note that this schema of changes is not compatible with any straightforward ranking of the six orders by functionality, as has traditionally been assumed. Pursuing these two lines of research in parallel, informed by further computational analyses, provides a path toward understanding the roots of syntax.

The computational methodology presented here is a valuable tool for this research program, and others like it. Neighbor joining is fast compared with alternative methods like maximum parsimony and maximum likelihood, so our approach can scale up to very large datasets. The distance-based approach permits easy integration of linguistic data of many different types. However, as it is developed thus far, the method is not without limitations. The method is unable to handle family-specific variation in word-order change dynamics. Dunn et al. find evidence for such variation (25); however, see refs. 37 and 38 for critiques. The method also does not feature any possibility of word-order change due to language contact (analogous to "horizontal transmission" in biology), or any systematic variation in change rates due to, e.g., population size or the development of writing. Recent work has shown that contact-induced change can have significant influence on phonology (39), and this may also be the case for word order.

There is also the issue of our language sample and the nonrepresentativeness of its word-order distribution. A more representative sample poses methodological difficulties; most of the languages not in our seven families are SOV, as expected, but they belong to a large number of relatively small families. To get more SOV than SVO languages, we would need to add a great many more families, and make age estimates for each of these. It is unclear how much impact this sampling problem has on our results. SVO languages outnumber SOV in our sample, but our methods do not conclude that SVO is more stable than SOV. Presumably this is due to the family-level phylogenetic structure of the data: There are families in our sample where SOV is 10 times more common than SVO, and families where the two orders are roughly equally common. This suggests that including large families that capture the different family-level variation in the relative frequency of SOV and SVO may be more important than capturing the global statistics. Exploring how a wider range of families can be incorporated into a computational analysis of the history of word order is an important direction for future research.

Materials and Methods

Constructing Trees. Our trees were generated using neighbor joining (29), which required the construction of pairwise distance matrices. We used four separate methods for generating these matrices. The parameters of all four methods were calibrated using two reference trees estimated from cognate data and used in previous research. The reference trees were for the Austronesian (40) and Indo-European (22) families.

The geographic method uses only the "great circle" distance between the location of each language as given by *The World Atlas of Language Structures* (1), normalized so that the most distant languages in any family have a distance of 1.0. We then take the logarithm of these distances and find a and b such that $a + b \log(d(l_1, l_2))$ has a cumulative density function (CDF) that best matches the CDF of the normalized pairwise distances for the Austronesian and Indo-European reference trees, as measured by the Kolmogorov–Smirnov coefficient. The resulting pairwise distances have correlations of 0.46 and 0.65 with the Austronesian (Au) and Indo-European (IE) reference trees, respectively. Using the logarithmic distance yields a higher correlation than linear distance, consistent with previous findings (39).

The genetic method uses genetic classifications of languages taken from Ethnologue (41). The classifications are used to assign a "genetic distance" between languages as follows. Let language A be classified as $A_1 \supset A_2 \supset \ldots A_n$ and language B as $B_1 \supset B_2 \supset \ldots B_m$. Without loss of generality, let $n \leq m$. We discarded $B_{n+1}, \ldots, B_m$ if necessary. If $A_i = B_i$ for $i = 1, \ldots, k$, i.e., A and B are classified identically for the first k refinements of their family, then the distance is $M - \sum_{i=1}^{k} \alpha^i$, where $\alpha = 0.69$ and M is the maximum value the sum can take, i.e., when $k = n$, so that identically classified languages have a distance of zero. Since $\alpha < 1$, each additional matching refinement contributes less to the sum, reflecting the fact that later refinements tend to be much more fine grained. Values of α, a, and b were chosen to fit $a + bd(l_1, l_2)$'s CDF against the Au and IE reference trees, yielding correlations of 0.64 and 0.87.

