



# Simulated Segment Evolution using the TKF91 Model

D. M. Goldstein, S. H. McCreight and J. P. Huelsenbeck

## 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

In this paper, we attempt to do the impossible!

## 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  $\Psi$ . 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  $1, 2, \dots, N$ . The interior nodes are labeled  $N + 1, N + 2, \dots, 2N - 1$  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.

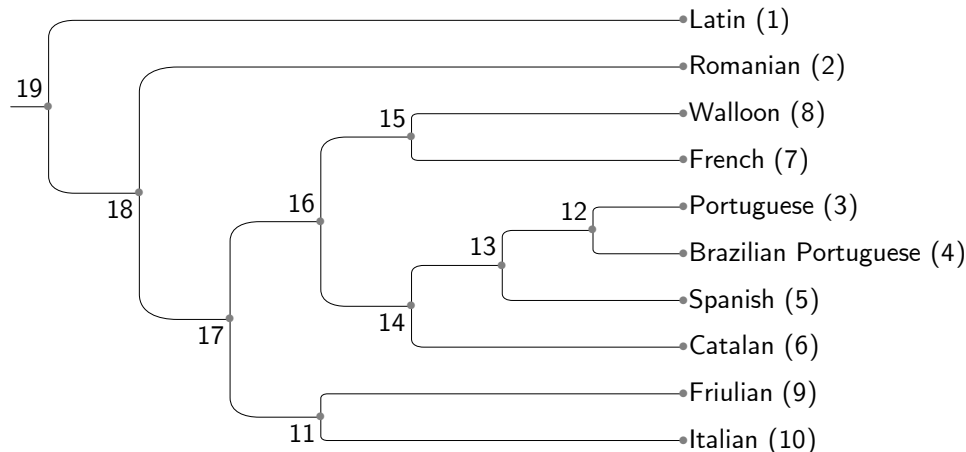


Figure 1: An example tree showing the relationships of  $N = 10$  languages.

A phylogeny is an information-rich graph. For one, it contains information on the relationships of the languages. This topological information is denoted  $\tau$ . 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,  $2N - 3$ :  $B(N) = (2N - 3)!! = 1 \times 3 \times \dots \times 2N - 3$ . Each topology is given a unique label,  $\tau_1, \tau_2, \dots, \tau_{B(N)}$ . 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  $5.86 \times 10^{96}$  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  $t = (t_{N+1}, t_{N+2}, \dots, t_{2N-1})$ . 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  $\mu$ . 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  $i$ th branch of the tree is  $\nu = (t_{\sigma_i} - t_i)\mu_i$ . In this study, we do not estimate the divergence times on the tree, but rather estimate the compound parameter representing the branch lengths (the  $\nu_i$  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	AAGCTTCACCGGCGCAATTATCCTCATAATCGCCACGGACTTACATCCT...
Gorilla	AAGCTTCACCGGCGCAGTTGTTCTTATAATTGCCACGGACTTACATCAT...
Human	AAGCTTCACCGGCGCAGTCATTCTCATAATCGCCACGGGCTTACATCCT...

These are partial mitochondrial sequences from Gojobori88<sup>1</sup>. 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 sequences, 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

---

<sup>1</sup> Takashi Gojobori. "To be determined", (1988)

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 ‘star’

Latin	- s t - e: l l a
French	e - t w a - l -
Spanish	e s t r e - j a
Italian	- s t - e l l a
Brazilian Portuguese	i s t r e - l ɐ
Portuguese	ɨ ʃ t r e - l ɐ
Catalan	ə s t - ɛ - l -
Walloon	- s t - ø: - l -
Friulian	- s t - ɛ - l e
Romanian	- s t - ɛa - -

The word for ‘star’ in Latin, for example, is /ste:lla/. The matrix for the word star also assumes that the fine-scale homology of the segments has been established. For example, the above matrix for the concept ‘star’ assumes that the first segment of the Romanian, Catalan, Portuguese, French and Friulian languages (respectively, s, ə, i, e and s) 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 star in Italian is also s. Why wasn't the first segment from Italian considered to be homologous to the first segments in Romanian, Catalan, Portuguese, French and Friulian? In this case, the alignment program chose the alignment that did not consider ‘s’ of Italian as homologous to the ‘s’ 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 star.

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  $\mathbf{S} = (\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_N)$ , where  $\mathbf{s}_i$  is the segmental string for the  $i$ th language. The segmental information for the  $N = 10$  languages for the cognate class above for ‘star’ is:

Latin	s t e:l l a
French	e t w a l
Spanish	e s t r e j a
Italian	s t e l l a
Brazilian Portuguese	i s t r e l v
Portuguese	i f t r e l v
Catalan	ə s t ε l
Walloon	s t ø:l
Friulian	s t ε l e
Romanian	s t ęa

Here, the segments for Latin would be  $\mathbf{s}_1 = (\text{ste:lla})$ . 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 star that can be formed using two different homology maps.

