Understanding Language Evolution Using an Event-Based Model

David M. Goldstein[1,+], **Shawn H. McCreight**[2,+] and **John P. Huelsenbeck**[3,+]

[1] Department of Linguistics, University of California, Los Angeles, UCLA, Los Angeles, CA 90095-1543, USA,
dgoldstein@humnet.ucla.edu
[2] Nytril LLC, 3060 San Pasqual St., Pasadena, CA 91107, USA,
shawn.mccreight@gmail.com
[3] Department of Integrative Biology, University of California, Berkeley, UC Berkeley, Berkeley, CA 94720, USA,
johnh@berkeley.edu
[+] these authors contributed equally to this work

Abstract

Linguistically phylogenies are standardly inferred on the basis of cognate relationships, which are discrete representations of ancestry. Although inference on the basis of such datasets has yielded important results, it suffers from an obvious fault: it ignores the phylogenetic signal in the segmental form of words. In this paper, we infer the phylogeny of Romance on the basis of segmental data...

Introduction

Modern languages are related to one another through a complicated history of divergence and word borrowing. The divergence of languages is caused by the slow change in spoken language as it is passed from parents to offspring. Over time, divergence causes languages to become increasingly different from one another, ultimately to the point where they are mutually unintelligible. Languages that were spoken by the same human group more recently in time are considered to be more closely related to each other than they are to groups that spoke the language more distantly in time; this relatedness information can be depicted by a tree-like diagram called a 'phylogeny.' Linguistic borrowing, by contrast, causes languages to become more similar to one another.

The phylogenetic relationships of languages are inferred from cognate words[1-7]— words that descend from a common ancestor, such as French *quatre*, Spanish *cuatro* and Romanian *patru*, all of which mean 'four.' The cognates that are used in a phylogenetic analysis of language are ones considered by linguists to be resistant to borrowing from other languages; words such as 'mother,' ' father,' and 'stone' are not commonly borrowed from other languages whereas words such as 'computer' or 'wi-fi' are more readily shared.

The study of language phylogeny depends critically not only on the choice of cognate words to use but also on the *coding* of these cognates so they can be read into software that was originally developed by biologists for the study of species phylogeny. As an example, consider how different words for the concept *hand* have been coded[3] (see Table 1, which are presented with phonemic IPA representations). Spanish *mano*, French *main*, and Italian *mano* all descend from a Latin ancestor *manus*. Words assigned the same state on the basis of shared segmental correspondences among words in a set of languages (which is part of a process linguists call the comparative method[8]). On the basis of such correspondences historical linguists identify words that descend from a common ancestor. Such decisions, even when well-informed, can significantly influence the results of a phylogenetic analysis.

Language	IPA	Coding
English	/hænd/	0
German	/hant/	0
French	/mɛ̃/	1
Spanish	/mano/	1
Italian	/maːno/	1
Russian	/rʊka/	2
Polish	/rɛŋka/	2

Table 1. Coding of lexical cognates for the word *hand*.

The coding procedure forces the linguist to treat the cognate word data in the same way biologists treat morphological characters in a phylogenetic analysis[9]. Consequently, linguistic phylogenetic analyses share the limitations of morphological phylogenetic analyses in biology. For one, because the state labels (0, 1, 2, *etc.*) are arbitrary, the linguist is limited to models that have a certain symmetry, so the probability of the observations will be the same regardless of the state label assignment. Moreover, the linguist is unlikely to include cognate words coded such that all of the languages have the same state. For this reason, one must condition on such words not finding their way into the data set in the first place. Finally, the state labels from one word to another have different inherent meanings. The state 0 from one cognate word is not equivalent to the state 0 from another word. For this reason, linguistic phylogenetic analysis is limited to estimating the language tree, and to some extent the divergence times between languages. Traditional phylogenetic analysis of linguistic clades provides little insight into how sounds change over time; it only models changes in the basic vocabulary of a language. When a set of languages share the same cognate state, phylogenetic inference is not possible. The differing states in Table 1 will distinguish the Romance languages (French, Spanish, and Italian) from the Slavic languages (Russian and Polish), but they have nothing to say about the relationships between the languages in these two clades. The segmental form of the words for 'hand' does, however, contain a phylogenetic signal. French /mẽ/ has, for instance, undergone more change than either the Spanish or Italian cognate form /mano/.

In this study, we treat the observations as the individual segments of words[10]. Specifically, we use phonemic representations coded with the International Phonetic Alphabet (IPA). We analyze the IPA information using a continuous-time Markov model that allows one of three events to occur in an instant of time[11] (1) a transition from one word segment to another; (2) the insertion of a single word segment; or (3) the deletion of a word segment. Our treatment allows us to analyze data in a manner more akin to molecular phylogenetic analyses in biology. Just as nucleotides are considered equivalent states regardless of their position in the genome, here we consider word segments to be equivalent across words. This means that we can learn about the rates of individual events by pooling information across different words. Moreover, our analysis allows us to use a richer set of models, none of which are limited by considerations of the labeling assigned to the states.

Our treatment of cognate words introduces several challenges, the most serious of which is establishing how individual word segments for a cognate are related. In a molecular phylogenetic analysis, the fine-scale homology of nucleotides sampled from different species is established using sequence alignment programs. However, even for molecular data in which long nucleotide sequences are used, there can be substantial uncertainty in the alignment[12]; alignment uncertainty is exacerbated for the cognate data because words typically have only a handful of word segments. We address this problem by marginalizing over word segment alignments. (That is, we consider all word alignments, weighting each by its probability under the model.) We do this by performing parameter estimation in a Bayesian framework, using Markov chain Monte Carlo[13-14] (MCMC) to sample model parameters, including word segment alignments, in proportion to their posterior probabilities.

The framework we develop allows the linguist to not only understand the phylogenetic relationships of languages, but to also learn about how words transform over time. We illustrate these points using a data set of 133 cognate classes comprising 1003 individual words from the Romance languages. Our approach lends itself to the study of large data sets; frees the linguist from the onerous and potentially error-laden task of coding states; and treats the observations in a manner that is more faithful to how the languages are spoken.

Methods

We use ancestral classes, in which the descendant forms are segmentally aligned. It is important to note that our definition of cognate refers only to segmental descent. It takes no account of semantics whatsoever. So the lexical items for 'ear' in Romance are paired with the Latin word *auricula* which is the diminute form of 'ear.'

Other things we have to mention:

1. The data are surface forms, not underlying forms.
2. We used the accusative singular for nouns in Latin.

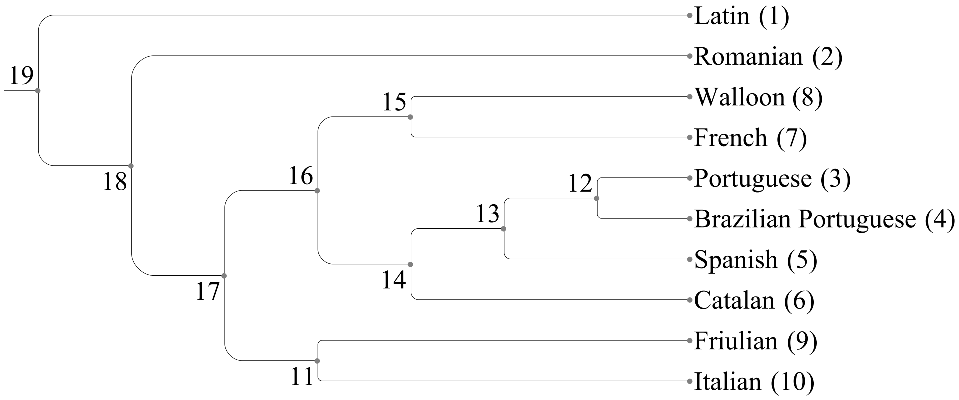
Here, we describe the details of the modeling assumptions we make for the linguistic characters, how we estimate the parameters of the linguists model, and how different models describing the evolution of words can be compared in a statistical framework.