The "feature" method is based on language feature data from WALS. Not all languages in WALS have a known value for every feature in the database. We identified 25 features that had at least one data point for either two Au or two IE languages in our reference trees. For any individual feature, we let the distance between two languages be a simple Hamming distance (i.e., 0 if the feature values are identical and 1 otherwise). If either language is missing data for the feature in question, we set the value to the mean distance for that feature in the relevant language family. We then performed a stochastic search through the 2^{25} possible combinations of these features, performing least squares regression on both reference trees. A set of ten features (listed in *Supporting Information*) maximized the lowest of the two correlations. We used weighted least squares regression with these features and the combined data of both reference trees, and then fit a and b in the usual way, yielding a pairwise distance measure with correlations of 0.29 and 0.49.

The "combination" method simply uses a linear combination of the three other methods, fitted against the reference trees using weighted least squares. The relative weights of the three methods are 0.18 for geographic, 0.77 for genetic, and 0.08 for feature. This combination yields correlations of 0.67 and 0.90 with the reference trees.

The methods described above yield one pairwise distance matrix for any given family, which we call the "base matrix." To generate a tree for a given method and family, we take the base matrix and add to each pairwise distance a random variate drawn from a $N(0, \sigma^2)$ distribution, with $\sigma = 2/30$ (so that the random noise almost never exceeds 20% of the maximum distance between any two languages). We then renormalize the matrix and pass it to the neighbor joining algorithm. We can repeat this process to sample as many distinct trees as we like.

The trees produced by neighbor joining were rooted at their midpoint. We then multiply each branch length by a change rate parameter, sampled from a lognormal distribution with a mean of 1.0 (following ref. 42). We interpret the resulting branch lengths as the product of the amount of time that separates two languages and the rate of word-order change during this time, as a proportion of the mean. This permits language change to occasionally occur faster or slower at certain places on the tree. After introducing rate variation, we scale all branch lengths such that the leaf farthest from each tree's root is at a distance corresponding to an estimate of the appropriate language family's age. The estimated language family age was randomly sampled for each tree instantiation, from a family-specific distribution. For all language families except Indo-European, tree ages were drawn from normal distributions, with means corresponding to published estimates of that family's age and variance set so that sampled ages rarely differ from the mean by more than 25%. The age estimates were 25,000 y for Afro-Asiatic (43), 7,000 y for Austronesian (44), 17,500 y for Niger-Congo and Nilo-Saharan (43), 7,500 y for Sino-Tibetan (45), and 8,000 y for Trans-New Guinea (46). Reflecting the controversy surrounding Indo-European dating, ages for IE trees were sampled from a sum of two normal distributions, one with mean 6,000 y (the Kurgan hypothesis of IE origins) and one with mean 8,750 y (the Anatolian hypothesis) (21). The DendroPy (47) and Newick (http://users-birc.au.dk/mailund/newick.html) software libraries were used for tree generation.

Inference. Given a tree T and a matrix Q, it is straightforward to compute the conditional probability of the known word-order data D, $P(D|T, Q)$. We began at the root of the tree and placed a uniform prior distribution over the word order at that node. Then we simply moved down the tree, computing a probability distribution for each node based on the parent node's distribution and rate matrix Q until we arrived at the leaves. Then $P(D|T, Q)$ was just the product of the probabilities of each individual leaf having taken its observed value.

We placed a prior distribution over Q that was unbiased with respect to the direction of transitions but was biased toward stability of word order. Recall that the 6 × 6 rate matrix can be expressed in terms of six waiting time rate parameters $\lambda_1, \ldots, \lambda_6$ and 30 transition probabilities t_{ij}. Higher values of λ_i make a word order less stable, so we put independent exponential prior distributions on each λ_i, with parameter values of 3.0. Thus, $P(Q) = 3^6 \exp\left(3\sum_{i=1}^{6}\lambda_i\right)$. With the prior defined, we can compute the posterior probability of a given matrix, $P(Q|T, D) \propto P(D|T, Q)P(Q)$.

