A Probabilistic Approach to Diachronic Phonology

Alexandre Bouchard-Côté* **Percy Liang*** **Thomas L. Griffiths†** **Dan Klein***
*Computer Science Division †Department of Psychology
University of California at Berkeley
Berkeley, CA 94720

Abstract

We present a probabilistic model of diachronic phonology in which individual word forms undergo stochastic edits along the branches of a phylogenetic tree. Our approach allows us to achieve three goals with a single unified model: (1) reconstruction of both ancient and modern word forms, (2) discovery of general phonological changes, and (3) selection among different phylogenies. We learn our model using a Monte Carlo EM algorithm and present quantitative results validating the model.

1 Introduction

Modeling how languages change phonologically over time (diachronic phonology) is a central topic in historical linguistics (Campbell, 1998). The questions involved range from reconstruction of ancient word forms, to the elucidation of phonological drift processes, to the determination of phylogenetic relationships between languages. However, this problem has received relatively little attention from the computational community. What work there is has focused on the reconstruction of phylogenies on the basis of a Boolean matrix indicating the properties of words in different languages (Gray and Atkinson, 2003; Evans et al., 2004; Ringe et al., 2002; Nakhleh et al., 2005).

In this paper, we present a novel framework, along with a concrete model and experiments, for the probabilistic modeling of diachronic phonology. We focus on the case where the words are etymological cognates across languages, e.g. French *faire* and Spanish *hacer* from Latin *facere* (to do). Given this information as input, we learn a model acting at the level of individual phoneme sequences, which can be used for reconstruction and prediction, Our model is fully generative, and can be used to reason about a variety of types of information. For example, we can observe a word in one or more modern languages, say French and Spanish, and query the corresponding word form in another language, say Italian. This kind of lexicon-filling has applications in machine translation. Alternatively, we can also reconstruct ancestral word forms or inspect the rules learned along each branch of a phylogeny to identify salient patterns. Finally, the model can be used as a building block in a system for inferring the topology of phylogenetic trees. We discuss all of these cases further in Section 4.

The contributions of this paper are threefold. First, the approach to modeling language change at the phoneme sequence level is new, as is the specific model we present. Second, we compiled a new corpus[1] and developed a methodology for quantitatively evaluating such approaches. Finally, we describe an efficient inference algorithm for our model and empirically study its performance.

1.1 Previous work

While our word-level model of phonological change is new, there have been several computational investigations into diachronic linguistics which are relevant to the present work.

The task of reconstructing phylogenetic trees

[1] nlp.cs.berkeley.edu/pages/historical.html

Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 887–896, Prague, June 2007.

for languages has been studied by several authors. These approaches descend from *glottochronology* (Swadesh, 1955), which views a language as a collection of shared cognates but ignores the structure of those cognates. This information is obtained from manually curated cognate lists such as the data of Dyen et al. (1997).

As an example of a cognate set encoding, consider the meaning "eat". There would be one column for the cognate set which appears in French as *manger* and Italian as *mangiare* since both descend from the Latin *mandere* (to chew). There would be another column for the cognate set which appears in both Spanish and Portuguese as *comer*, descending from the Latin *comedere* (to consume). If this were the only data, algorithms based on this data would tend to conclude that French and Italian were closely related and that Spanish and Portuguese were equally related. However, the cognate set representation has several disadvantages: it does not capture the fact that the cognate is closer between Spanish and Portuguese than between French and Spanish, nor do the resulting models let us conclude anything about the regular processes which caused these languages to diverge. Also, the existing cognate data has been curated at a relatively high cost. In our work, we track each word using an automatically obtained cognate list. While our cognates may be noisier, we compensate by modeling phonological changes rather than boolean mutations in cognate sets.

There has been other computational work in this broad domain. Venkataraman et al. (1997) describe an information theoretic measure of the distance between two dialects of Chinese. Like our approach, they use a probabilistic edit model as a formalization of the phonological process. However, they do not consider the question of reconstruction or inference in multi-node phylogenies, nor do they present a learning algorithm for such models.

Finally, for the specific application of cognate prediction in machine translation, essentially transliteration, there have been several approaches, including Kondrak (2002). However, the phenomena of interest, and therefore the models, are extremely different. Kondrak (2002) presents a model for learning "sound laws," general phonological changes governing two completely observed aligned cognate lists. His model can be viewed as a special case of ours using a simple two-node topology.