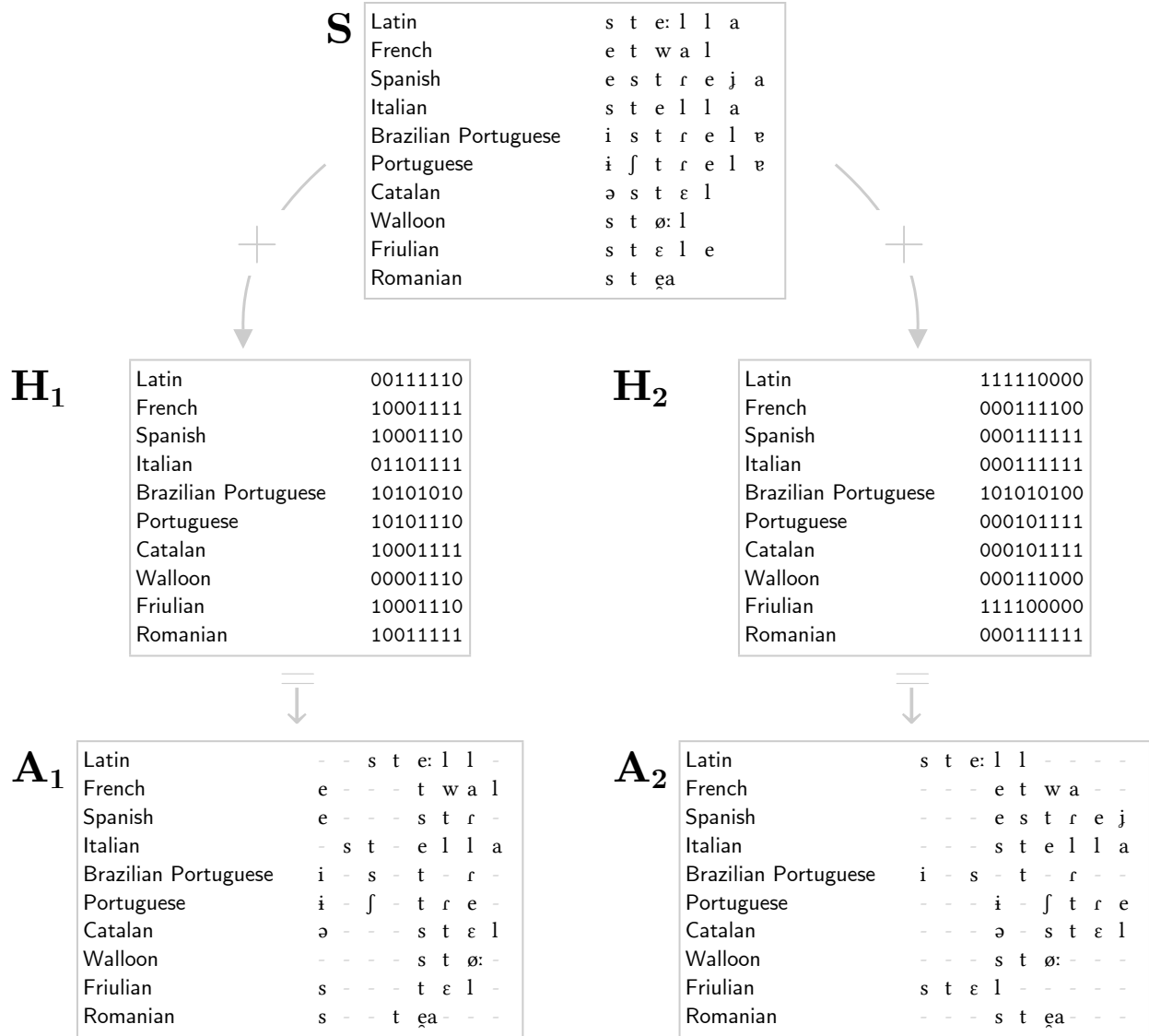


Figure 2: Alignments ( $\mathbf{A}$ ) are formed from the observed segments ( $\mathbf{S}$ ) and a homology map ( $\mathbf{H}$ ).

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 *star*, 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 of states. 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
From	A	$q_{AA}$	$q_{AB}$	$q_{AC}$	$q_{AD}$	$q_{AE}$
	B	$q_{BA}$	$q_{BB}$	$q_{BC}$	$q_{BD}$	$q_{BE}$
	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,  $\beta$ , 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  $\nu$  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  $\pi$  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  $\lambda$  and  $\mu$ , 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!

Thorne et al. (1991) described a model of DNA sequence evolution that allowed single nucleotides to be inserted and deleted at rates  $\lambda$  and  $\mu$ . 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 the right of the immortal link. In the model described by Thorne et al. (1991) 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  $\lambda / \mu$ . We follow the convention of Thorne et al. (1991) in this study. In fact, the model we use is precisely the same as the Thorne et al. (1991) 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 Thorne 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 (**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:

$\tau_1, \dots, \tau_{B(N)}$	Tree Topologies
$\nu_1, \dots, \nu_{B(N)}$	Branch length parameters



$\vartheta$	Parameters associated with the rate matrix $\mathbf{Q}$
$\lambda, \mu$ ( $\lambda < \mu$ )	The insertion and deletion rates of segments
$\mathbf{H}$	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

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

Note that the likelihood is marginalized over all possible alignments,

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

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:

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

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) and Hastings (1970) 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  $\vartheta$ . If this is the first cycle of the Markov chain, initialize  $\vartheta$ , perhaps by choosing a value from the prior distribution.
2. Propose a new value for  $\vartheta$  denoted  $\vartheta'$ . 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,  $\vartheta' = h(\vartheta, u)$ . The probability of proposing the new value is  $q(\vartheta \rightarrow \vartheta')$  whereas the probability of the imagined reverse move, not actually made in computer memory, is  $q(\vartheta' \rightarrow \vartheta)$
3. Calculate the probability of accepting the proposed value:

$$R = \min\left(1, \frac{f(X|\vartheta')}{f(X|\vartheta)} \times \frac{f(\vartheta')}{f(\vartheta)} \times \frac{q(\vartheta' \rightarrow \vartheta)}{q(\vartheta \rightarrow \vartheta')}\right)$$

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  $\vartheta = \vartheta'$ . Otherwise, the proposed state is said to be rejected and the chain remains in state  $\vartheta$
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,  $M_1$  and  $M_2$  with marginal likelihoods,  $P(\mathbf{S}|M_1)$  and  $P(\mathbf{S}|M_2)$  (note the marginal likelihoods are calculated for the same observations). The ratio of the marginal likelihoods,

$$BF_{12} = \frac{P(\mathbf{S}|M_1)}{P(\mathbf{S}|M_2)}$$

called the Bayes Factor, measures the relative support of the two models; a Bayes Factor less than one favors  $M_1$  whereas the opposite 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). The second model assumes that the rate of change to (something) is potentially different than the rate of change to (something else):

$$\begin{aligned} M_1: & \quad q_{ij} = \alpha \text{ (i.e., all elements of rate matrix are equal)} \\ M_2: & \quad q_{ij} = \begin{cases} \alpha & \text{if } i \rightarrow j \text{ is a something} \\ \beta & \text{if } i \rightarrow j \text{ is a something else} \end{cases} \end{aligned}$$

## CONCLUSION

1000 Generations

10 Languages

80 Concepts

83 Unique Segments

83 Unique Characters

## AUTHORS

David M. Goldstein

Los Angeles, CA 90095-1543 USA

E-mail: [dgoldstein@humnet.ucla.edu](mailto:dgoldstein@humnet.ucla.edu)

<https://linguistics.ucla.edu/person/david-goldstein>

Shawn H. McCreight

3060 San Pasqual St., Pasadena, CA 91107 USA

E-mail: [shawn.mccreight@gmail.com](mailto:shawn.mccreight@gmail.com)

<https://nytril.com>

John P. Huelsenbeck

3040 Valley Life Sciences Building #3140, Berkeley, CA 94720-3140 USA

E-mail: [johnh@berkeley.edu](mailto:johnh@berkeley.edu)

