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
dgoldstein@humnet.ucla.edu
[2] Nytril LLC, 3060 San Pasqual St., Pasadena, CA 91107, USA
shawn.mccreight@gmail.com
[3] Department of Integrative Biology
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 2, 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	/ˈmano/	1
Italian	/ˈmaːno/	1
Russian	/rʊka/	2
Polish	/rɛŋka/	2

Table 2. 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 2 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 /ˈmano/.

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 220 cognate classes comprising 1546 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.

Results

Understanding Word Transformation

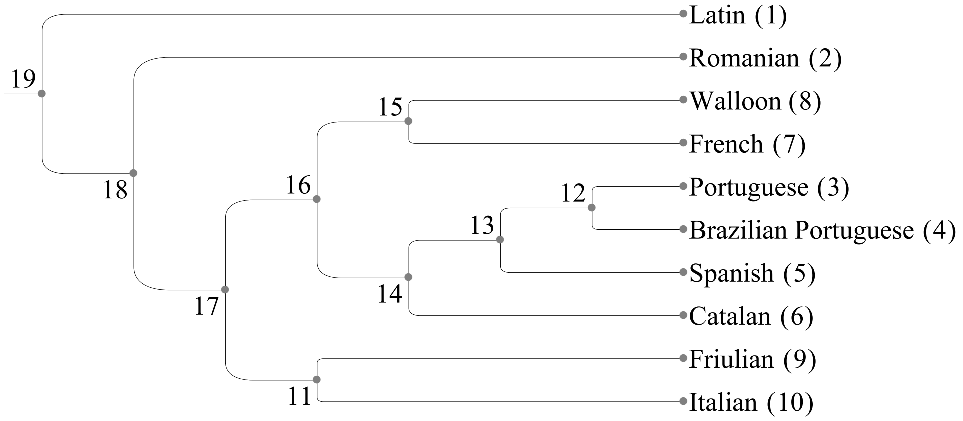
Marginal likelihoods under the three models of word transition were approximated using the stepping-stone path-sampling method[15]. The marginal likelihoods were NUMBER, NUMBER, and NUMBER under the Jukes-Cantor, linguistically informed, and GTR models, respectively. Hence, the linguistically informed model is best supported by the data.

The linguistically informed model allows different rates for different groups of word segments, Figure 1 summarizes the estimated rates of change from one segment grouping to another. Note that rates of change are much greater from one word segment to another when the change is within the word segment group than it is when the change is between word segments in different groups.

Word Segment Relationships

Our method treats the relationships of word segments from one language to another as a random variable. Figure [NUMBER] shows alignments for ten of the words we analyzed. The pictured alignments form a 95% credible set, with alignments ordered from highest to lowest posterior probability. Note that insertions and deletions appear to occur more frequently at the end of the words. Also note that the alignments tend to assign word segments from the same segment group to the same column in the alignment.

Romance Language Phylogeny



Discussion

Despite a wealth of textual data from Latin and its medieval descendants, the phylogeny of Romance remains a challenge [16-18]. The topology of our tree agrees with the prevailing views of the Romance phylogeny in most aspects. For instance, it recognizes that French and Walloon have undergone the greatest segmental change. In addition, it identifies the Ibero-Romance (Spanish, European Portuguese, and Brazilian Portuguese) and Gallo-Romance (French, Walloon, and Catalan) clades. In recent years a consensus appears to have emerged among specialists that Romanian was one of the first clades to form. This view has been based on studies using both cognate data [5-1-7] and sound changes [19]. Our results suggest that Romanian did not emerge as early as is thought. Rather it is Italian that was the first to form. Furthermore, our results demonstrate that the branching of Romance languages did not take place along an East-West fault line, since Romanian, an Eastern Romance language, emerges between Ibero- and Gallo-Romance. Finally, our study refutes an idea that has long circulated in the linguistics literature, namely that the diversification of Romance correlates with the date of Roman colonization. According to this theory, Romanian would be the last language to emerge since the Roman colony of Dacia was the last to be established.