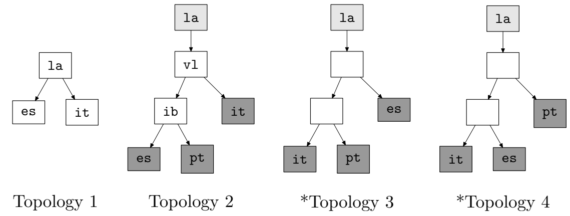


Figure 1: Tree topologies used in our experiments. *Topology 3 and *Topology 4 are incorrect evolutionary tree used for our experiments on the selection of phylogenies (Section 4.4).

There is also a rich literature (Huelsenbeck et al., 2001) on the related problems of evolutionary biology. A good reference on the subject is Felsenstein (2003). In particular, Yang and Rannala (1997), Mau and Newton (1997) and Li et al. (2000) each independently presented a Bayesian model for computing posteriors over evolutionary trees. A key difference with our model is that independence across evolutionary sites is assumed in their work, while the evolution of the phonemes in our model depends on the environment in which the change occurs.

2 A model of phonological change

Assume we have a fixed set of *word types* (cognate sets) in our vocabulary V and a set of languages L. Each word type i has a *word form* w_{il} in each language $l \in L$, which is represented as a sequence of phonemes and might or might not be observed. The languages are arranged according to some tree topology T (see Figure 1 for examples). One might consider models that simultaneously induce the topology and cognate set assignments, but let us fix both for now. We discuss one way to relax this assumption and present experimental results in Section 4.4.

Our generative model (Figure 3) specifies a distribution over the word forms $\{w_{il}\}$ for each word type $i \in V$ and each language $l \in L$. The generative process starts at the root language and generates all the word forms in each language in a top-down manner. One appealing aspect about our model is that, at a high-level, it reflects the actual phonological process that languages undergo. However, important phenomena like lexical drift, borrowing, and other non-phonological changes are not modeled.

Our generative model can be summarized as follows:

For each word $i \in V$:
  $w_{i\text{ROOT}} \sim$ LanguageModel
For each branch $(k \to l) \in T$:
  $\theta_{k \to l} \sim$ Dirichlet(α) [choose edit params.]
  For each word $i \in V$:
    $w_{il} \sim$ Edit$(w_{ik}, \theta_{k \to l})$ [sample word form]

In the remainder of this section, we describe each of the steps in the model.

2.1 Language model

For the distribution $w \sim$ LanguageModel, we used a simple bigram phoneme model. The phonemes were partitioned into *natural classes* (see Section 4 for details). A root word form consisting of n phonemes $x_1 \cdots x_n$ is generated with probability

$$p_{\text{lm}}(x_1) \prod_{j=2}^{n} p_{\text{lm}}(x_j \mid \text{NaturalClass}(x_{j-1})),$$

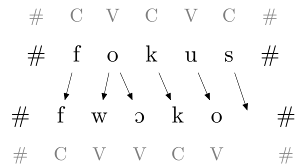
where p_{lm} is the distribution of the language model.

2.2 Edit model

The stochastic edit model $y \sim$ Edit(x, θ) describes how a single old word form $x = x_1 \cdots x_n$ changes along one branch of the phylogeny with parameters θ to produce a new word form y. This process is parameterized by rule probabilities $\theta_{k \to l}$, which are specific to branch $(k \to l)$.

The generative process is as follows: for each phoneme x_i in the old word form, walking from left to right, choose a rule to apply. There are three types of rules: (1) *deletion* of the phoneme, (2) *substitution* with another phoneme (possibly the same one), or (3) *insertion* of another phoneme, either before or after the existing one. The probability of applying a rule depends on a *context* $(\text{NaturalClass}(x_{i-1}), \text{NaturalClass}(x_{i+1}))$. Figure 2 illustrates the edits on an example. The context-dependence allows us to represent phenomena such as the fact that s is likely to be deleted only in word-final contexts.

The edit model we have presented approximately encodes a limited form of classic rewrite-driven segmental phonology (Chomsky and Halle, 1968). One



Edits applied

f	→	f	/ # _ V
o	→	w ɔ	/ C _ C
k	→	k	/ V _ V
u	→	o	/ C _ C
s	→		/ V _ #

Rules used

Figure 2: An example of edits that were used to transform the Latin word *FOCUS* (/fokus/) into the Italian word *fuoco* (/fwɔko/) (fire) along with the context-specific rules that were applied.

could imagine basing our model on more modern phonological theory, but the computational properties of the edit model are compelling, and it is adequate for many kinds of phonological change.