Phylogeny relating languages. We assume that modern-day languages are related by an unknown evolutionary tree, called a 'Phylogeny' and denoted Ψ. The phylogeny contains information on the topological relationships among N sampled languages as well as information on the divergence times of the languages or on the amount of change that occurred between the languages.

Figure 1 shows an example of a phylogeny for $N = 10$ languages. In the terminology of evolutionary biology, the tree is composed of 'nodes' and 'branches.' (By contrast, mathematicians call nodes and branches 'vertices' and 'edges,' respectively.) The nodes represent the tips of the tree, each of which is assigned a language, and the points on the tree where the languages diverge from one another. Each language-assigned tip node is labeled $1, 2, ..., N$. The interior nodes are labeled $N + 1, N + 2, ..., 2N - 1$ in preorder sequence (i.e., ordered sequentially from the tips to the root). The root node is always assigned the label $2N - 1$. We denote the ancestor of node i as $\sigma(i)$. In the tree of Figure 1, the ancestor of node 3 is $\sigma(3) = 12$.

The branches connect the nodes of the tree and are represented as lines in Figure 1. The branch is assigned the label from its descendant node, so for example in the tree of Figure 1, the branch connecting nodes 12 and 13 is assigned the label 12.

Figure 1. An example tree showing the relationships of $N = 10$ languages.



A phylogeny is an information-rich graph. For one, it contains information on the relationships of the languages. This topological information is denoted τ. Figure 1 for example, suggests that French and Walloon are each others' closest relatives. They are more closely related to each other than they are to any another language on the tree because they share a more recent common ancestor with each other than they do with any other language. This common ancestor is the node numbered 15 in Figure 1. French and Walloon, together, are more closely related to Friulian in the tree of Figure 1. Both French and Walloon are equally related to Friulian because they both share the same common ancestor with Friulian, at the node numbered 9 in Figure 1. It is important to realize that there are many possible ways in which the languages can be related to one another, with the tree of Figure 1 depicting only one of the possibilities. In fact, for the case in which $N = 10$ languages are considered, there are 34459425 possible trees relating the languages. In general, the number of possible rooted trees is the product of the odd numbers up to, and including, $2N - 3$: $\mathscr{B}(N) = (2N - 3)!! = 1 \times 3 \times ... \times 2N - 3$. Each topology is given a unique label, $\tau_1, \tau_2, ..., \tau_{\mathscr{B}(N)}$. The number of possible tree topologies becomes quite large very quickly — it is a factorial, after all. A linguist interested in the relationships among $N = 60$ languages, for example, would contend with $5.86x10^{96}$ possible topologies, each depicting a unique and different way the languages can be related. For comparison, the number of atoms in the known universe is on the order of 10^{80}.

Ideally, the linguist would not only be able to estimate the correct topology relating the languages of interest, but also the times at which the languages diverged. The interior nodes of the tree represent language divergence events that occurred at specific times in the past and are denoted $t = (t_{N+1}, t_{N+2}, ..., t_{2N-1})$. The tip nodes are all assigned the time $t = 0$. Below, we will discuss in more detail a stochastic model of language change. However, the model we use, along with every other stochastic model for phylogenies, has an all-important parameter that describes the rate at which the language changes. This parameter is called the substitution rate and is denoted μ. Without external information that constrains the divergence times, such as one language divergence time that is considered known, perhaps from textual information, it is impossible to estimate the divergence times. The problem is that one obtains the same net divergence between two languages from a high rate of language evolution and a short divergence time separating the languages, or a low rate of language evolution and a long time separating the languages. In fact, the expected number of evolutionary events that occurred between two languages that diverged at time t is $\nu = 2t\mu$. (The factor of two is introduced because the path between the two languages is the time from one language to the common ancestor, and then back up the tree to the other language.) In this paper, we allow each of the $2N - 2$ branches of the tree to have an independent substitution rate. Hence, the expected number of evolutionary events that occur along the ith branch of the tree is $\nu = (t_{\sigma_i} - t_i) \times \mu_i$. In this study, we do not estimate the divergence times on the tree, but rather estimate the compound parameter representing the branch lengths (the ν_i which are in units of expected number of substitutions per segment (see below).

To summarize, we assume languages only diverge from one another, ignoring events such as word assimilation. We represent the divergence as a phylogeny containing information on both the topology and branch lengths, together denoted $\Psi = (\tau, \nu)$. One of the goals of this study is to estimate these parameters from the data collected from each language.

Data. The similarities of words from different languages are informative about how the languages are related. In this study, we use statistical methods developed in the field of evolutionary biology to estimate the relationships of species based on either the morphological characteristics of the species or the DNA sequences sampled from the same gene and compared across the species. The methods assume that the characteristics compared across species are homologous. Homology, in evolutionary biology, is similarity in some characteristic that is caused by common ancestry. Consider as an example the following DNA sequences sampled from three species,

Chimpanzee	AAGCTTCACCGGCGCAATTATCCTCATAATCGCCCACGGACTTACATCCT...
Gorilla	AAGCTTCACCGGCGCAGTTGTTCTTATAATTGCCCACGGACTTACATCAT...
Human	AAGCTTCACCGGCGCAGTCATTCTCATAATCGCCCACGGGCTTACATCCT...

These are partial mitochondrial sequences from Gojobori88[15]. In the original paper, the complete data had $N = 12$ primate species and the sequences were each $S = 898$ nucleotides in length. In a phylogenetic analysis of DNA sequenes, homology is assumed at two levels. First, one assumes that the sequences that are compared are homologous. Typically, homology at this level is established by sequence similarity and synteny of the gene (e.g., the gene that is compared across species is in the same, or at least similar, position along the chromosome when compared across species, which is another way of saying the gene that is compared has the same neighboring genes in all of the species in the analysis).

Not only must the DNA sequences be homologous, but the fine-scale homology of the sequences must also be established. The DNA sequences, above, are in an aligned form in which the fine-scale homology has been established; each column of the alignment is considered to be homologous. So, for example, the first column of the alignment which happens to be the nucleotide A in all three species is assumed to be homologous; it is assumed that the common ancestor of gorillas, chimpanzees, and humans had the same gene that also had a position that was homologous to the first column in the alignment. Fine-scale homology is established using computer programs in a process called ‘alignment.’ Importantly, phylogenetic methods assume that the homology established by the alignment program is correct.

Linguistic information, of course, is not like biological information. In the past, linguists attempted to find homologous words, called cognates. Typically, words are chosen that are thought to be resistant to assimilation. Variation in the cognate words is carefully scrutinized by the linguist and encoded in a way that computer software, developed with biological character data in mind, can read and produce sensible results. The encoding process produces variants on a cognate word with the variant states coded as 0 or 1 (or sometimes more, if there are more than two states for the word).