There is a long debate—known as the "Neogrammarian controversy"—about the units involved in sound change. Leonard Bloomfield famously declared "phonemes change"[20]. According to this view, sound change is abrupt and affects all words at once. So a change from /a/ to /o/ would occur more or less simultaneously in all words with the input phoneme. Other scholars have emphasized the role of acoustic and auditory phonetics in sound change[21-23], while yet another view contends that sound change occurs by gradually making its way through the lexicon in a process known as *lexical diffusion*. [24,25] Our method cannot be located within any one of these traditions, but rather has affinities with the phonemic approach and lexical diffusion. Since our analyses are based on phonemic representations, it is phonemes that undergo transitions. Rates of change are, however, inferred from phonemes in the context of word forms. Consider again the change from /a/ to /o/. Our method does not infer a single transition in this case. It instead infers a rate of change that is based on the frequencies of /a/ and /o/ in the data and how often the former transitions to the latter.

Although the rates at which linguistic phenomena change is poorly understood, one thing that is clear is that rates of change differ among different components of grammar. In general, changes in the basic vocabulary of a language take place at a slower rate than phonological change. One can compare for instance dialects of American and British English. After approximately four centuries, a number of phonological differences distinguish these two dialects, but differences in basic vocabulary are either minimal or have yet to occur. Rates inferred from lexical cognate relationships are thus going to underestimate rates of change and divergence times.

Our event-based modeling approach to linguistic history has the power to transform the discipline by offering an inroad into questions that were previously intractable. For instance, questions pertaining to the relationship between phonological change and the phonemic inventory of a language can now be addressed. Does the frequency of a phoneme in the words of a language affect its diachronic stability? To what extent is phonemic change sensitive to the size and structure of the phonemic inventory of a language? What is the role of "natural classes" in phonological change? This last question can be addressed in particular through comparison of models with different numbers of rate parameters. The method can also be extended in various ways, for instance, to handle more complex phonological changes (such as those involving context dependency); to estimate divergence times and diversification rates in addition to topology; or to model the history of words along a phylogenetic tree.

- maybe conclude on something that talks about a more -omics approach to linguistics, perhaps by combining automatic cognate discovery with the more event-based modeling approach we pursue here. Also, might mention that this framework does allow for more complicated questions to be addressed, such as context dependence of segment change.

Methods

A detailed description of the model and analyses can be found in the supplemental material, [reference]. What follows is an outline, sufficient in detail to provide an understanding of the model and experiments performed in this study.

Data

We selected 106 concepts from the 200-word Swadesh list of basic vocabulary. The initial set of data was downloaded from Wiktionary[26], which was then manually checked for accuracy and augmented. For each concept, lexical items descending from a common ancestor were grouped together into cognate sets. Membership in a particular cognate set depends solely on shared descent; meaning is irrelevant. For instance, the Latin adjective *gravis* means 'heavy', but its French descendant *grave* has lost this sense and now means 'serious.' They are assigned to the same cognate set because segmentally French *grave* descends from Latin *gravis*. There are in total 106 cognate sets, 1546 word forms, and 90 segments. Each word is represented phonemically with the IPA alphabet. Phonemic representations were used in lieu of phonetic representations since phonetic data is harder to come by and for ancient languages such as Latin non-existent.

Latin is a highly inflectional language, which means that there are multiple word forms for verbs, nouns, and adjectives. For verbs, infinitive forms were selected. For nouns, singular forms were used in either the nominative or the accusative case, since these are the two case forms that were most often ancestral to Romance descendants. For adjectives, masculine singular nominative forms were typically used. The selection of word forms is not without consequences, since it impacts the rates of phonological change. For instance, most masculine singular nominative adjectives in Latin end in -*us*. By contrast, feminine singular nominative adjectives end in -*a*. Use of the feminine forms would increase the number of transitions originating in this vowel.