<https://vcresearch.berkeley.edu/faculty/john-huelsenbeck>

## APPENDICES

1. Program Output
2. Statistics
3. Character Assignments
4. Partition Assignments

## 5. Concepts Used

Cited works

---

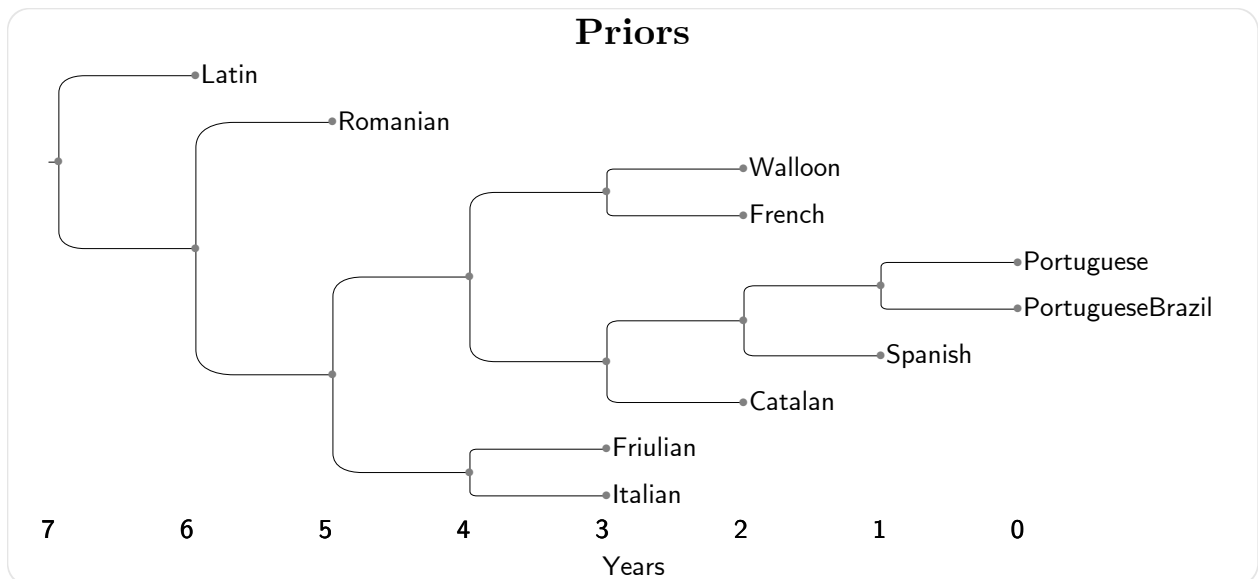
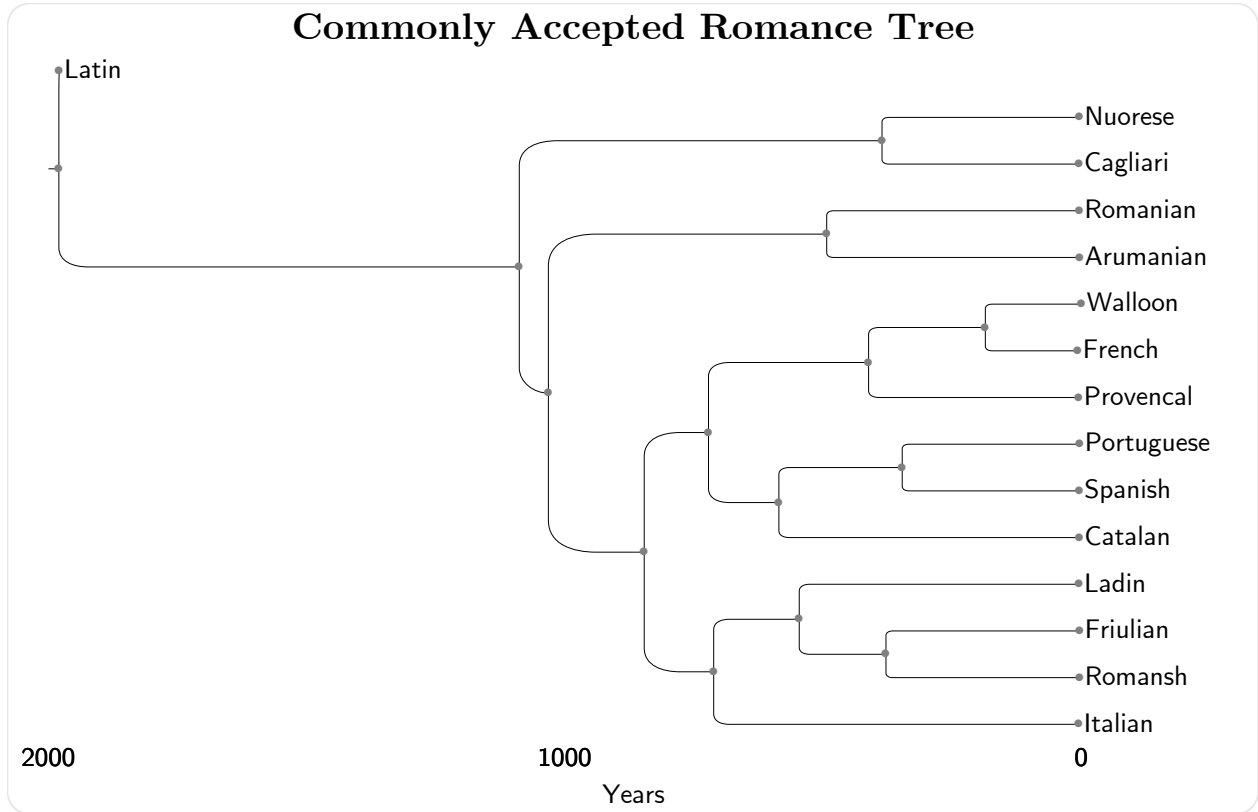
- <sup>1</sup> Gojobori, Takashi. "To be determined", (1988)

Bibliography

---

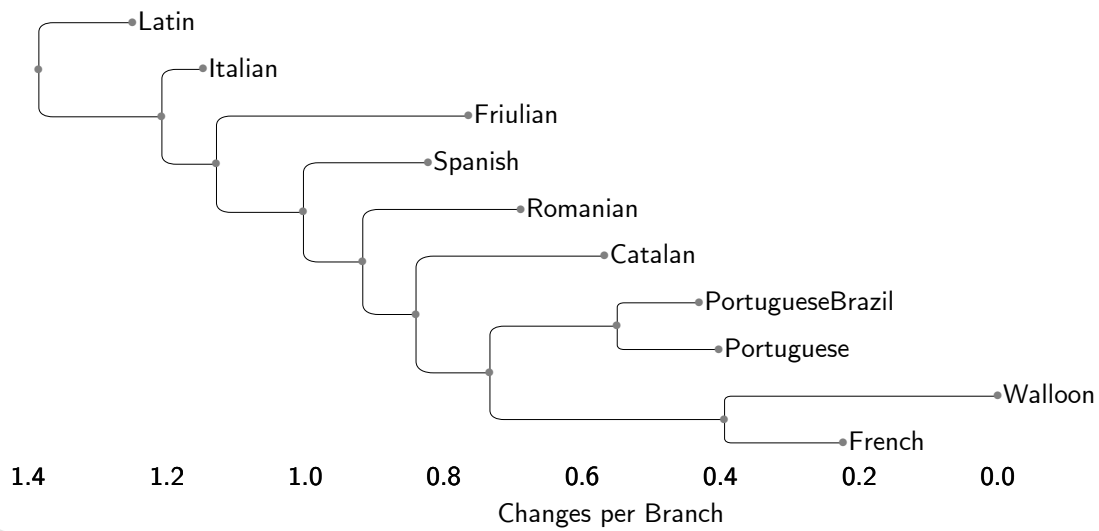
- <sup>1</sup> Lunter, G A, Miklós, I, Song, Y S, and Hein, J "An Efficient Algorithm for Statistical Multiple Alignment on Arbitrary Phylogenetic Trees", (2003)
- <sup>2</sup> Wikipedia1. *Distinctive Feature*. : 2020[https://en.wikipedia.org/wiki/Distinctive\\_feature](https://en.wikipedia.org/wiki/Distinctive_feature).
- <sup>3</sup> Wikipedia2. *Extended Speech Assessment Methods Phonetic Alphabet* . : 2016<https://en.wikipedia.org/wiki/X-SAMPA>.
- <sup>4</sup> InternationalPhoneticAlphabet.org3. *IPA Symbols Chart Complete* . : 2019<http://www.internationalphoneticalphabet.org/ipa-charts/ipa-symbols-chart-complete>.
- <sup>5</sup> InternationalPhoneticAlphabet.org4. *IPA Symbols Chart Complete* . : 2019<http://www.internationalphoneticalphabet.org/ipa-charts/ipa-symbols-chart-complete>.
- <sup>6</sup> Wikipedia5. *International Phonetic Alphabet* . : 2019[https://en.wikipedia.org/wiki/International\\_Phonetic\\_Alphabet](https://en.wikipedia.org/wiki/International_Phonetic_Alphabet).
- <sup>7</sup> Höhna, Sebastian, Landis, Heath, Boussau, Lartillot, Moore, Huelsenbeck, John P., and Ronquist 6. *RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language*. : Systematic Biology 2016<http://www.revbayes.com>. , accessed 2021
- <sup>8</sup> Wichmann, Søren, Holman, Eric W, and Brown, Cecil H 7. *The ASJP Database (version 18)*. : ASJP2018<https://asjp.clld.org>.
- <sup>9</sup> Gojobori, Takashi. "To be determined", (1988)