In this study, we take a different approach. Like others, we concentrate on the so-called ‘basic vocabulary’ of a language (SWADESH REFERENCE), since the lexical items that instantiate concepts in this domain are less prone to horizontal transmission (i.e., linguistic borrowing). In contrast to every study of linguistic phylogenetics that we are aware of, however, our investigation draws inferences from segmental information. For each concept in our dataset, homologous lexical items are assigned to the same class, which we refer to as a *cognate class*. The word forms within each cognate class are phonemic representations based on the International Phonetic Alphabet (IPA) (IPA REFERENCE). Consider the following word forms for one of the cognate classes for the concept ‘null’

Latin
French
Spanish
Italian
Brazilian Portuguese
Portuguese
Catalan
Walloon
Friulian
Romanian

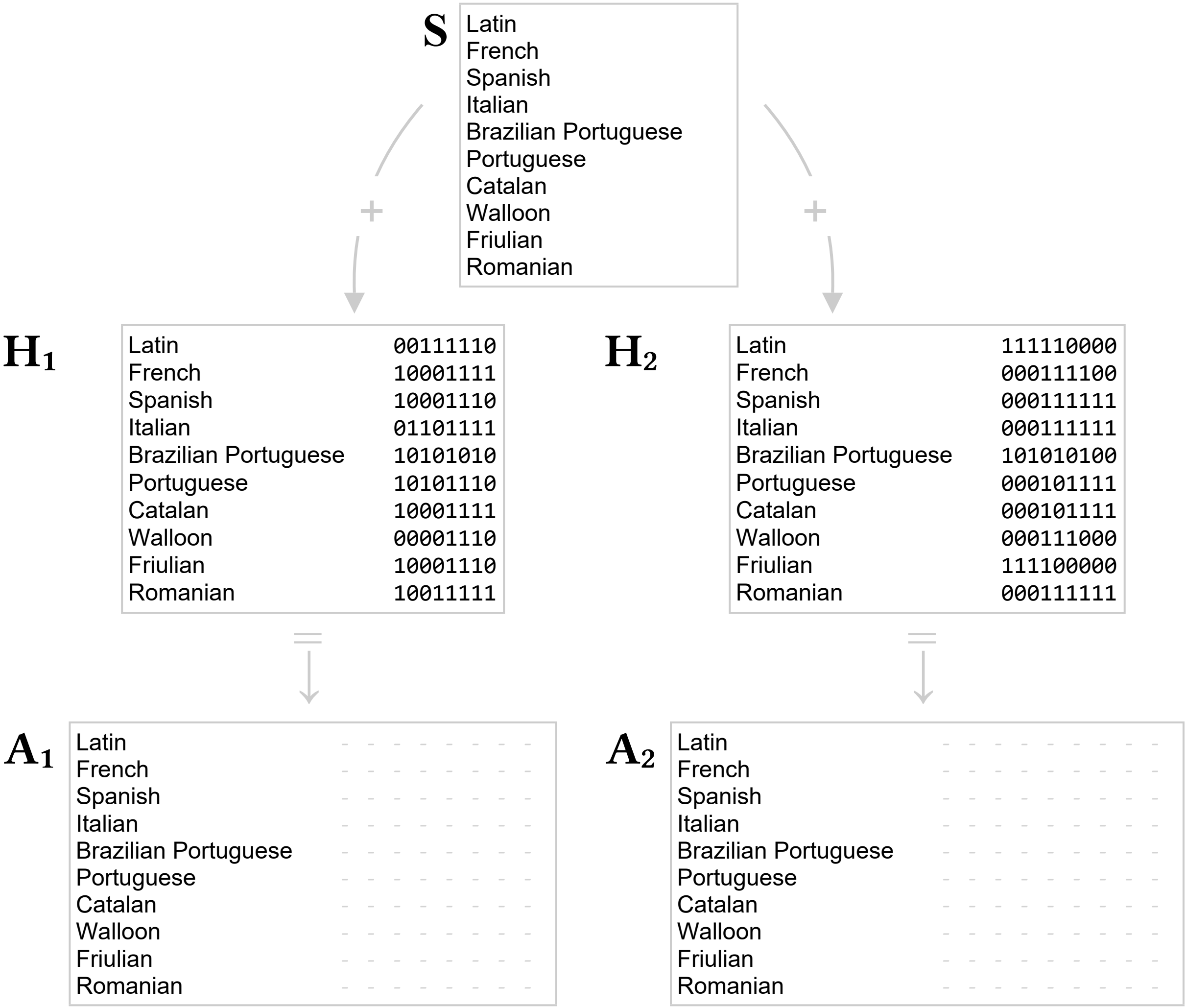
The word for ‘’ in Latin, for example, is //. The matrix for the word null also assumes that the fine-scale homology of the segments has been established. For example, the above matrix for the concept ‘null’ assumes that the first segment of the French, Portuguese, Catalan, Friulian and Romanian languages (respectively,) are homologous. The dashes indicate that there is no homologous segment at that potential position in the word. Note that the first segment for the word null in Italian is also . Why wasn't the first segment from Italian considered to be homologous to the first segments in French, Portuguese, Catalan, Friulian and Romanian? In this case, the alignment program chose the alignment that did not consider ‘’ of Italian as homologous to the ‘’ of French based on the settings chosen by the user. It may be that other alignments are nearly as good as the one that is illustrated. Figure 2 shows two possible alignments of the word null.

Note that a segmental alignment, denoted $\mathbf{A}$, is constructed by combining the word segment information for the languages of interest with information on the homology of the segments. The segmental information for the N languages of interest is denoted $\mathbf{S} = (\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_N)$, where $\mathbf{s}_i$ is the segmental string for the ith language. The segmental information for the $N = 10$ languages for the cognate class above for ‘null’ is:

Latin
French
Spanish
Italian
Brazilian Portuguese
Portuguese
Catalan
Walloon
Friulian
Romanian

Here, the segments for Latin would be $\mathbf{s}_1 = ()$. The alignment of the segments is accomplished by the use of a map, $\mathbf{H}$, describing the homology of the word segments. The alignment is formed by combining the word segment information with the homology map, $\mathbf{A} = (\mathbf{S}, \mathbf{H})$. Figure 2 shows an example of two alignments for the word null that can be formed using two different homology maps.

Figure 2. Alignments ($\mathbf{A}$) are formed from the observed segments ($\mathbf{S}$) and a homology map ($\mathbf{H}$).


S
Latin
French
Spanish
Italian
Brazilian Portuguese
Portuguese
Catalan
Walloon
Friulian
Romanian
+
+
H1
H2
=
=
A1
Latin - - - - - - - - -
French - - - - - - - - -
Spanish - - - - - - - - -
Italian - - - - - - - - -
Brazilian Portuguese - - - - - - - - -
Portuguese - - - - - - - - -
Catalan - - - - - - - - -
Walloon - - - - - - - - -
Friulian - - - - - - - - -
Romanian - - - - - - - - -
A2
Latin - - - - - - - - - -
French - - - - - - - - - -
Spanish - - - - - - - - - -
Italian - - - - - - - - - -
Brazilian Portuguese - - - - - - - - - -
Portuguese - - - - - - - - - -
Catalan - - - - - - - - - -
Walloon - - - - - - - - - -
Friulian - - - - - - - - - -
Romanian - - - - - - - - - -

	$\mathbf{H}_1$
Latin	00111110
French	10001111
Spanish	10001110
Italian	01101111
Brazilian Portuguese	10101010
Portuguese	10101110
Catalan	10001111
Walloon	00001110
Friulian	10001110
Romanian	10011111

	$\mathbf{H}_2$
Latin	111110000
French	000111100
Spanish	000111111
Italian	000111111
Brazilian Portuguese	101010100
Portuguese	000101111
Catalan	000101111
Walloon	000111000
Friulian	111100000
Romanian	000111111

Ultimately, the linguist observes the phonetic (in the case of contemporary languages) or graphemic (in the case of corpus languages) forms of words, on the basis of which phonemic representations are posited. The alignment, by contrast, is not directly observed. There are many different ways in which the segments from a cognate word can be homologous. The example from the word null, above, shows only one such way. In this study, we develop the statistical and analytical machinery that allow us to marginalize over the segmental alignments. Our method considers all possible segmental alignments of the word forms in a cognate class, weighting each such possibility by its probability under a model. In this way, our method does not condition on any specific segmental alignment.