In addition to being collected and classified, the data were also manually aligned, so that putatively historically related segments belong to the same column. The manual alignments served as the starting point for the MCMC sampling procedure (see below), so our analyses did not condition on any particular alignment of word segments being correct. Table 1 provides an illustrative example of the alignment for the concept 'what'. The ancestral Latin form /ˈkʷ-id/ begins with a voiceless velar stop with a secondary labial articulation /ʷ/. In all of the descendant languages, the secondary articulation is lost. In some languages, it becomes a consonant in its own right (i.e., the kʷ of French and Walloon). To accommodate this change, the Latin form has two segmental slots before the vowel (i.e., kʷ-id), one of which anticipates the development of /k/, the other its secondary articulation /ʷ/.

Language	Phonemic Representation	Alignment			
Latin	/kʷid/	kʷ	-	i	d
French	/kwa/	k	w	a	-
Spanish	/ke/	k	-	e	-
Italian	/ke/	k	-	e	-
Brazilian Portuguese	/ki/	k	-	i	-
Portuguese	/kɨ/	k	-	ɨ	-
Catalan	/kɛ/	k	-	ɛ	-
Walloon	/kwɛː/	k	w	ɛː	-
Friulian	/ʧe/	ʧ	-	e	-
Romanian	/ʧe/	ʧ	-	e	-

Table 1. Manual alignment of the words for the concept 'what'.

Our method models insertions, deletions, and transitions among segments. An effort was made to exclude word forms that involve changes of sets of segments (such as occur in morphological change). Metathesis, a change in which two segments swap positions, is modeled as separate transitions from the input and output segments. For instance, the Latin verb *exprimere* begins with a string /eks/, which in Romanian has become /sk/. The initial vowel has been lost and /k/ and /s/ have undergone metathesis. Under our model, this change is treated as two separate transitions, one from /k/ to /s/ and another from /s/ to /k/.

Model and Statistical Inference

We assume that modern languages are related to one another through an unknown phylogenetic tree, $\Psi = (\tau, \nu)$, that contains information on the relatedness of the languages (τ) and the expected number of sound substitutions ν that occur along each branch of the tree. Language evolves along the branches of the tree according to the TKF91 model[11] which allows either a word segment transition, insertion, or deletion to occur in an instant of time. Insertions and deletions occur at rates λ and μ respectively, with $\lambda < \mu$. Word segment substitution is modeled as a continuous-time Markov model with the IPA word segments as the states of the process (here, we use 90 word segments). The rates of change between all pairs of states are contained in the rate matrix $\mathbf{Q}$, which has parameters $\boldsymbol{\theta}$.

We perform estimation in a Bayesian framework, basing parameter estimates on the joint posterior probability distribution,

$$f(\Psi, \lambda, \mu, \theta|\mathbf{S}) = \frac{f(\mathbf{S}|\Psi, \lambda, \mu, \theta)}{f(\Psi, \lambda, \mu, \theta)f(\mathbf{S})}$$

where $\mathbf{S}$ are the observed word segments for the cognate words. The likelihood function, $f(\mathbf{S}|\Psi, \lambda, \mu, \theta)$, is calculated using the algorithm described by Lunter *et al.*[27]. We use priors for the model parameters that are standard in phylogenetics, with the exception of the rate parameters for the insertions and deletions, which are assumed to follow independent and identically distributed exponential distributions with $\lambda < \mu$. We numerically approximate the posterior distribution of the parameters using MCMC. Specifically, we constructed a Markov chain that has as its states the parameters of the model and a stationary distribution that is the posterior distribution of interest. Samples from this chain when at stationarity are valid, albeit dependent, samples from the posterior distribution. Besides sampling the phylogenetic parameters of the model, the chain also samples word segment alignments[28].

The marginal likelihood, $f(\mathbf{S})$, plays the key role in choosing among models. We numerically approximate the marginal probability for a model using path sampling techniques[15-1-29].