In addition to simple edits, we can model some classical changes that appear to be too complex to be captured by a single left-to-right edit model of this kind. For instance, bleeding and feeding arrangements occur when one phonological change introduces a new context, which triggers another phonological change, but the two cannot occur simultaneously. For example, vowel raising e → i / _ c might be needed before palatalization t → c / _ i. Instead of capturing such an interaction directly, we can break up a branch into two segments joined at an intermediate language node, conflating the concept of historically intermediate languages with the concept of intermediate stages in the application of sequential rules.

However, many complex processes are not well-represented by our basic model. One problematic case is chained shifts such as Grimm's law in Proto-Germanic or the Great Vowel Shift in English. To model such dependent rules, we would need to use a more complex prior distributions over the edit parameters. Another difficult case is prosodic changes, such as unstressed vowel neutralizations, which would require a representation of suprasegmental features. While our basic model does not account for these phenomena, extensions within the generative framework could capture such richness.

3 Learning and inference

We use a Monte Carlo EM algorithm to fit the parameters of our model. The algorithm iterates between a stochastic E-step, which computes recon-

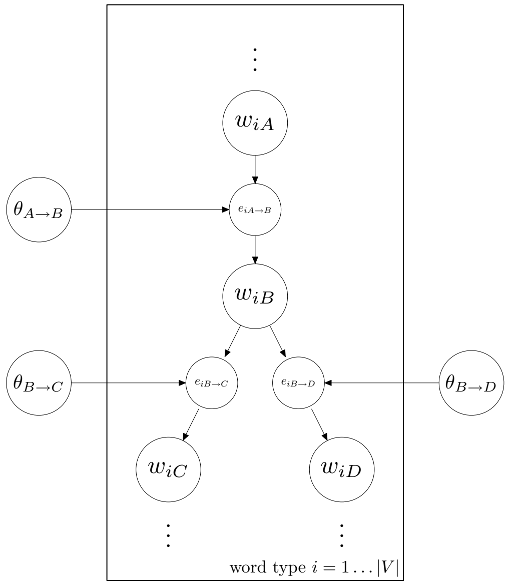


Figure 3: The graphical model representation of our model: θ are the parameters specifying the stochastic edits e, which govern how the words w evolve. The plate notation indicates the replication of the nodes corresponding to the evolving words.

structions based on the current edit parameters, and an M-step, which updates the edit parameters based on the reconstructions.

3.1 Monte Carlo E-step: sampling the edits

The E-step needs to produce expected counts of how many times each edit (such as o $\to$ ɔ) was used in each context. An exact E-step would require summing over all possible edits involving all languages in the phylogeny (all unobserved $\{e\}, \{w\}$ variables in Figure 3). Unfortunately, unlike in the case of HMMs and PCFGs, our model permits no tractable dynamic program to compute these counts exactly.

Therefore, we resort to a Monte Carlo E-step, where many samples of the edit variables are collected, and counts are computed based on these samples. Samples are drawn using Gibbs sampling (Geman and Geman, 1984): for each word form of a particular language w_{il}, we fix all other variables in the model and sample w_{il} along with its corresponding edits.

In the E-step, we fix the parameters, which renders the word types conditionally independent, just as in an HMM. Therefore, we can process each word type in turn without approximation.

First consider the simple 4-language topology in Figure 3. Suppose that the words in languages A, C and D are fixed, and we wish to infer the word at language B along with the three corresponding sets of edits (remember the edits fully determine the words). There are an exponential number of possible words/edits, but it turns out that we can exploit the Markov structure in the edit model to consider all such words/edits using dynamic programming, in a way broadly similar to the forward-backward algorithm for HMMs.

Figure 4 shows the lattice for the dynamic program. Each path connecting the two shaded endpoint states represents a particular word form for language B and a corresponding set of edits. Each node in the lattice is a state of the dynamic program, which is a 5-tuple $(i_A, i_C, i_D, c_1, c_2)$, where i_A, i_C and i_D are the cursor positions (represented by dots in Figure 4) in each of the word forms of A, C and D, respectively; c_1 is the natural class of the phoneme in the word form for B that was last generated; and c_2 corresponds to the phoneme that will be generated next.