## Appendix 1 - Program Output

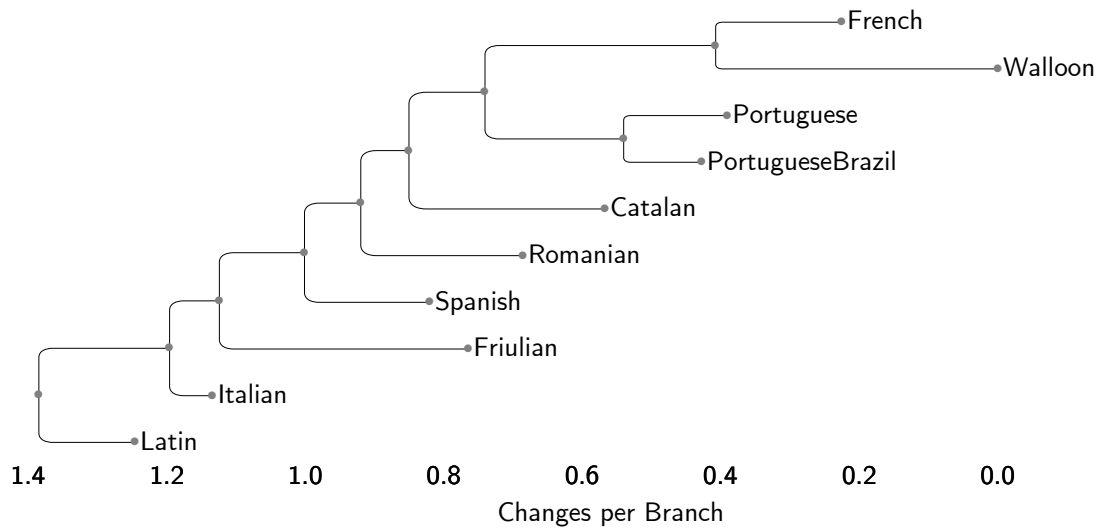


## Appendix 1 - Program Output

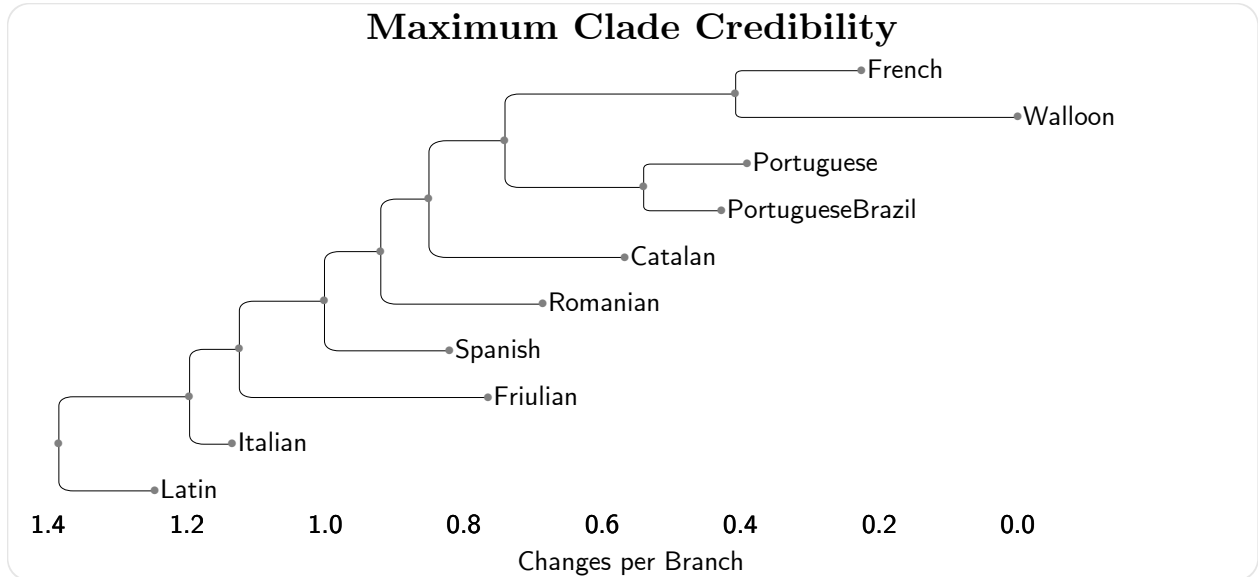
### Consensus



### MAP

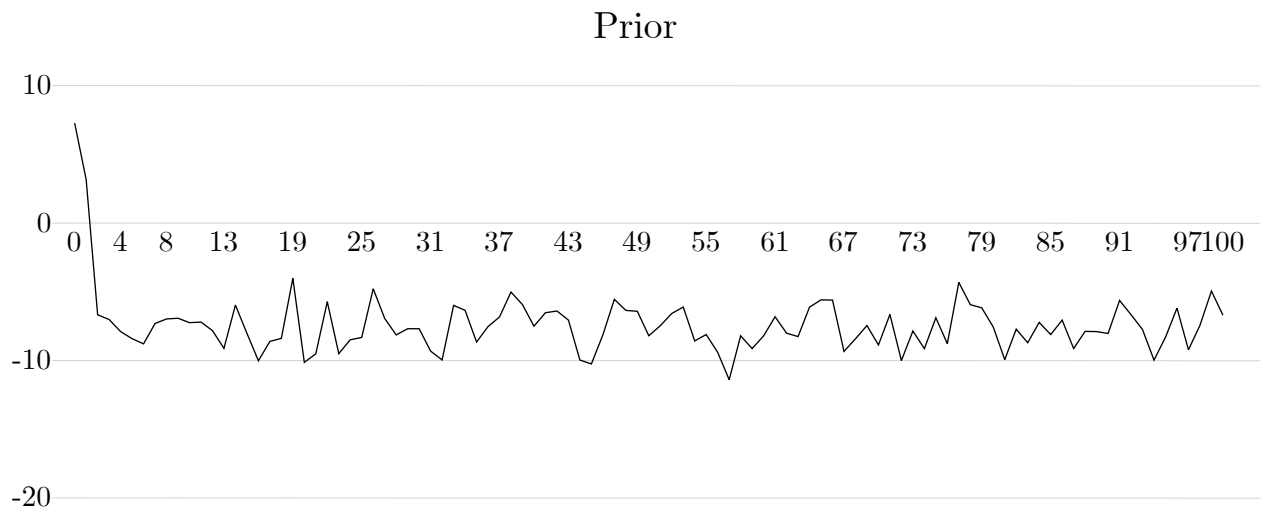
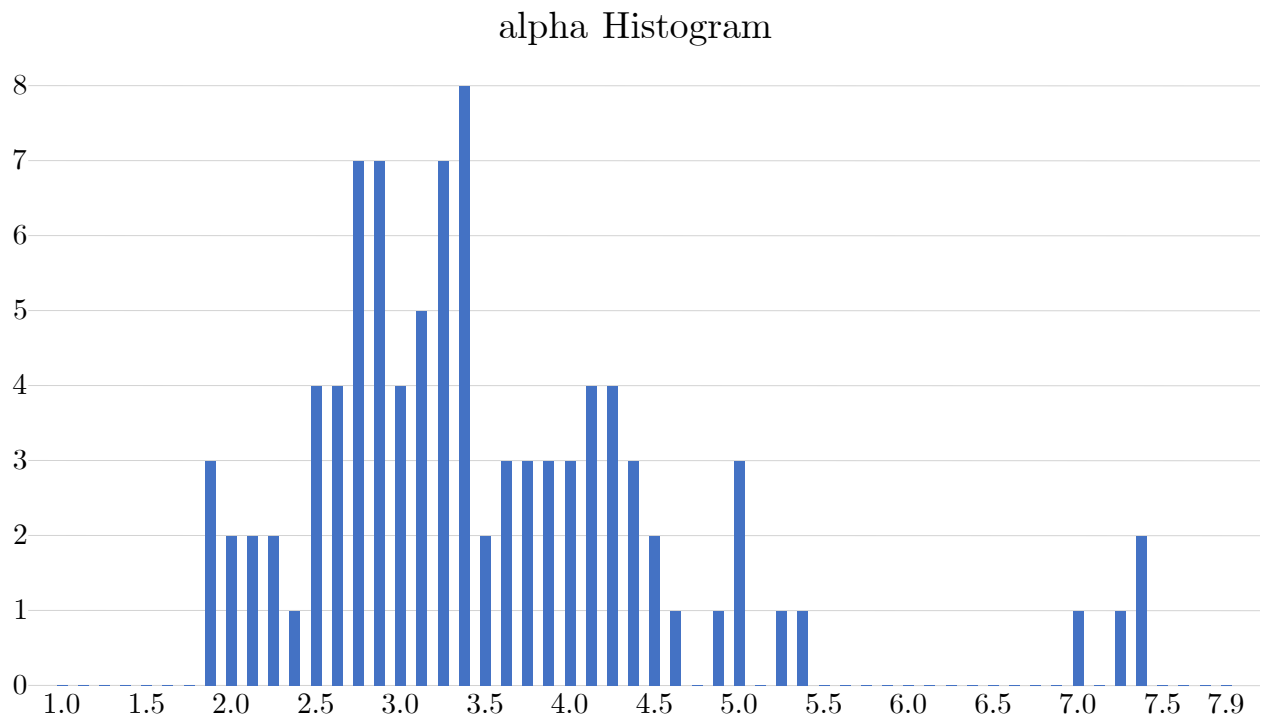