We examined three models of segmental transition. The first (denoted 'JC69') is isomorphic to the Jukes-Cantor[30] model of molecular evolution that constrains all rates of change to be equal and has no free parameters to estimate; the second ('Linguistically Informed') model allows different rates among five groups of word segments (Nasal Vowel, Vowel, Nasal Consonant, Non Sylabic Sonorant and Consonant) and has an intermediate number of parameters (90 parameters for the equilibrium distribution and 15 parameters describing rates among word segment groups); and the third ('GTR') is a model isomorphic to the general time reversible model of molecular evolution[31], and has a large number of parameters (90 parameters for the equilibrium distribution and 3402 exchangability parameters). For both the second and third models, we consider the equilibrium frequencies of the model to be random variables with a flat Dirichlet prior distribution.

Data Curation and Analyses

Our analyses is coordinated using a program written in the Nytril programming language[32]. We created an automated framework for the experiment that comprises a reusable IPA library, word data, quality tests, external MCMC software (written in C++), and typeset output for papers and presentations.

Words segments from the 10 different human languages, coded in standard unicode IPA, are organized into data structures. For example, the concept 'dog' is coded in the following way:

```
Latin           = "k-a-nem";              // canis
French          = "ʃjɛ˜----";             // chien
Spanish         = "k-a-n--";              // can ˈpero perro
Italian         = "k-a-ne-";              // cane
PortugueseBrazil = "k-ɐ˜w˜--;"            // cão, cachorro"; Alt. ["kɐʃoʁu"]
Portuguese      = "k-ɐ˜w˜--;"             // cão, cachorro"; Alt. ["kɐʃoʁu"]
Catalan         = "k-a----";              // ca /ˈgos/ <gos>
Walloon         = "ʧ-ẽ----";              // tchén
Friulian        = "c-a-n--";              // cjan [caŋ]
Romanian        = "k-ɨjne-";              // câine
```

Here, the comments following the '//' are meant to be read by a reviewer, but have no effect on the analysis. A challenge for programmers working in linguistics is that the IPA representation of words are in Unicode (UCS), not ANSI. This requires care with structures, file formats, text encodings and diacritic compositional form. Furthermore, each text string can comprise several UCS code-points, so that there is not a one-to-one mapping between string elements and segments. For instance the the following segment uses 4 unicode code-points to represent a single sound.

ɐ̃w̃

`0250, 0303, 0077, 0303`

Mid Front Nasal Vowel
Diphthong

Since our analysis requires each word to be seperated into segments, a parser converts word strings into arrays of objects that facilitate the manipulation of segments at a high level. A library of of such objects endows each segment with linguistic features that can be used for analysis. For instance, the partition models are coded using properties such as 'consonant' and 'vowel', rather than working with strings of characters directly.

After the raw input is coded, the control program aggregates and partitions the words and builds the data files for the MCMC program to perform Bayesian phylogenetic analysis under the TKF91 model. The MCMC analyses runs for one million cycles and repeats to check for convergence. Samples taken during the first 10% of the chain are discarded as the burn-in phase. When the analysis is complete, the results are read back into the control program. The resulting tables, trees and graphical figures are combined with narrative elements created by the authors to make the final paper and presentation materials.

Acknowledgements

J.P.H. was supported through NSF (1759909) and the Koret Foundation. We thank Gerton Lunter for help with the TKF91 model.

Author contributions statement

D.M.G. collected data, S.H.M. developed the control software, and J.P.H. developed the MCMC software. All authors wrote the manuscript.

Figure 1. For the 'Linguistically Informed' model, states were grouped into five sets: Nasal Vowel (1), Vowel (2), Nasal Consonant (3), Non Sylabic Sonorant (4) and Consonant (5). Here, the area of the circles is proportional to the estimated equilibrium frequencies for each group. The width of the arrows is proportional to the estimated rates. Note that rates of change are much greater from one word segment to another when the change is within the word segment group than it is when the change is between word segments in different groups.