Each state transition involves applying a rule to A's current phoneme (which produces 0–2 phonemes in B) and applying rules to B's new 0–2 phonemes. There are three types of rules (deletion, substitution, insertion), resulting in $3^0+3^2+3^4 = 91$ types of state transitions. For illustration, Figure 4 shows the simpler case where B only has one child C. Given these rules, the new state is computed by advancing the appropriate cursors and updating the natural classes c_1 and c_2. The weight of each transition $w(s \to t)$ is a product of the language model probability and the rule probabilities that were chosen.

For each state s, the dynamic program computes $W(s)$, the sum of the weights of all paths leaving s,

$$W(s) = \sum_{s \to t} w(s \to t) W(t).$$

To sample a path, we start at the leftmost state, choose the transition with probability proportional to its contribution in the sum for computing $W(s)$, and repeat until we reach the rightmost state.

We applied a few approximations to speed up the sampling of words, which reduced the running time by several orders of magnitude. For example, we pruned rules with low probability and restricted the

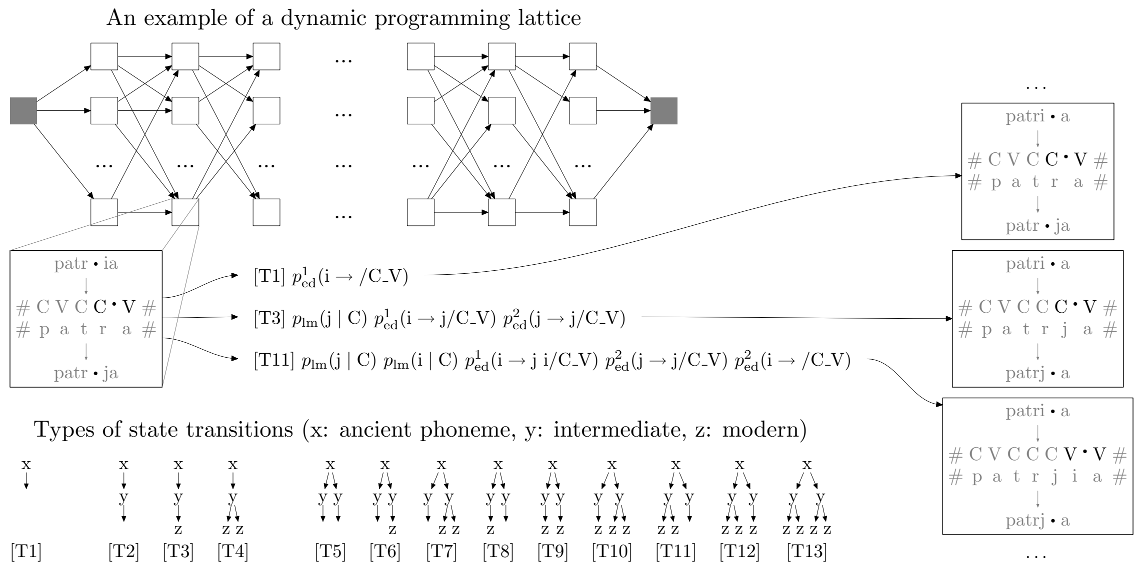


Figure 4: The dynamic program involved in sampling an intermediate word form given one ancient and one modern word form. One lattice node is expanded to show the dynamic program state (represented by the part not grayed out) and three of the many possible transitions leaving the state. Each transition is labeled with the weight of the transition, which is the product of the relevant model probabilities. At the bottom, the 13 types of state transitions are shown.

state space of the dynamic program by limiting the deviation in cursor positions.

3.2 M-step: updating the parameters

The M-step is standard once we have computed the expected counts of edits in the E-step. For each branch $(k \rightarrow l) \in T$ in the phylogeny, we compute the maximum likelihood estimate of the edit parameters $\{\theta_{k \rightarrow l}(x \rightarrow \beta \,/\, c_1 \,\_\, c_2)\}$. For example, the parameter corresponding to $x = $ /e/, $\beta = $ /e s/, $c_1 = $ ALVEOLAR, $c_2 = $ # is the probability of inserting a final /s/ after an /e/ which is itself preceded by an alveolar phoneme. The probability of each rule is estimated as follows:

$$\theta_{k \rightarrow l}(x \rightarrow \beta \,/\, c_1 \,\_\, c_2) = \frac{\#(x \rightarrow \beta \,/\, c_1 \,\_\, c_2) + \alpha(x \rightarrow \beta \,/\, c_1 \,\_\, c_2) - 1}{\sum_{\beta'} \#(x \rightarrow \beta' \,/\, c_1 \,\_\, c_2) + \alpha(x \rightarrow \beta' \,/\, c_1 \,\_\, c_2) - 1},$$

where α is the concentration hyperparameter of the Dirichlet prior. The value $\alpha - 1$ can be interpreted as the number of pseudocounts for a rule.