Language evolution model. We assume that cognate words evolve along the branches of a phylogenetic tree through substitution of one segment by another, insertion of a new segment, or deletion of a segment.

> The process of linguistic change is more complex than this sentence allows. There are cases where entire words disappear and emerge. There are also cases in which the form of word can change with addition or deletion of a block of segments (called a morpheme). If we want to restrict our scope to the forms of change mentioned at the beginning of this paragraph (i.e., to segmental transitions), we can do that, but that will impact the data that I collect. We would also need to make it explicit that we are by design excluding certain types of well-known linguistic change.

Substitution of one segment for another is modeled using a continuous-time Markov model in which the possible states are the set of segments in the phonemic representations. At the heart of a continuous-time Markov chain is a rate matrix describing the rate of change between all pairs ofstates. As an example, consider a simplified Markov process with only five segments as states. The rates of change between the pairs of states can be represented in table form as

				To		
		A	B	C	D	E
	A	q_{AA}	q_{AB}	q_{AC}	q_{AD}	q_{AE}
	B	q_{BA}	q_{BB}	q_{BC}	q_{BD}	q_{BE}
From	C	q_{CA}	q_{CB}	q_{CC}	q_{CD}	q_{CE}
	D	q_{DA}	q_{DB}	q_{DC}	q_{DD}	q_{DE}
	E	q_{EA}	q_{EB}	q_{EC}	q_{ED}	q_{EE}

where $q_{ij} \geq 0 (i \neq j)$ is the rate of change from segment i to segment j. The diagonal elements of the rate matrix q_{ii} are specified such that each row sums to zero (i.e., $q_{ii} = -\sum_{j \neq i} q_{ij}$); this negative value can be interpreted as the rate at which the process moves away from state i. Note that the information on rates of change between all pairs of states is not typically represented in table form, but rather in matrix form as

$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} q_{AA} & q_{AB} & q_{AC} & q_{AD} & q_{AE} \\ q_{BA} & q_{BB} & q_{BC} & q_{BD} & q_{BE} \\ q_{CA} & q_{CB} & q_{CC} & q_{CD} & q_{CE} \\ q_{DA} & q_{DB} & q_{DC} & q_{DD} & q_{DE} \\ q_{EA} & q_{EB} & q_{EC} & q_{ED} & q_{EE} \end{pmatrix} \beta$$

Here, we introduce an additional parameter, β, that scales the rate matrix such that the average rate of segmental substitution is one.

A continuous-time Markov model has a simple physical interpretation. Specifically, when the process is in state i, one waits an exponentially-distributed time with parameter $-q_{ii}$ until the next segmental substitution occurs. The change, when it occurs, is to state j with probability $-q_{ij} / q_{ii}$. Several important quantities can be calculated using the information contained in the rate matrix, $\mathbf{Q}$. For one, the probability the process ends in state j conditional on starting in state i after a period ν can be calculated through exponentiation of the rate matrix, $\mathbf{P}(\nu) = e^{\mathbf{Q}\nu}$. One can also calculate the equilibrium distribution of the process --- denoted π and interpreted as the probability of capturing the process in a particular state after a very long time (formally, an infinite amount of time) has passed --- by solving the system of equations defined by $\pi\mathbf{Q} = \mathbf{0}$. Both the transition probabilities and equilibrium probability distribution play an important role in calculating the likelihood (see below).

Insertions and deletions of single segments occur at rates λ and μ, respectively. Consider, for a moment, the long-term behavior of a process in which $\lambda > \mu$. On average, segments would be inserted more frequently than they would be deleted. Word forms, then, would evolve to become a mouthful, growing without bounds. The opposite situation occurs when $\lambda < \mu$, in which segments are deleted at a higher rate than they are inserted. In this case, words would be whittled down to nothing; a language speaker would not find the words to describe anything, even important people in the person's life, such as the person who gave birth to him or her!

TKF91[11] described a model of DNA sequence evolution that allowed single nucleotides to be inserted and deleted at rates λ and μ. They introduced a convention for thinking about a DNA sequence in which nucleotides were connected by invisible links. Each nucleotide paired with the link to its right. The left-most nucleotide had to its left a special link, termed the immortal link. Importantly, when a nucleotide was inserted, it was inserted to the right of a link and brought along its own link (to its right). Deletions removed a nucleotide and the paired link. Importantly, the immortal link is never deleted. So, in the event that $\lambda < \mu$, the process does not actually go extinct. Rather, a nucleotide (and its link) can be inserted to to the right of the immortal link. In the model described by TKF91 they constrain the insertion rate to be less than the deletion rate $\lambda < \mu$. The equilibrium distribution of a sequence length is then geometrically-distributed with parameter λ / μ. We follow the convention of TKF91 in this study. In fact, the model we use is precisely the same as the TKF91 model, but with a different continuous-time Markov model used to describe segmental transitions (instead of the four-state process describing nucleotide substitutions used by Throne et al., 1991).

The overall substitution and insertion/deletion process can be described as follows. In a sequence n segments in length, of which n_i of them are of segment type i, the time until the next event occurs is exponentially-distributed with parameter

$$r = -\sum_i (n_i q_{ii}) + n\lambda + (n - 1)\mu$$

When an event occurs, it is a substitution with probability $-\sum_i (n_i q_{ii}) / r$, an insertion with probability $n\lambda / r$, and a deletion with probability $(n - 1)\mu / r$. Importantly, the process allows the alignment map $(\mathbf{H})$ to be treated as a parameter of the model.

Bayesian estimation of language evolution model parameters. We estimate the parameters of the language-evolution model in a Bayesian framework. Bayesians base inferences on the posterior probability distribution of a parameter, which can be calculated using Bayes' theorem as

$$P(\text{Parameter(s)} \mid \text{Observations}) = \frac{P(\text{Observations} \mid \text{Parameter(s)})\ P(\text{Parameter(s)})}{P(\text{Observations})}$$

where the vertical bar indicates a conditional statement. In words, the posterior probability distribution of the parameters is equal to the likelihood $[P(\text{Parameter(s)} \mid \text{Observations})]$ times the prior probability distribution $[P(\text{Parameter(s)})]$, divided by the marginal likelihood $[P(\text{Observations})]$.