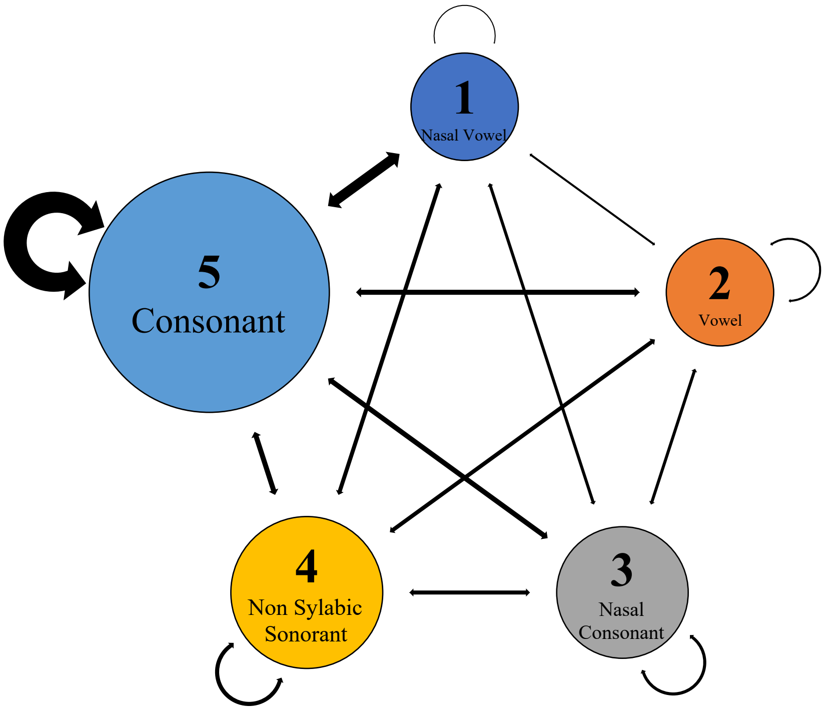
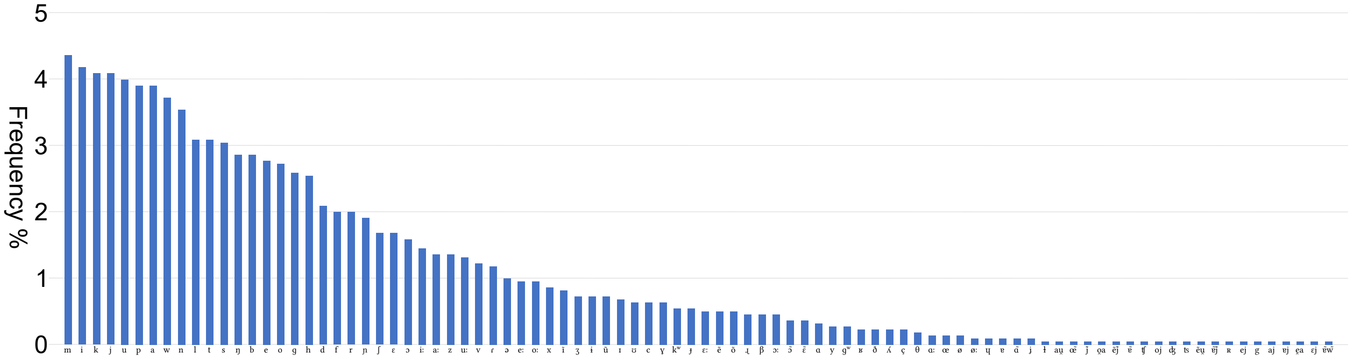


Figure 2. Frequency of occurance of segments in the lexicon[33]


Frequency %
m i k j u p a w n l t s ŋ b e o ɡ h d f r ɲ ʃ ɛ ɔ iː aː z uː v ɾ ə eː oː x ĩ ʒ ɨ ũ ɪ ʊ c ɣ kʷ ɟ ɛː ẽ õ ɻ β ɔː ɔ̃ ɛ̃ ɑ y ɡʷ ʁ ð ʎ ç θ ɑː œ ø øː ɥ ɐ ɑ̃ ʝ ɬ au̯ œ̃ j̃ o̯a ẽj̃ ɐ̃ ʧ oj ʤ ʦ ẽu̯ ɐ̃j̃ ʀ ej ɡ aj ɐj e̯a ɛj ɐ̃w̃