4 Experiments

In this section we show the results of our experiments with our model. The experimental conditions are summarized in Table 1, with additional information on the specifics of the experiments presented in Section 4.5. We start with a description of the corpus we created for these experiments.

Experiment	Topology	Heldout
Latin reconstruction (4.2)	1	`la`:293
Italian reconstruction (4.2)	1	`it`:117
Sound changes (4.3)	2	None
Phylogeny selection (4.4)	2, 3, 4	None

Table 1: Conditions under which each of the experiments presented in this section were performed. The *topology* indices correspond to those displayed in Figure 1. Note that by conditional independence, the topology used for Spanish reconstruction reduces to a chain. The heldout column indicates how many words, if any, were heldout for edit distance evaluation, and from which language.

4.1 Corpus

In order to train and evaluate our system, we compiled a corpus of Romance cognate words. The raw data was taken from three sources: the `wiktionary.org` website, a Bible parallel corpus (Resnik et al., 1999) and the Europarl corpus (Koehn, 2002). From an XML dump of the Wiktionary data, we extracted multilingual translations, which provide a list of word tuples in a large number of languages, including a few ancient languages.

The Europarl and the biblical data were processed and aligned in the standard way, using combined GIZA++ alignments (Och and Ney, 2003).

We performed our experiments with four languages from the Romance family (Latin, Italian, Spanish, and Portuguese). For each of these languages, we used a simple in-house rule-based system to convert the words into their IPA representations.[2] After augmenting our alignments with the *transitive closure*[3] of the Europarl, Bible and Wiktionary data, we filtered out non-cognate words by thresholding the ratio of edit distance to word length.[4] The preprocessing is constraining in that we require that all the elements of a tuple to be cognates, which leaves out a significant portion of the data behind (see the row *Full entries* in Table 2). However, our approach relies on this assumption, as there is no explicit model of non-cognate words. An interesting direction for future work is the joint modeling of phonology with the determination of the cognates, but our simpler setting lets us focus on the properties of the edit model. Moreover, the restriction to full entries has the side advantage that the Latin bottleneck prevents the introduction of too many neologisms, which are numerous in the Europarl data, to the final corpus.

Since we used automatic tools for preparing our corpus rather than careful linguistic analysis, our cognate list is much noiser in terms of the presence of borrowed words and phonemeic transcription errors compared to the ones used by previous approaches (Swadesh, 1955; Dyen et al., 1997). The benefit of our mechanical preprocessing is that more cognate data can easily be made available, allowing us to effectively train richer models. We show in the rest of this section that our phonological model can indeed overcome this noise and recover meaningful patterns from the data.

[2]The tool and the rules we used are available at `nlp.cs.berkeley.edu/pages/historical.html`.

[3]For example, we would infer from an `la-es` bible alignment *confessionem-confesión* (confession) and an `es-it` Europarl alignment *confesión-confessione* that the Latin word *confessionem* and the Italian word *confessione* are related.

[4]To be more precise we keep a tuple $(w_1, w_2, \ldots, w_p)$ iff $\frac{d(w_i,w_j)}{\bar{l}(w_i,w_j)} \leq 0.7$ for all $i, j \in \{1, 2, \ldots, p\}$, where $\bar{l}$ is the mean length $\frac{|w_i|+|w_j|}{2}$ and d is the Levenshtein distance.

Name	Languages	Tuples	Word forms
Raw sources of data used to create the corpus			
Wiktionary	`es,pt,la,it`	5840	11724
Bible	`la,es`	2391	4782
Europarl	`es,pt`	36905	73773
	`it,es`	39506	78982
Main stages of preprocessing of the corpus			
Closure	`es,pt,la,it`	40944	106090
Cognates	`es,pt,la,it`	27996	69637
Full entries	`es,pt,la,it`	586	2344

Table 2: Statistics of the dataset we compiled for the evaluation of our model. We show the languages represented, the number of tuples and the number of word forms found in each of the source of data and pre-processing steps involved in the creation of the dataset we used to test our model. By *full entry*, we mean the number of tuples that are jointly considered cognate by our preprocessing system and that have a word form known for each of the languages of interest. These last row forms the dataset used for our experiments.