In this study, parameters include:

$\tau_1, \ldots, \tau_{\mathcal{B}(N)}$	Tree Topologies
$\nu_1, \ldots, \nu_{\mathcal{B}(N)}$	Branch length parameters
θ	Parameters associated with the rate matrix **Q**
$\lambda, \mu\ (\lambda < \mu)$	The insertion and deletion rates of segments
H	The map describing the alignment of the segments

We assign prior probability distributions to all parameters of the model (see Table 1). The posterior probability distribution of the linguistic model parameters is then

$$P(\tau, \nu, \theta, \lambda, \mu | \mathbf{S}) = \frac{P(\mathbf{S} | \tau, \nu, \theta, \lambda, \mu)\ P(\tau, \nu, \theta, \lambda, \mu)}{P(\mathbf{S})}$$

Note that the likelihood is marginalized over all possible alignments,

$$P(\mathbf{S} | \tau, \nu, \theta, \lambda, \mu) = \sum_{\mathbf{H}} P(\mathbf{A} | \tau, \nu, \theta, \lambda, \mu)\ P(\mathbf{A} | \mathbf{H}, \mathbf{S})\ P(\mathbf{H})$$

where the sum is over all possible alignment maps, which implies that our inferences are not conditioned on any particular alignment of segments being correct. Similarly, the marginal likelihood accounts for all possible combinations of model parameters:

$$P(\mathbf{S}) = \sum_{\tau} \int_{\nu} \int_{\theta} \int_{\lambda < \mu} P(\mathbf{S} | \tau, \nu, \theta, \lambda, \mu)\ P(\tau, \nu, \theta, \lambda, \mu)\ d\nu\ d\theta\ d\lambda\ d\mu$$

where the integrals represent integration over all possible combinations of branch lengths, rate matrix parameters, and insertion/deletion rates.

We calculate the likelihood on a per-word basis using the algorithm described by Lunter et al. (2003) that conditions on an alignment. Although the posterior probability distribution can be written down, and individual components such as the prior probability or likelihood for a particular combination of parameters can be calculated, analytically solving the high dimensional summations and integrals required for the posterior probability is unfeasible. Instead, we numerically approximate the joint posterior probability distribution of the parameters using Markov chain Monte Carlo.

Markov chain Monte Carlo. The aim with Markov chain Monte Carlo (MCMC) is to construct a Markov chain that has as its possible states the parameter values of the statistical model and a stationary distribution that is the posterior probability distribution of the parameters. Metropolis et al. (1953)[13] and Hastings (1970)[14] described rules that allow the scientist to construct such a chain. When at stationarity, samples from this chain form valid, albeit dependent, samples from the posterior probability distribution. The

Metropolis-Hastings algorithm constructs the Markov chain using the following algorithm:

1. The current state of the chain is denoted θ. If this is the first cycle of the Markov chain, initialize θ, perhaps by choosing a value from the prior distribution.
2. Propose a new value for θ denoted θ'. The proposal mechanism is up to the programmer, but must involve the generation of random numbers u such that the proposed value is a function of the current value and the random numbers, $\theta' = \mathrm{h}(\theta, u)$. The probability of proposing the new value is $q(\theta \longrightarrow \theta)$whereas the probability of the imagined reverse move, not actually made in computer memory, is $q(\theta' \longrightarrow \theta)$
3. Calculate the probability of accepting the proposed value:
$$R = \min\left(1, \frac{f}{(X|\theta')f(X|\theta)} \times \frac{f}{(\theta')f(\theta)} \times \frac{q(\theta' \longrightarrow \theta)}{q(\theta \longrightarrow \theta')}\right)$$
In words, the acceptance probability is the product of the likelihood, prior, and proposal ratios.
4. Generate a uniform(0,1) random variable, u. If $u < R$, accept the proposed state, setting $\theta = \theta'$. Otherwise, the proposed state is said to be rejected and the chain remains in state θ
5. Return to Step # 1.

The proposals we implement in this study are all typical for phylogenetic models. (Some details here)The unique aspect of this study is a proposal mechanism for the alignments of the segments for various words. Here, we use the proposal mechanism described by Lunter et al. (2005).

(A bit on interpretation of MCMC results here.)

Model comparison. In a Bayesian analysis, parameter estimates are based on the joint posterior probability distribution of the parameters, which we numerically approximate using the Metropolis-Hastings algorithm. Often, however, the linguist is interested in the comparison of two or more models with the goal of evaluating which of the models best explains the observations. Bayesian model comparison is based on the marginal likelihoods of the models. Consider two different linguistics models, $\mathscr{M}_1$ and $\mathscr{M}_2$ with marginal likelihoods, $P(\mathbf{S}|\mathscr{M}_1)$ and $P(\mathbf{S}|\mathscr{M}_2)$ (note the marginal likelihoods are calculated for the same observations). The ratio of the marginal likelihoods,

$$BF_{12} = \frac{P(\mathbf{S}|\mathscr{M}_1)}{P(\mathbf{S}|\mathscr{M}_2)}$$

called the Bayes Factor, measures the relative support of the two models; a Bayes Factor less than one favors $\mathscr{M}_1$ whereas the oppose is true for a Bayes factor greater than one. Unlike in frequentist statistics, one does not obtain p-values in a Bayesian comparison of models. Rather, the Bayes Factor is interpreted as is, or on a log scale. Jeffreys (1961) provided a table to help with the interpretation of Bayes Factors:

BF	$\log_{10} BF$	Interpretation
1 - 3.2	0 - $\frac{1}{2}$	Not worth a bare mention
3.2 - 10	$\frac{1}{2}$ - 1	Substantial
10 - 100	1 - 2	Strong
> 100	> 2	Decisive

In a Bayesian analysis, there is no need to penalize parameter rich models for having more parameters. Rather, the penalization is built into the comparison; the additional parameters in a complicated model are each assigned a prior probability distribution. A parameter-rich model has lower prior probability for any combination of model parameters than a simpler model. Hence, there is no need to compare the Bayes factor to a null distribution as there is in frequentist hypothesis testing.

The main limitation of Bayesian model comparison is numerically approximating the marginal likelihoods of the models. This can be done in numerous ways. For example, one can construct a Markov chain that jumps between models, even if the models differ in dimensions, using a generalization of MCMC described by Green (1995; reversible-jump MCMC). Alternatively, one can numerically approximate the marginal likelihoods using what is called path-sampling in which numerous MCMC chains explore a path between the prior and posterior distributions (citations).

In this study, we compare two models. The first model is the simplest one we could devise, assuming that the rate of change between all segments is equal. Our first model is isomorphic to the earliest model of DNA substitution, called the Jukes and Cantor model in molecular evolution (Jukes and Cantor, 1969)[16]. The second model assumes that the rate of change to (something) is potentially different than the rate of change to (something else):

$\mathscr{M}_1$: $q_{ij} = \alpha$ (i.e., all elements of rate matrix are equal)

$$\mathscr{M}_2 : \quad q_{ij} = \begin{cases} \alpha \text{ if } i \longrightarrow j \text{ is a something} \\ \beta \text{ if } i \longrightarrow j \text{ is a something else} \end{cases}$$

Figure 3. For the ‘Linguistically Informed’ model, states were grouped into 12 sets. Here, the area of the circles is proportional to the occurance frequencies for each group. The width of the lines is proportional to the rates of transition between each partition.