The Metropolis–Hastings algorithm allowed us to draw samples from the posterior and compute an approximation to the posterior mean. Our proposal step operated directly on Q and randomly selected from a set of moves including adding normally distributed values of mean zero to randomly selected matrix elements and swapping randomly selected rows or columns. For each tree, we collected 1,000 samples after a burn-in of 5,000 samples, with an intersample lag of 100 iterations.

ACKNOWLEDGMENTS. We thank Russell Gray for providing the Indo-European and Austronesian reference trees, Will Chang, David Hall, and Stephan Meylan for comments on an early version of this paper, and three anonymous reviewers for valuable feedback. This work was supported by National Science Foundation Grant IIS-1018733.

1. Dryer M, Haspelmath M, eds (2011) *The World Atlas of Language Structures Online* (Max Planck Digital Library, Munich).
2. Comrie B (1981) *Language Universals and Linguistic Typology* (Blackwell, Oxford).
3. Mallinson G, Blake BJ (1981) *Language Typology* (North-Holland, Amsterdam).
4. Krupa V (1982) Syntactic typology and linearization. *Language* 58(3):639–645.
5. Manning AD, Parker F (1989) The SOV> . . . > OSV frequency hierarchy. *Lang Sci* 11(1):43–65.
6. Tomlin RS (1986) *Basic Word Order: Functional Principles* (Croom Helm, London).
7. Vennemann T (1973) Explanation in syntax. *Syntax Semant* 2:1–50.
8. Li CN, ed (1977) *Mechanisms of Syntactic Change* (Univ of Texas Press, Austin).
9. Givón T (1979) *On Understanding Grammar* (Academic, London).
10. Newmeyer FJ (2000) On the reconstruction of 'proto-world' word order. *The Evolutionary Emergency of Language*, eds Knight C, Studdert-Kennedy M, Hurford JR (Cambridge Univ Press, Cambridge, UK), pp 372–388.
11. Gell-Mann M, Ruhlen M (2011) The origin and evolution of word order. *Proc Natl Acad Sci USA* 108(42):17290–17295.
12. Gershkoff-Stowe L, Goldin-Medow S (2002) Is there a natural order for expressing semantic relations? *Cognit Psychol* 45(3):375–412.
13. Sandler W, Meir I, Padden C, Aronoff M (2005) The emergence of grammar: Systematic structure in a new language. *Proc Natl Acad Sci USA* 102(7):2661–2665.
14. Goldin-Meadow S, So WC, Ozyürek A, Mylander C (2008) The natural order of events: How speakers of different languages represent events nonverbally. *Proc Natl Acad Sci USA* 105(27):9163–9168.
15. Langus A, Nespor M (2010) Cognitive systems struggling for word order. *Cognit Psychol* 60(4):291–318.
16. Nichols J (2011) Monogenesis or polygenesis: A single ancestral language for all humanity? *The Oxford Handbook of Language Evolution*, eds Gibson KR, Tallerman M (Oxford Univ Press, Oxford), pp 558–572.
17. Tily H, Frank M, Jaeger TF (2011) The learnability of constructed languages reflects typological patterns. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (Cognitive Science Society, Austin, TX), pp 1364–1369.
18. Fedzechkina M, Jaeger TF, Newport EL (2012) Language learners restructure their input to facilitate efficient communication. *Proc Natl Acad Sci USA* 109(44): 17897–17902.
19. Culbertson J, Smolensky P, Wilson C (2013) Cognitive biases, linguistic universals, and constraint-based grammar learning. *Top Cogn Sci* 5(3):392–424.
20. Tily H, Jaeger TF (2011) Complementing quantitative typology with behavioral approaches: Evidence for typological universals. *Linguist Typol* 15(2):497–508.
21. Gray RD, Atkinson QD (2003) Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 426(6965):435–439.
22. Bouckaert R, et al. (2012) Mapping the origins and expansion of the Indo-European language family. *Science* 337(6097):957–960.