Language	Baseline	Model	Improvement
Latin	2.84	2.34	9%
Spanish	3.59	3.21	11%

Table 3: Results of the edit distance experiment. The *language* column corresponds to the language held-out for evaluation. We show the mean edit distance across the evaluation examples.

4.2 Reconstruction of word forms

We ran the system using Topology 1 in Figure 1 to demonstrate the the system can propose reasonable reconstructions of Latin word forms on the basis of modern observations. Half of the Latin words at the root of the tree were held out, and the (uniform cost) Levenshtein edit distance from the predicted reconstruction to the truth was computed. Our baseline is to pick randomly, for each heldout node in the tree, an observed neighboring word (i.e. copy one of the modern forms). We stopped EM after 15 iterations, and reported the result on a Viterbi derivation using the parameters obtained. Our model outperformed this baseline by a 9% relative reduction in average edit distance. Similarly, reconstruction of modern forms was also demonstrated, with an improvement of 11% (see Table 3).

To give a qualitative feel for the operation of the system (good and bad), consider the example in Figure 5, taken from this experiment. The Latin *dentis* /dɛntis/ (teeth) is nearly correctly reconstructed as /dɛntes/, reconciling the appearance of the /j/ in the

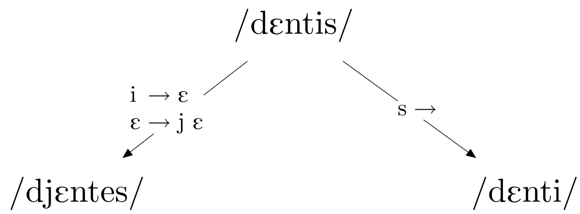


Figure 5: An example of a Latin reconstruction given the Spanish and Italian word forms.

Spanish and the disappearance of the final /s/ in the Italian. Note that the /is/ vs. /es/ ending is difficult to predict in this context (indeed, it was one of the early distinctions to be eroded in vulgar Latin).

While the uniform-cost edit distance misses important aspects of phonology (all phoneme substitutions are not equal, for instance), it is parameter-free and still seems to correlate to a large extent with linguistic quality of reconstruction. It is also superior to held-out log-likelihood, which fails to penalize errors in the modeling assumptions, and to measuring the percentage of perfect reconstructions, which ignores the degree of correctness of each reconstructed word.

4.3 Inference of phonological changes

Another use of our model is to automatically recover the phonological drift processes between known or partially known languages. To facilitate evaluation, we continued in the well-studied Romance evolutionary tree. Again, the root is Latin, but we now add an additional modern language, Portuguese, and two additional hidden nodes. One of the nodes characterizes the least common ancestor of modern Spanish and Portuguese; the other, the least common ancestor of all three modern languages. In Figure 1, Topology 2, these two nodes are labelled `vl` (Vulgar Latin) and `ib` (Proto-Ibero Romance) respectively. Since we are omitting many other branches, these names should not be understood as referring to actual historical proto-languages, but, at best, to collapsed points representing several centuries of evolution. Nonetheless, the major reconstructed rules still correspond to well known phenomena and the learned model generally places them on reasonable branches.

Figure 6 shows the top four general rules for each of the evolutionary branches in this experiment, ranked by the number of times they were used in the derivations during the last iteration of EM. The `la`, `es`, `pt`, and `it` forms are fully observed while the `vl` and `ib` forms are automatically reconstructed. Figure 6 also shows a specific example of the evolution of the Latin *VERBUM* (word/verb), along with the specific edits employed by the model.

While quantitative evaluation such as measuring edit distance is helpful for comparing results, it is also illuminating to consider the plausibility of the learned parameters in a historical light, which we do here briefly. In particular, we consider rules on the branch between `la` and `vl`, for which we have historical evidence. For example, documents such as the *Appendix Probi* (Baehrens, 1922) provide indications of orthographic confusions which resulted from the growing gap between Classical Latin and Vulgar Latin phonology around the 3rd and 4th centuries AD. The *Appendix* lists common misspellings of Latin words, from which phonological changes can be inferred.