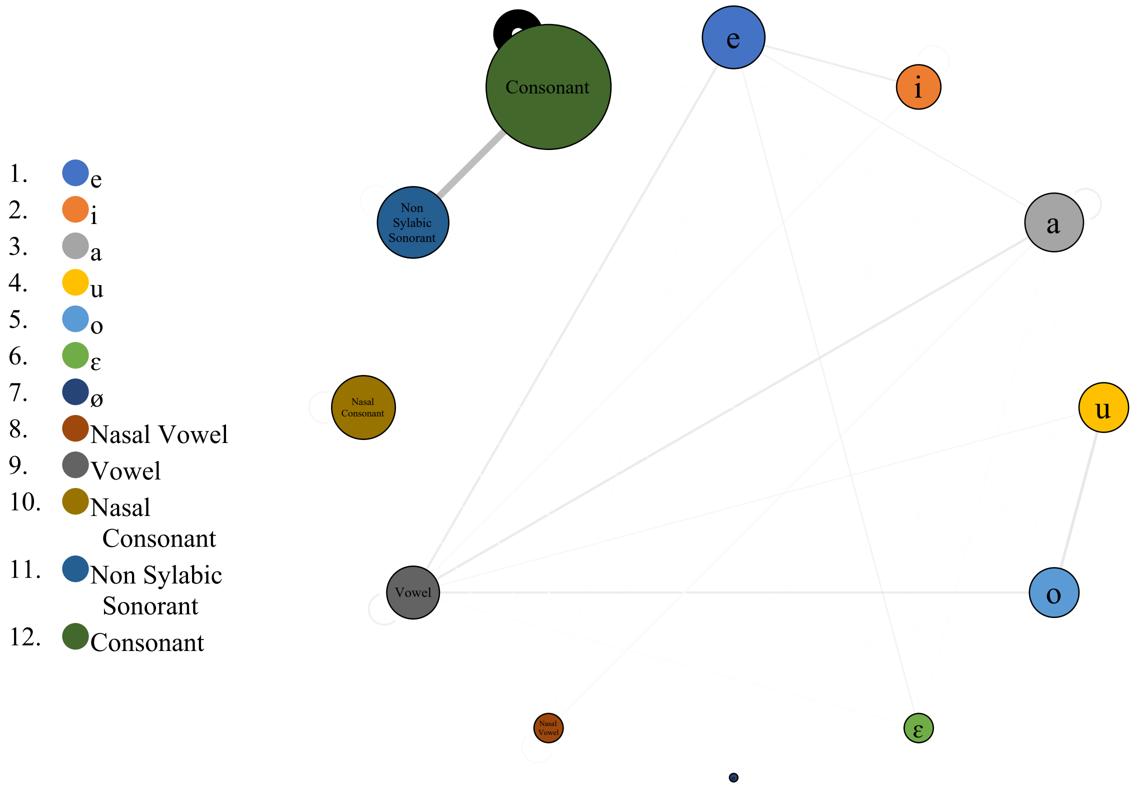
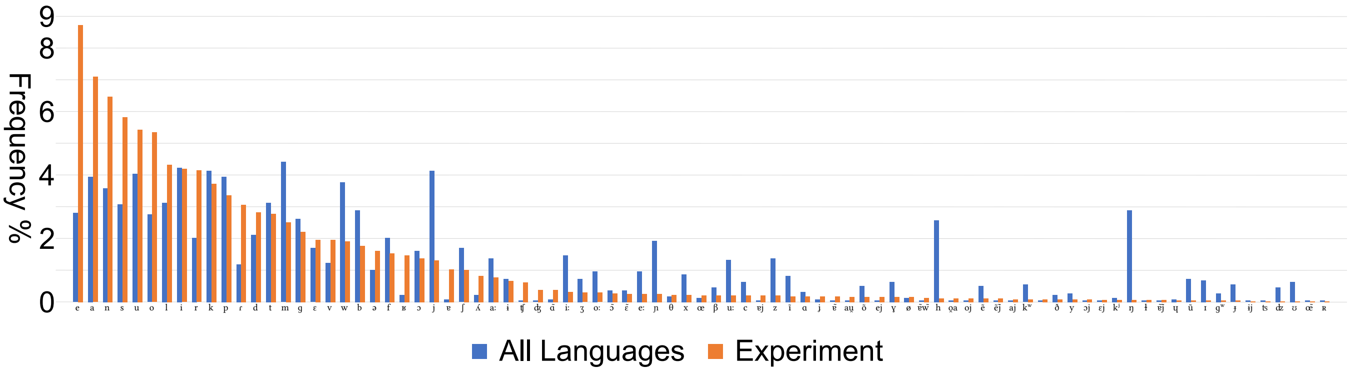


Figure 4. Frequency of occurance of segments in the lexicon[17]