## Appendix 1 - Program Output





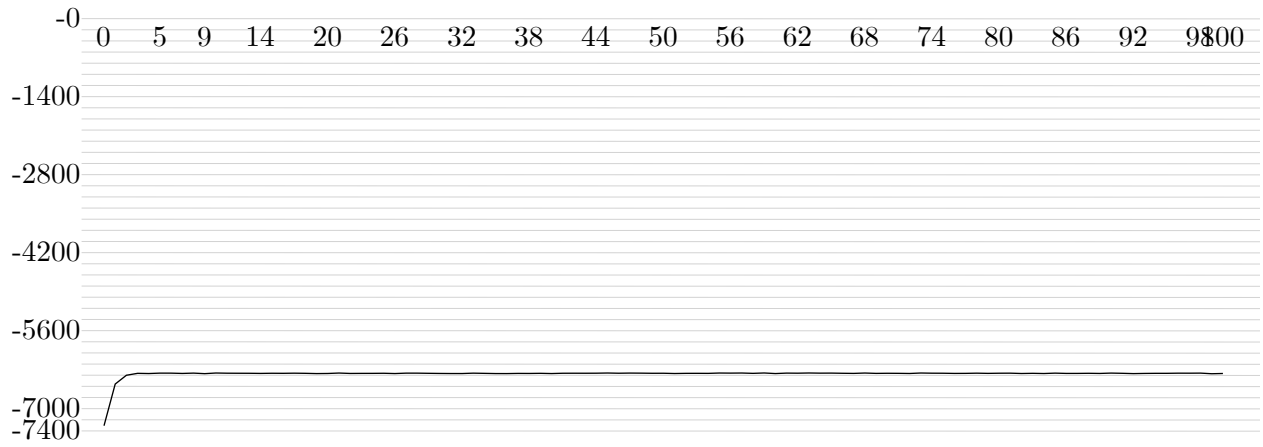
## Appendix 2 - Statistics



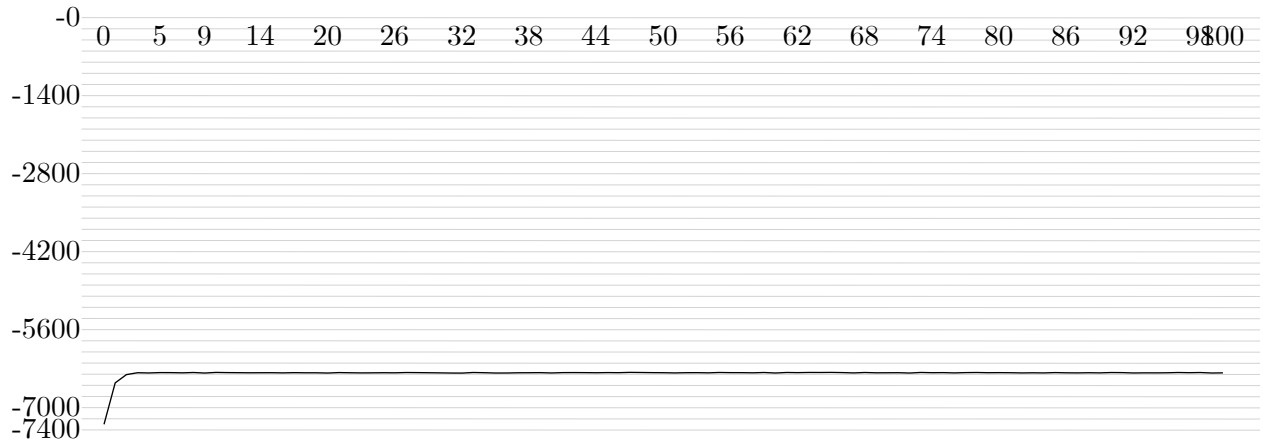
## Appendix 2 - Statistics

---

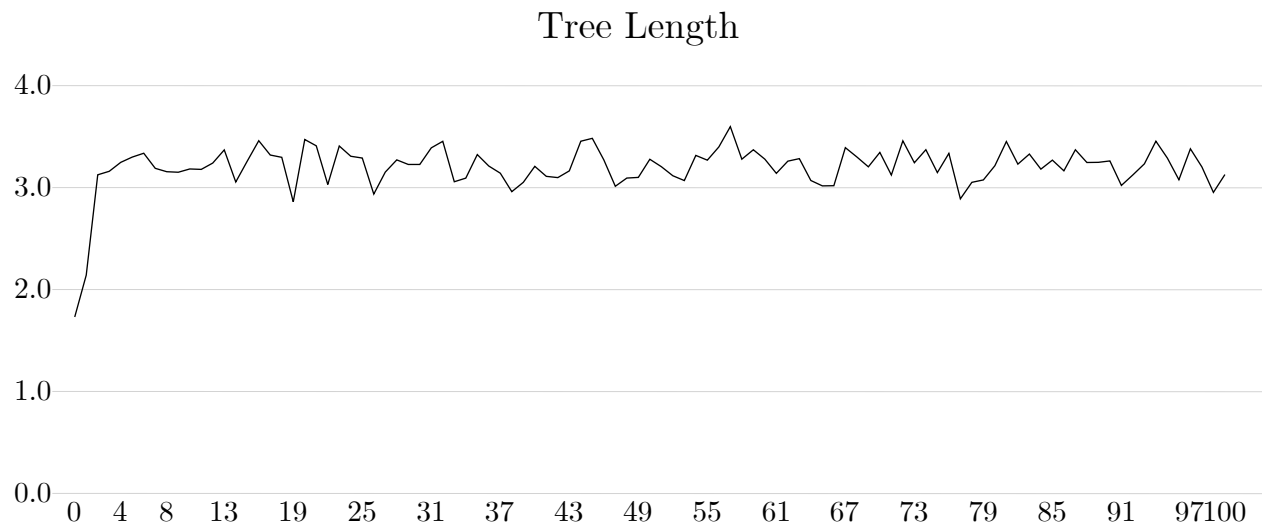
### Likelihood



### Posterior



## Appendix 2 - Statistics



## Appendix 3 - Character Assignments

---

Each segment gets a different number

0	e	1	g	2	o:	3	ʒ	4	ə	5	j	6	o	7	i	8	w	9	ɔ
10	dʒ	11	ɪ	12	j	13	n	14	s	15	u	16	ʃ	17	oj	18	v		
19	b	20	k	21	m	22	h	23	i:	24	a	25	ɐ	26	l	27	a:		
28	ʌ	29	k <sup>w</sup>	30	ẽj	31	ẽj	32	ʈʃ	33	d	34	ɨ	35	ɛ				
36	ɛ:	37	ʈʃ	38	ã	39	ẽ	40	t	41	ẽ	42	u:	43	œ	44	ũ		
45	õ	46	ɳ	47	ø	48	ø:	49	aj	50	r	51	e:	52	ɤ	53	ɑ		
54	ɾ	55	ej	56	ẽ	57	θ	58	ĩ	59	g	60	ɲ	61	õ	62	ɣ		
63	p	64	ã	65	ɐj	66	z	67	x	68	β	69	ɕa	70	ʃ	71	ð		
72	y	73	ɹ	74	ɔa	75	ɥ	76	c	77	ts	78	ɬ	79	ɛj				
80	f	81	g <sup>w</sup>	82	ẽɥ														

## Appendix 4 - Partition Assignments

---

David's basic rules

### 1 Nasal Vowel

---

ẽ ĩ ẽĩ ã ẽ ã œ ã õ ẽ ĩ ã ã

### 2 Vowel

---

e o: ə o i ɔ ɪ u oɪ i: a ɐ a: i ɛ ɛ: u: ø ø: aɪ e: ɑ ej vj ɛa y ɔa ej ẽu

### 3 Nasal Consonant

---

n m ŋ ɲ

### 4 Non Sylabic Sonorants

---

w j l r

### 5 Consonants

---

g ʒ j dʒ s ʃ v b k h ʌ k<sup>w</sup> tʃ d ʈ t ɸ ɾ θ g ɣ p z x β ʃ ʈ ɾ ɸ c ts ɭ f  
g<sup>w</sup>

## Appendix 5 - Concepts Used

	I	We	YouPlural	Here
Latin	- e g o:	n o: - s	w o: - s	- e k k u m h i: k
French	- ʒ - ə	n u - -	v u - -	i - - s - - i -
Spanish	- j - o	n o - s	b o - s	- a - k - - i -
Italian	- i - o	n o: i -	v o: i -	- - k w - - i -