On the `la` to `vl` branch, rules for word-final deletion of classical case markers dominate the list (rules ranks 1 and 3 for deletion of final /s/, ranks 2 and 4 for deletion of final /m/). It is indeed likely that these were generally eliminated in Vulgar Latin. For the deletion of the /m/, the *Appendix Probi* contains pairs such as *PASSIM NON PASSI* and *OLIM NON OLI*. For the deletion of final /s/, this was observed in early inscriptions, e.g. *CORNELIO* for *CORNELIOS* (Allen, 1989). The frequent leveling of the distinction between /o/ and /u/ (rules ranked 5 and 6) can be also be found in the *Appendix Probi*: *COLUBER NON COLOBER*. Note that in the specific example shown, the model lowers the orignal /u/ and then re-raises it in the `pt` branch due to a latter process along that branch.

Similarily, major canonical rules were discovered in other branches as well, for example, /v/ to /b/ fortition in Spanish, /s/ to /z/ voicing in Italian, palatalization along several branches, and so on. Of course, the recovered words and rules are not perfect. For example, reconstructed Ibero /trinta/ to Spanish /treinta/ (thirty) is generated in an odd fashion using rules /e/ to /i/ and /n/ to /in/. Moreover, even when otherwise reasonable systematic sound changes are captured, the crudeness of our fixed-granularity contexts can prevent the true context

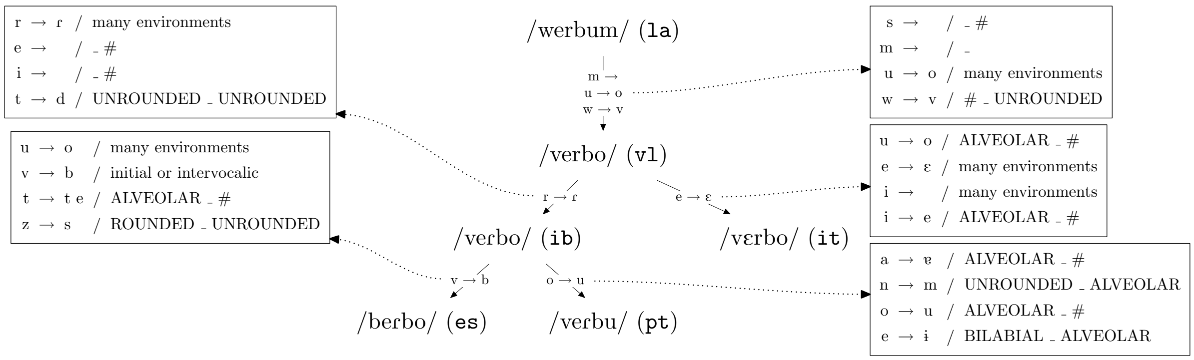


Figure 6: The tree shows the system's hypothesised derivation of a selected Latin word form, *VERBUM* (word/verb) into the modern Spanish, Italian and Portuguese pronunciations. The Latin root and modern leaves were observed while the hidden nodes as well as all the derivations were obtained using the parameters computed by our model after 15 iterations of EM. Nontrivial rules (i.e. rules that are not identities) used at each stage are shown along the corresponding edge. The boxes display the top four nontrivial rules corresponding to each of these evolutionary branches, ordered by the number of time they were applied during the last E round of sampling. Note that since our natural classes are of fixed granularity, some rules must be redundantly discovered, which tends to flood the top of the rule lists with duplicates of the top few rules. We summarized such redundancies in the above tables.

from being captured, resulting in either rules applying with low probability in overly coarse environments or rules being learned redundantly in overly fine environments.

4.4 Selection of phylogenies

In this experiment, we show that our model can be used to select between various topologies of phylogenies. We first presented to the algorithm the universally accepted evolutionary tree corresponding to the evolution of Latin into Spanish, Portuguese and Italian (Topology 2 in Figure 1). We estimated the log-likelihood L^* of the data under this topology. Next, we estimated the log-likelihood L' under two defective topologies (*Topology 3 and *Topology 4). We recorded the log-likelihood ratio $L^* - L'$ after the last iteration of EM. Note that the two likelihoods are comparable since the complexity of the two models is the same.[5]

We obtained a ratio of $L^* - L' = -4458 - (-4766) = 307$ for Topology 2 versus *Topology 3, and $-4877 - (-5125) = 248$ for Topology 2 versus *Topology 4 (the experimental setup is described in Table 1). As one would hope, this log-likelihood ratio is positive in both cases, indicating that the system prefers the true topology over the wrong ones.