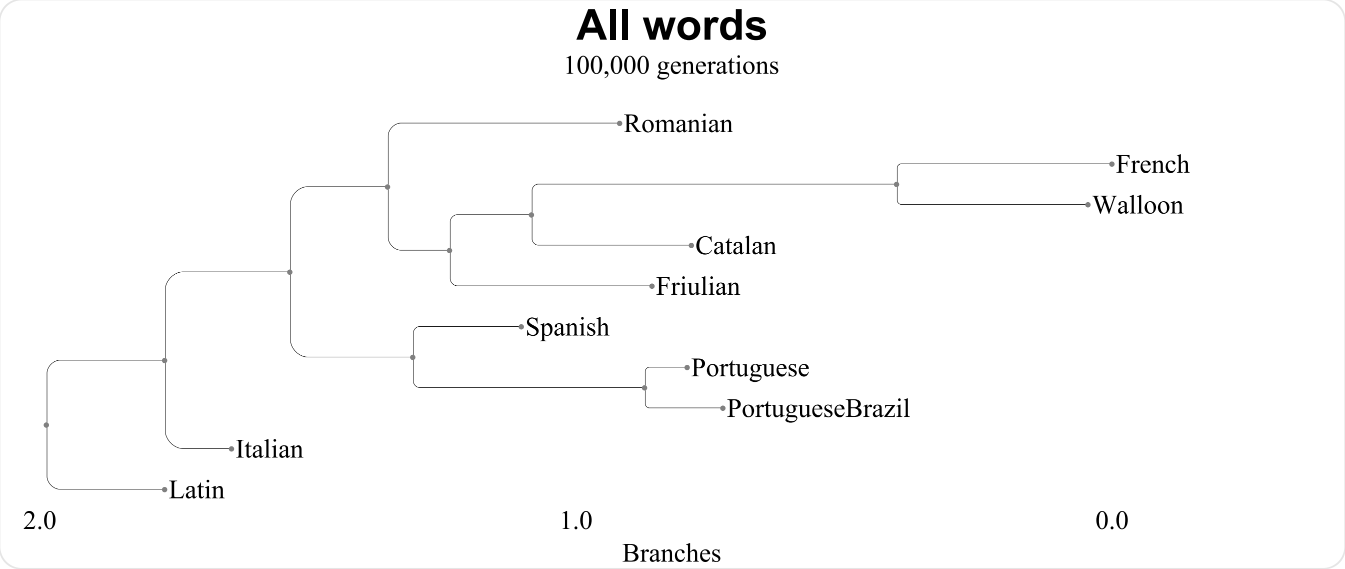
Conclusion

10 Languages
133 Concepts
85 Unique Segments

References

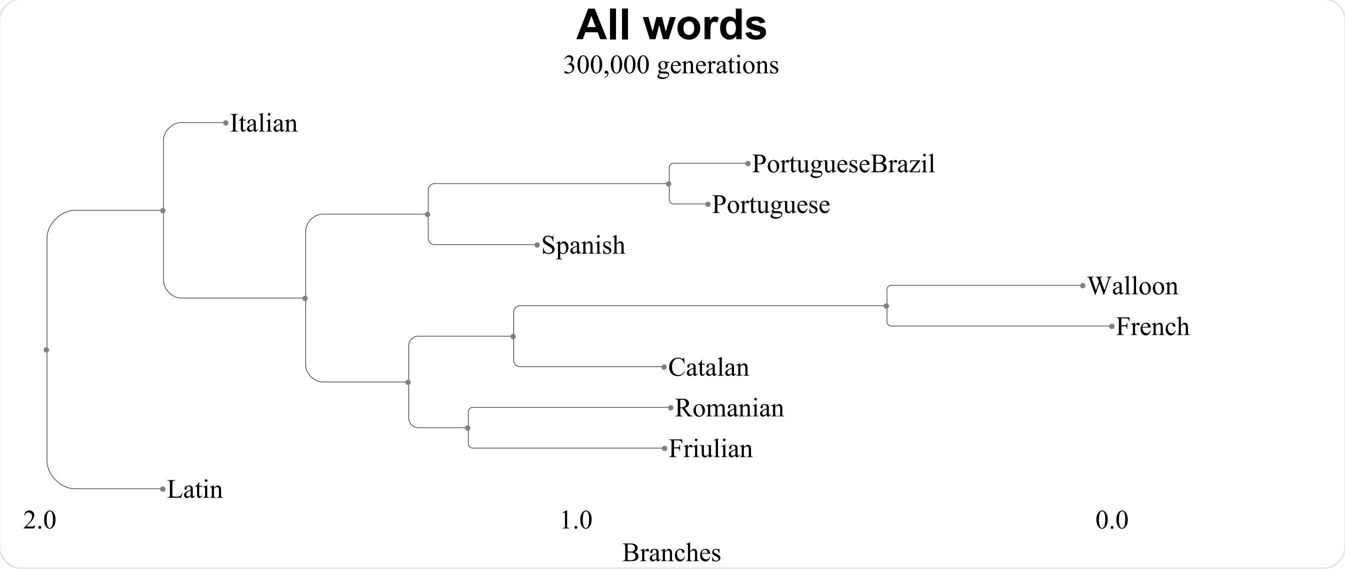
1. Gray, R. D. , and Jordan, F. M. Language trees support the express-train sequence of Austronesian expansion. *Nature* **405**, 1052-1055, DOI: 10.1038/35016575 (2000)
2. Holden, C. J. Bantu language trees reflect the spread of farming across sub-Saharan Africa. *Proceedings of the Royal Society B* **269**, 793-799, DOI: 10.1098/rspb.2002.1955 (2002)
3. Ringe, D. A. , Warnow, T. , and Taylor, A. Indo-European and computational cladistics.*Transactions of the Philological Society* **100**, 59-129, DOI: 10.1111/1467-968X.00091 (2002)
4. Gray, R. D. , and Atkinson, Q. D. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* **426**, 435-439, DOI: 10.1038/nature02029 (2003)
5. Bouckaert, R. R. , et al.Mapping the origins and expansion of the Indo-European language family. *Science* **337**, 957-960, DOI: 10.1126/science.1219669 (2012)
6. Bowern, C. , and Atkinson, Q. D. Computational phylogenetics and the internal structure of Pama-Nyungan. *Language* **88**, 817-845, DOI: 10.1353/lan.2012.0081 (2012)
7. Chang, W. , Cathcart, C. A. , Hall, D. P. , and Garrett, A. J. Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language* **91**, 194-244, DOI: 10.1353/lan.2015.0005 (2015)
8. Weiss, M. The comparative method. In Bowern, C. , and Evans, B. (Eds.) *The Routledge handbook of historical linguistics*, 127-145, DOI: 10.4324/9781315794013.ch4 (Routledge, London 2015)
9. Lewis, P. O. A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic Biology* **50**, 913-925 (2001)
10. Bouchard-Côté, A. , Hall, D. , Griffiths, T. L. , and Klein, D. Automated reconstruction of ancient languages using probabilistic models of sound change. *Proceedings of the National Academy of Sciences, U.S.A.* **110**, 4224-4229 (2013)
11. Thorne, J. L. , Kishino, H. , and Felsenstein, J. An evolutionary model for maximum likelihood alignment of DNA sequences. *Journal Of Molecular Evolution* **33**, 114-124 (1991)
12. Wong, K. M. , Suchard, M. A. , and Huelsenbeck, J. P. Alignment uncertainty and genomic analysis.*Science* **319**, 473-476 (2008)
13. Metropolis, N. , Rosenbluth, A. W. , Rosenbluth, M. N. , Teller, A. H. , and Teller, E. Equation of state calculations by fast computing machines. *Journal Of Chemical Physics* **21**, 1087-1092 (1953)
14. Hastings, W. K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97-109 (1970)
15. Gojobori, T. To be determined.(1988)
16. Jukes, T. H. , and Cantor, C. R. Evolution of protein molecules. In Munro, H. N. (Ed.), *Mammalian Protein Metabolism*, 21-123 (Academic Press, 1969)
17. PHOIBLE 2.0. In Moran, S. , and McCloy, D. (Eds.) Retrieved August 19, 2021 from www.phoible.org (Max Planck Institute for the Science of Human History, 2019)