23. Dunn M, Terrill A, Reesink G, Foley RA, Levinson SC (2005) Structural phylogenetics and the reconstruction of ancient language history. *Science* 309(5743):2072–2075.
24. Pagel M (2009) Human language as a culturally transmitted replicator. *Nat Rev Genet* 10(6):405–415.
25. Dunn M, Greenhill SJ, Levinson SC, Gray RD (2011) Evolved structure of language shows lineage-specific trends in word-order universals. *Nature* 473(7345):79–82.
26. Dediu D, Levinson SC (2012) Abstract profiles of structural stability point to universal tendencies, family-specific factors, and ancient connections between languages. *PLoS ONE* 7(9):e45198.
27. Bouchard-Côté A, Hall D, Griffiths TL, Klein D (2013) Automated reconstruction of ancient languages using probabilistic models of sound change. *Proc Natl Acad Sci USA* 110(11):4224–4229.
28. Pagel M, Atkinson QD, Calude AS, Meade A (2013) Ultraconserved words point to deep language ancestry across Eurasia. *Proc Natl Acad Sci USA* 110(21): 8471–8476.
29. Saitou N, Nei M (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4(4):406–425.
30. Wheeler TJ (2009) Large-scale neighbor-joining with NINJA. *Proceedings of the 9th Workshop on Algorithms in Bioinformatics* (Springer, Berlin), pp 375–389.
31. Felstenstein J (2004) *Inferring Phylogenies* (Sinauer, Sunderland, MA), pp 196–221.
32. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equations of state calculations by fast computing machines. *J Chem Phys* 21(6):1087–1092.
33. Hastings WK (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57(1):97–109.
34. Pearl J (1982) Reverend Bayes on inference engines: A distributed hierarchical approach. *Proceedings of the American Association of Artificial Intelligence National Conference on AI* (Am Assoc of Artificial Intelligence, Pittsburgh), pp 133–136.
35. Huelsenbeck JP, Bollback JP (2001) Empirical and hierarchical Bayesian estimation of ancestral states. *Syst Biol* 50(3):351–366.
36. Schwarz G (1978) Estimating the dimension of a model. *Ann Stat* 6(2):461–464.
37. Levy R, Daumé H III (2011) Computational methods are invaluable for typology, but the models must match the questions: Commentary on Dunn et al. (2011). *Linguist Typol* 15(2):393–399.
38. Croft W, Bhattacharya T, Kleinschmidt D, Smith DE, Jaeger TF (2011) Greenbergian universals, diachrony and statistical analyses. *Linguist. Typol* 15(2):433–453.
39. Jaeger TF, Graff P, Croft W, Pontillo D (2011) Mixed effect models for genetic and areal dependencies in linguistic typology. *Linguist Typol* 15(2):281–320.
40. Gray RD, Drummond AJ, Greenhill SJ (2009) Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* 323(5913):479–483.
41. Lewis MP, Simons GF, Fennig CD (2013) *Ethnologue: Languages of the World* (SIL International, Dallas, TX), 17th Ed.
42. Drummond AJ, Ho SYW, Phillips MJ, Rambaut A (2006) Relaxed phylogenetics and dating with confidence. *PLoS Biol* 4(5):e88.
43. Ehret C (2000) *Language and History in African Languages: An Introduction*, eds Heine B, Nurse D (Cambridge Univ Press, Cambridge, UK).
44. Melton T, Clifford S, Martinson J, Batzer M, Stoneking M (1998) Genetic evidence for the proto-Austronesian homeland in Asia: mtDNA and nuclear DNA variation in Taiwanese aboriginal tribes. *Am J Hum Genet* 63(6):1807–1823.
45. Handel Z (2008) What is Sino-Tibetan? Snapshot of a field and a language family in flux. *Lang Linguist Compass* 2(3):422–441.
46. Pawley A (2005) The chequered career of the Trans New Guinea hypothesis. *Papuan Pasts*, eds Pawley A, Attenborough R, Golson J, Hide R (Pacific Linguistics, Canberra, Australia), pp 67–107.
47. Sukumaran J, Holder MT (2010) DendroPy: A Python library for phylogenetic computing. *Bioinformatics* 26(12):1569–1571.