Brazilian	- e - w	n ɔ j - s	v ɔ j s	- a - k - - i -
Portuguese				
Portuguese	- e - w	n ɔ - f	v o - f	- ɐ - k - - i -
Catalan	- ʒ - ɔ		b u - s	- a - k - - i -
Walloon	- ʒ - ɪ	n o - -	v o - s	
Friulian	- j - o	n o: - -	v o: - -	- a - k - - i -
Romanian	j e - w	n oj -	v oj -	
	There	There-Ibi	There-Adlbi	Who-Qui
Latin	i l l a: k	i b i:		k <sup>w</sup> - i s
French	- - l a -			k - i -
Spanish	a - j a -		a i	
Italian	- - l a -	- v i		k - i -
Brazilian	- - l a -		a i	
Portuguese				
Portuguese	- - l a -		ɐ i	
Catalan	ə - ʎ a -			k - i -
Walloon	- - l a -			k - i: -
Friulian				k w i -
Romanian	- - l a -			
	Who-Quem	What	When-Quando	How-Quomodo
Latin	k <sup>w</sup> - e m -	k <sup>w</sup> - i d	k <sup>w</sup> - a n d o:	k <sup>w</sup> o: m o d o
French		k w a -	k - ã - -	k ɔ m - -
Spanish	k j e n -	k - e -	k w a n d o	k o m o - -
Italian		k - e -	k w a n d o	k o m e - -
Brazilian	k - ẽj - -	k - i -	k w ẽ - d u	k o m u - -
Portuguese				
Portuguese	k - ẽj - -	k - i -	k w ẽ - d u	k o m u - -
Catalan		k - ɛ -	k w a n	k ɔ m - -
Walloon		k w ɛ: -	k w a n d	k ɔ m ẽ - -
Friulian		fj - e -	k w a n t	fj e m u: t -
Romanian	fj - i n e	fj - e -	k - i n d	k u m - -
	One	Two-Duos	Two-Duae	Two-Duas
Latin	u: n u m	d u o: s	d u - aj	d u a s
French	œ - - -	d - ø -		
Spanish	u n o -	d - o s		
Italian	u n o -		d u - e	
Brazilian	ũ - - -	d - oj s		d u ɐ s
Portuguese				
Portuguese	ũ - - -	d - oj f		d u ɐ f
Catalan	u n - -	d - o s		
Walloon	õ - - -	d - ø: -		
Friulian	u ŋ - -	d - oj -		
Romanian	u n u -	d - oj -	d o w ə	