Appendix 1 - Experimental Runs



All words

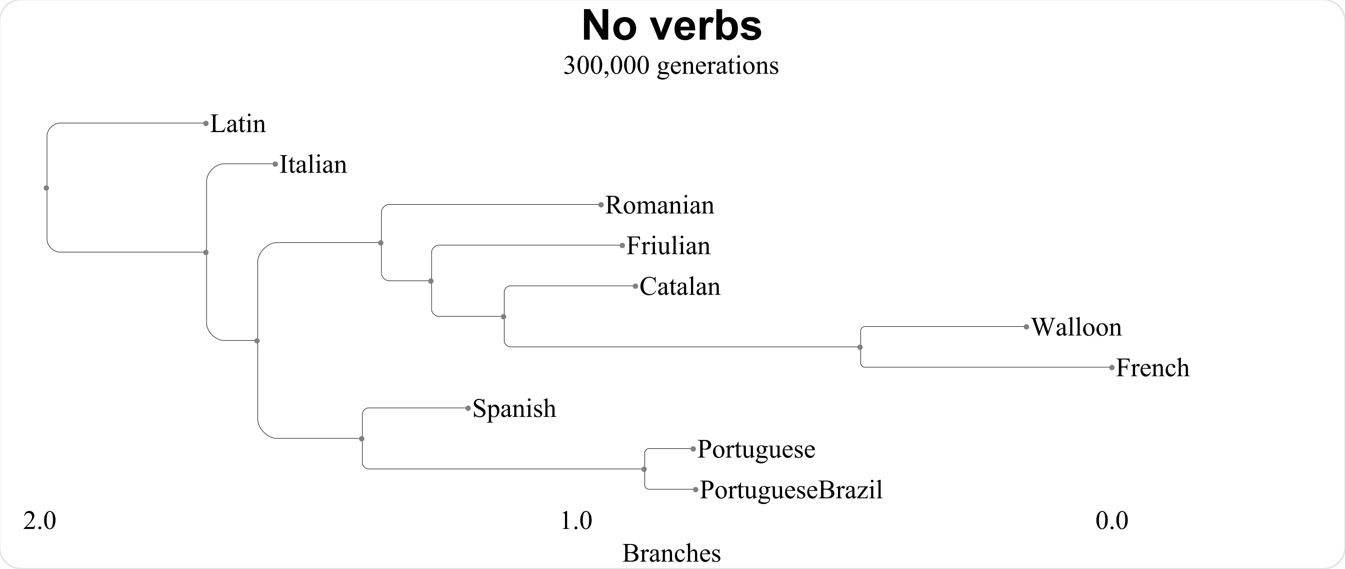
100,000 generations



All words

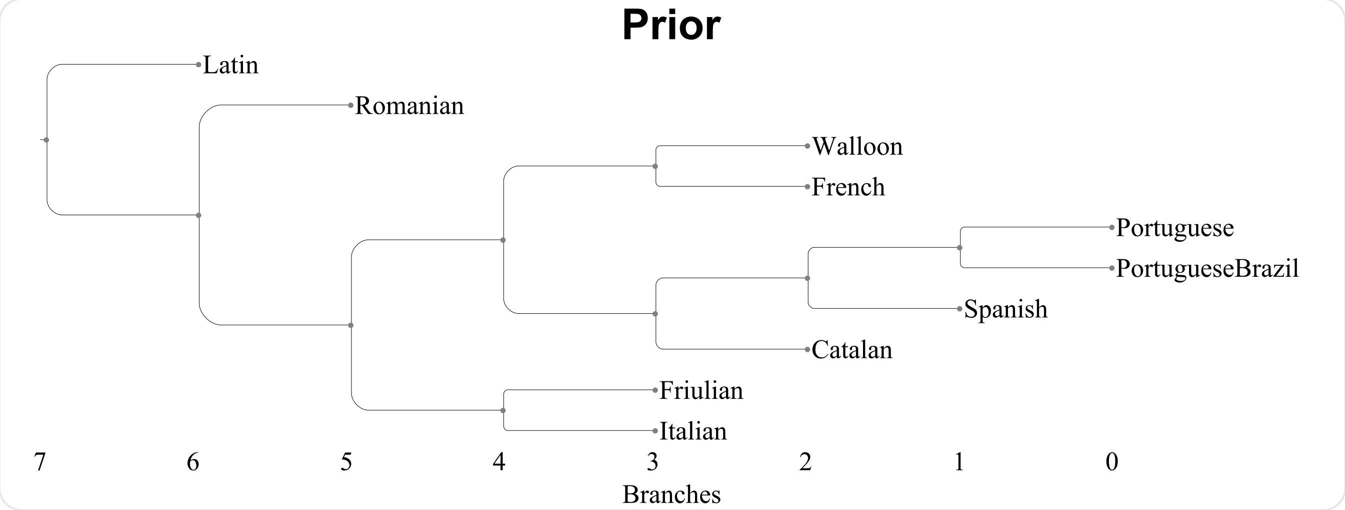
300,000 generations

Appendix 1 - Experimental Runs

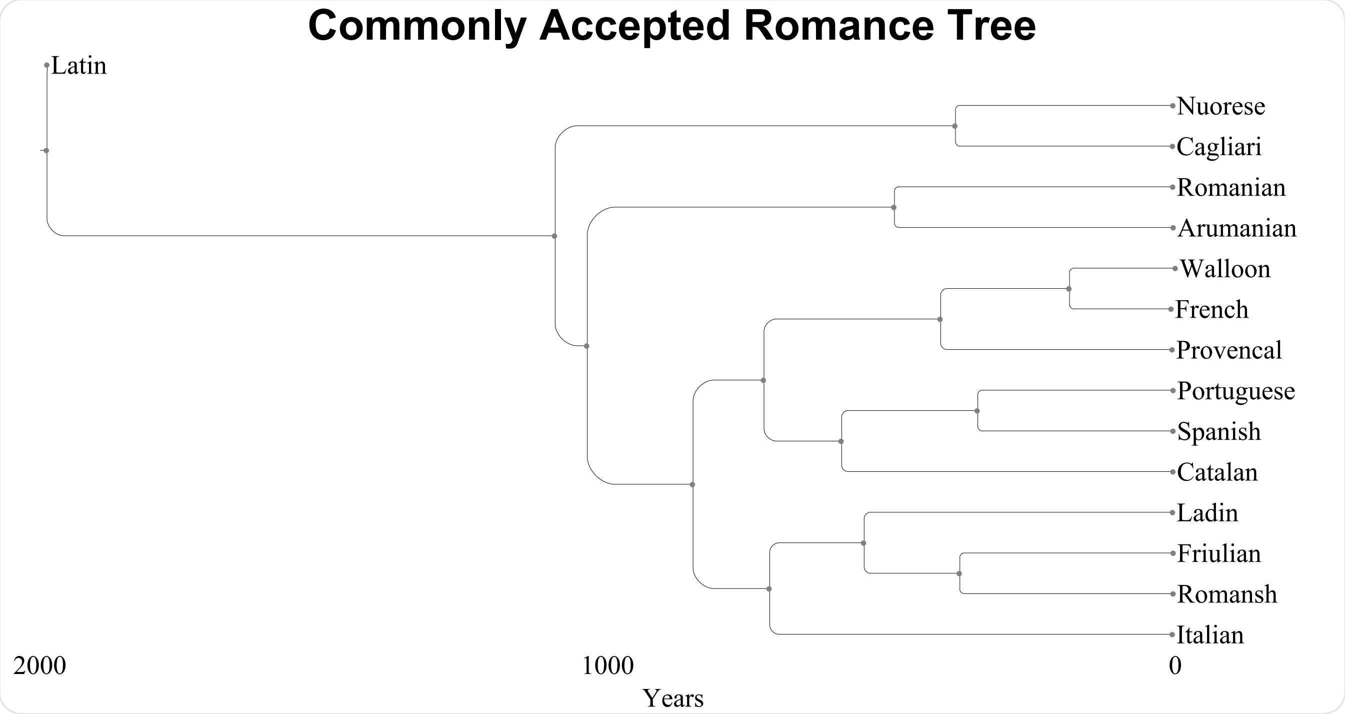


No verbs

300,000 generations



Prior


Commonly Accepted Romance Tree
Latin
Nuorese
Cagliari
Romanian
Arumanian
Walloon
French
Provencal
Portuguese
Spanish
Catalan
Ladin
Friulian
Romansh
Italian
2000
1000
0
Years

Appendix 2 - Partition Assignments

Model: “Linguistically Informed”

1) e	e eː
2) i	i iː
3) a	a aː
4) u	u uː
5) o	o oː
6) ɛ	ɛ
7) ø	ø
8) Nasal Vowel	ɑ̃ ĩ ɐ̃ ẽj̃ ɔ̃ õ ɛ̃ ɐ̃w̃ ẽ œ̃ ũ ɐ̃j̃
9) Vowel	ɨ ej ɐj au̯ ɐ ə ɨj ɔ œ ɔj o̯a oj ɑ ɛj ɪ e̯a aj y ʊ
10) Nasal Consonant	n m ɲ ŋ
11) Non Sylabic Sonorant	w l j r
12) Consonant	f t ʧ p s k θ ʃ h b v z ʎ x ɾ c d ʣ ʁ ʤ ɡʷ ɡ ʝ β ʀ kʲ ʒ ɟ ɣ ð kʷ ɥ ʦ ɫ