## Appendix 5 - Concepts Used

	Three	Five	Big	Big-Grandem
Latin	t r - e: s	k <sup>w</sup> i: n k <sup>w</sup> - e	m a g n u m	g r a n d e m
French	t ʁ w a -	s ẽ - k - -		g ʁ ã - - -
Spanish	t r - e s	θ i n k - o	m a - ɲ o -	g r a n d e -
Italian	t r - e -	ʃ i ŋ k w e	m a ɲ ɲ o -	g r a n d e -
Brazilian	t r - e s	s ã - k - u		g r ẽ - ɕ i -
Portuguese				
Portuguese	t r - e ʃ	s ã - k - u		g r ẽ - d i -
Catalan	t r - ε s	s i ŋ k - -		g r a n - - -
Walloon	t r - ø: -	s ẽ - k - -		g ʁ ã - - -
Friulian	t r - e: -	ʃ i ŋ k - -		g r a n t - -
Romanian	t r - ej -	ʃ i n ʃ - -		
	Big-Majjorem	Long-Longum	Long-Largum	Wide
Latin	m a j j o r e m	l o - n g u m	l a r g u m	l a: t u m
French		l ɔ̃ - - - -	l a ʁ ʒ - -	l e - - -
Spanish		l o - n g o -	l a r ɣ o -	l a t o -
Italian		l u - n g o -	l a r g o -	l a t o -
Brazilian		l õ - g u -	l a ʁ g u -	l a t u -
Portuguese				
Portuguese		l õ - g u -	l a r g u -	l a t u -
Catalan		ʎ o - ŋ - - -	ʎ a r k - -	
Walloon		l ɔ̃ - - - -		
Friulian		l u n k - -	l a r k - -	
Romanian	m a - - - r e -	l u n g - -	l a r g - -	l a t - -
	Wide-Amplum	Fish-Piskem	Fish-Pescion	Bird-Avem
Latin	a m p l u m	p i s k e m		a w e m
French	ã - p l - -		p w a s ɔ̃	
Spanish	a n ʃ - o -	p e θ - - -		a b e -
Italian	a m p j o -	p e ʃ ʃ e -		
Brazilian	ẽ - ʃ - u -	p e j ʃ - i -		a v i -
Portuguese				
Portuguese	ẽ - ʃ - u -	p e j ʃ - i -		a v i -
Catalan	a m p l ə -	p e ʃ - - -		a w - -
Walloon			p - ε h ɔ̃	
Friulian		p ε s - - -		
Romanian		p e ʃ t e -		
	Bird-Avicellum	Bird-Passerem	ToDrink	ToEat
Latin	- a w i - k - e l l u m	p a s s e r e m	b i b - e r e	k o m e d e r e
French	w a - - - z - o - - -		b - - w a ʁ -	
Spanish		p a - x a r o -	b e b - e r -	k o m e - - - r -
Italian	- - u - t ʃ - ε l l o	p a s s e r o -	b e - - - r e	
Brazilian		p a - s ɐ r u -	b e b - e ʁ -	k o m e - - -
Portuguese				
Portuguese		p a - s ɐ r u -	b i β - e r -	k u m - - e r -
Catalan	- - u - - s - e - ʎ - -		b ε w - - r ə	
Walloon	- - u: - - h - ε: - - -		b - - - ø: ʁ -	
Friulian	- - u - - ʃ j ε - l - -		b e v i - -	
Romanian		p a - s ə r e -	b - - - ɛ a - -	

## Appendix 5 - Concepts Used

	ToEat-Manducare	ToBite	ToBite-Morsicare	ToSuck
Latin	m a n d u k a : r e	m o r d e : r e	m o r s i k a : r e	s u : g e r e
French	m ã - ʒ - - e - -	m o ʁ d - ʁ -		s y s e - -
Spanish	m a n d u k a r -	m o r ð e r -	m o - s - k a r -	
Italian	m a n dʒ - - a : r e	m ɔ r d e r e	m o r s i k a r e	
Brazilian	m ê - ʒ - - a ʁ -	m o ʁ d e ʁ -	m u - s ə g a - -	s u g a ʁ -
Portuguese				
Portuguese	m ê - ʒ - - a r -	m u r ð e r -	m u - s ə g a r -	s u g a r -
Catalan	m ə ɲ ʒ - - a - -		m u - s ə γ a - -	
Walloon	m a ɲ - - i : - -			s y s i : - -
Friulian	m a ɲ ʃ - - a - -			
Romanian	m i n - - k a - -		m u r s e k a - -	s u dʒ e - -
	ToSpit	ToSpit-Exconspuere	ToSpit-Sputare	ToKnow
Latin	k o n s p u e r e		- s p u t e r e	s k i : r e
French	k ɔ̃ - s p y e - -			
Spanish		e s k u - p i r	e s p u t a r -	
Italian			- s p u t a : r e	
Brazilian	k u - s p i ʁ -			
Portuguese				
Portuguese	k u - ʃ p i r -			
Catalan		ə s k u - p i - -		
Walloon				
Friulian				
Romanian		- s k u j p a		ʃ t i - -
	ToKnow-Sapere	ToKnow-Cognoscere	ToThink	ToThink-Pensare
Latin	s a p - e r e	k o g n o : s k e r e	k - o : g i t a : r e	p e n s a : r e
French	s a v w a ʁ -	k ɔ̃ - n ɛ - t ʁ -	k ɥ - - i d e - -	p ã - s e - -
Spanish	s a b - e r -	k o - n o θ e r -	k w - - i d a r -	p e n s a r -
Italian	s a p - e : r e	k o - n o ʃ f e r e	k - o dʒ i t a : r e	p e n s a : r e
Brazilian	s a b - e ʁ -	k o - ɲ e s - e ʁ -	k - u ʒ i t a ʁ -	p ê - s a ʁ -
Portuguese				
Portuguese	s ɐ b - e r -	k u - ɲ i s - e r -	k - u ʒ i t a r -	p ê - s a r -
Catalan	s ə β - ɛ - -	k u - n ɛ ʃ - ə - -	k - u - j d a - -	p ə n s a - -
Walloon	s a v - ø : ʁ -	k i - n ɔ h - - -		p ê - s e - -
Friulian	s a v - e : - -	k o - ɲ o s - i - -		p e ɲ s a : - -
Romanian		k u - n ɔ a ʃ t e - -	k - u dʒ e t a - -	p ə - s a - -
	ToFear	ToFear-Tremere	ToSleep	ToLive
Latin	t i m e : r e	t r e m - e r e	d - o r m i : r e	w i : w e r e
French		k ʁ ê - d - ʁ -	d - ɔ ʁ m i ʁ -	v i v - ʁ -
Spanish	t e m e r -	t r e m - e r -	d - o r m i r -	b i b i r -
Italian	t e m e : r e	t r e m - a : r e	d - o r m i : r e	v i : v e r e
Brazilian	t e m e ʁ -	t r e m - e ʁ -	d - o ʁ m i ʁ -	v i v e ʁ -
Portuguese				
Portuguese	t i m e r -	t r i m - e r -	d - u r m i r -	v i v e r -
Catalan	t e m ə - -		d - u r m i - -	b i w - r ə
Walloon			d w a ʁ m i - -	
Friulian	t e m i - -	t r i m - a : - -	d - u r m i : - -	v i v i - -
Romanian	t e m e - -		d - o r m i - -	v i - a - -



## Appendix 5 - Concepts Used

	ToDie	ToSwim	ToSwim-Navigare	ToFly
Latin	m o r i: - -	n - a t a: r e	n a v i g a: r e	w o l a: r e
French	m u ʁ i ʁ -		n a - - ʒ e - -	v o l e - -
Spanish	m o r i r -	n - a d a r -	n a b e g a r -	b o l a r -
Italian	m o r i: r e	n w o t a: r e	n a v i g a r e	v o l a: r e
Brazilian	m o ʁ e ʁ -	n - a d a ʁ -	n a v i ɣ a ʁ -	v o - a ʁ -
Portuguese				
Portuguese	m u ʁ e r -	n - ɐ ð a r -	n ɐ v i ɣ a r -	v w - a r -
Catalan	m u r i - -	n - ə d a - -	n ə b ə g a - -	b u l a - -
Walloon	m ɔ r i - -		n ɛ - - ɕ i: - -	v ɔ l e - -
Friulian	m u r i: - -	n - a d a - -	n a v i g a - -	- - - - - -
Romanian	m u r i - -		n a v i g a - -	- - - - - -
	ToFly-Exvolare	ToCome	ToFall	ToFall-Tumbare
Latin		w e n i: r e	k - a d e r e	
French		v ə n i ʁ -	ʃ w a - - ʁ -	t ɔ - b e -
Spanish		b e n i r -	k - a - e r -	t u m b a r
Italian	z v o l a: r e	v e n i: r e	k - a d e: r e	
Brazilian		v - - i ʁ -	k - ɐ - i r -	t ɔ - b a ʁ
Portuguese				
Portuguese		v - - i r -	k - ɐ - i r -	t ɔ - b a r
Catalan		b ə n i - -	k - a w - r ə	t u m b a -
Walloon		v i n i - -	ʃ - ɛ: - - ʁ -	t u m - e -
Friulian	z v o l a - -	v i ɲ i: -	ʃ j a d e - -	
Romanian	z b u r a - -	v e n i - -	k ə d ɛ a - -	t u m b a -
	ToSing-AncestralClass	ToPlay-Ludere	ToPlay-Jiocare	ToFreeze
Latin	k a n t a: r e	l u: d e r e	j o - k a: r e	g e l a: r e
French	ʃ ã - t e - -		ʒ - w - e - -	ʒ ə l e - -
Spanish	k a n t a r -	l u d i r -	x u - g a r -	- e l a r - -
Italian	k a n t a: r e		ɕ o - k a: r e	ɕ e l a r e
Brazilian	k ẽ - t a r -		ʒ o - g a ʁ -	ʒ i l a r -
Portuguese				
Portuguese	k ẽ - t a r -		ʒ u - ɣ a r -	ʒ i l a r -
Catalan	k ə n t a - -		ʒ u - g a - -	ʒ ə l a - -
Walloon	ʃ ã - t e - -		ɕ u w - e - -	ɕ a l e - -
Friulian	c a n t a: - -		ɕ u j - a - -	
Romanian	k i n t a - -		ʒ u - k a - -	
	ToFreeze-Glaciare	ToFreeze-Inglaciare	Sun	Sun-Soliculus
Latin	g l a - k i a: r e	i n g l a k i a: r e	s o: l e m	s o l i c u l u s
French	g l a - s - e - -			s ɔ l ɛ j - - - -
Spanish			s o l - -	
Italian	g j a t ʃ j a r e		s o l e -	
Brazilian			s ɔ w - -	
Portuguese				
Portuguese			s ɔ ʔ - -	
Catalan	g l ə - s - a - -		s ɔ l - -	
Walloon			s ɔ l ɔ -	
Friulian	g l a - ʃ - a - -			s o r e - - l i -
Romanian		i n g - e t s - a - -	s ɔ a r e -	

## Appendix 5 - Concepts Used

	Moon	Star	Water	Rain
Latin	l u : n a	- s t - e : l l a	a - k <sup>w</sup> - a	p l u - w i a
French	l y n -	e - t w a - l -	o - - -	p l ʁ - - i -
Spanish	l u n a	e s t r e - j a	a - ɣ w a	j - u - b j a
Italian	l u n a	- s t - e l l a	a k k w a	p j o d dʒ - a
Brazilian	l u - ɐ	i s t r e - l ɐ	a - g w ɐ	ʃ - u - v - ɐ
Portuguese				
Portuguese	l u - ɐ	i ʃ t r e - l ɐ	a - g w ɐ	ʃ - u - v - ɐ
Catalan	ʎ u n ə	ə s t - ε - l -	aj - ɣ w ə	p l u - ʒ - ə
Walloon	l ø n -	- s t - ø : - l -	ɛ : - - w -	p l u : - ɸ - -
Friulian	l u n e	- s t - ε - l e	a - g - e	p l o - - j e
Romanian	l u n ə	- s t - ɕa - -	a - p - ə	p l ɔa - - j e
	Lake	Sea	Salt	Earth
Latin	l a k u m	m a r e	s a : l e m	t - e r r a
French	l a k - -	m ɛ ɸ -	s ɛ l - -	t - e - ɸ -
Spanish	l a g o -	m a r -	s a l - -	t j e - r a
Italian	l a g o -	m a r e	s a l e -	t - ε r r a
Brazilian	l a g u -	m a ɸ -	s a w - -	t - e - ɸ ɐ
Portuguese				
Portuguese	l a ɣ u -	m a r -	s a l - -	t - ε - ɸ ɐ
Catalan	ʎ a k - -	m a r -	s a l - -	t - ε - r ə
Walloon	l a k - -	m e ɸ -	s e - - -	t - ɛ : - ɸ -
Friulian	l a : t - -	m a : r -	s a : l - -	t j ε - r e
Romanian	l a k - -	m a r e	s a r e -	ts - a - r ə
	Sky	Wind	Fire-Focus	Green
Latin	k - aj l u m	w - e n t u m	f - o k u m	w i r i d e m
French	s j ɛ l - -	v - ã - - -	f - ø - - -	v e ɸ - - -
Spanish	θ j e l o	b j e n t o -	f w e g o -	b e r - d e -
Italian	ʧ - ε l o -	v - ε n t o -	f w ɔ k o -	v e r - d e -
Brazilian	s - ε w - -	v - ě - t u -	f - o g u -	v e r - d i -
Portuguese				
Portuguese	s - ε w - -	v - ě - t u -	f - o ɣ u -	v e r - d i -
Catalan	s - ε l - -	b - e n - - -	f - ɔ k - -	v ɛ r - t - -
Walloon	s - i : ɸ - -	v - ě - - - -	f - ø : - - -	v ɛ ɸ - - -
Friulian	ʧ - i : l - -	v - i n t - -	f - u : k - -	v ɛ r - t - -
Romanian	ʧ - e r - -	v - i n t - -	f - o k - -	v e r - d e -
	Year	Name-Nomen	Name-Nominem	Tongue
Latin	a n n u m	n o : m e n	n o : m i n - e m	l i n g <sup>w</sup> - a
French	ã - - -	n ɔ̃ - - -		l ã - g - -
Spanish	a - ɲ o -		n o m - b r e -	l e ɲ g w a
Italian	a n n o -	n o : m e -		l i n g w a
Brazilian	ɐ - n u -	n o m i -		l ĩ - g w ɐ
Portuguese				
Portuguese	ɐ - n u -	n o m i -		l ĩ - g w ɐ
Catalan	a - ɲ - -	n ɔ m - -		
Walloon	ã - - - -	n ɔ̃ - - -		l ẽɥ - - - -
Friulian	a - ɲ - -	n ɔ ɲ - -		l ε n g - ε
Romanian	a n - - -	n u m e -		l i m b - ə