While it may seem, at the first glance, that this result is limited in scope, knowing the relative arrangement of all groups of four nodes is actually sufficient for constructing a full-fledged phylogenetic tree. Indeed, *quartet-based* methods, which have been very popular in the computational biology community, are precisely based on this fact (Erdos et al., 1996). There is a rich literature on this subject and approximate algorithms exist which are robust to misclassification of a subset of quartets (Wu et al., 2007).

4.5 More experimental details

This section summarizes the values of the parameters we used in these experiments, their interpretation, and the effect of setting them to other values.

The Dirichlet prior on the parameters can be interpreted as adding pseudocounts to the corresponding edits. It is an important way of infusing parsimony into the model by setting the prior of the self-substitution parameters much higher than that of the other parameters. We used 6.0 as the prior on the self-substitution parameters, and for all environments, 1.1 was divided uniformly across the other edits. As long as the prior on self-substitution is kept within this rough order of magnitude, varying them has a limited effect on our results. We also initialized the parameters with values that encourage self-substitutions. Again, the results were robust to perturbation of initialization as long as the value for self-substitution dominates the other parameters.

The experiments used two natural classes for vowels (rounded and unrounded), and six natural

[5] If a word was not reachable in one of the topology, it was ignored in both models for the computation of the likelihoods.

classes for consonants, based on the place of articulation (alveolar, bilabial, labiodental, palatal, postalveolar, and velar). We conducted experiments to evaluate the effect of using different natural classes and found that finer ones can help if enough data is used for training. We defer the meticulous study of the optimal granularity to future work, as it would be a more interesting experiment under a log-linear model. In such a model, contexts of different granularities can coexist, whereas such coexistence is not recognized by the current model, giving rise to many duplicate rules.

We estimated the bigram phoneme model on the words in the root languages that were not heldout. Just as in machine translation, the language model was found to contribute significantly to reconstruction performance. We tried to increase the weight of the language model by exponentiating it to a power, as is often done in NLP applications, but we did not find that it had any significant impact on performance.

In the reconstruction experiments, when the data was not reachable by the model, the word used in the initialization was used as the prediction, and the evolution of these words were ignored when re-estimating the parameters. Words were initialized by picking at random, for each unobserved node, an observed node's corresponding word.

5 Conclusion

We have presented a novel probabilistic model of diachronic phonology and an associated inference procedure. Our experiments indicate that our model is able to both produce accurate reconstructions as measured by edit distance and identify linguistically plausible rules that account for the phonological changes. We believe that the probabilistic framework we have introduced for diachronic phonology is promising, and scaling it up to richer phylogenetic may indeed reveal something insightful about language change.

6 Acknowledgement

We would like to thank Bonnie Chantarotwong for her help with the IPA converter and our reviewers for their comments. This work was supported by a FQRNT fellowship to the first author, a NDSEG fellowship to the second author, NSF grant number BCS-0631518 to the third author, and a Microsoft Research New Faculty Fellowship to the fourth author.

References

W. Sidney Allen. 1989. *Vox Latina: The Pronunciation of Classical Latin*. Cambridge University Press.

W.A. Baehrens. 1922. *Sprachlicher Kommentar zur vulgärlateinischen Appendix Probi*. Halle (Saale) M. Niemeyer.

L. Campbell. 1998. *Historical Linguistics*. The MIT Press.

N. Chomsky and M. Halle. 1968. *The Sound Pattern of English*. Harper & Row.

I. Dyen, J.B. Kruskal, and P. Black. 1997. FILE IE-DATA1. Available at http://www.ntu.edu.au/education/langs/ielex/IE-DATA1.

P. L. Erdos, M. A. Steel, L. A. Szekely, and T. J. Warnow. 1996. Local quartet splits of a binary tree infer all quartet splits via one dyadic inference rule. Technical report, DIMACS.

S. N. Evans, D. Ringe, and T. Warnow. 2004. Inference of divergence times as a statistical inverse problem. In P. Forster and C. Renfrew, editors, *Phylogenetic Methods and the Prehistory of Languages*. McDonald Institute Monographs.

Joseph Felsenstein. 2003. *Inferring Phylogenies*. Sinauer Associates.

S. Geman and D. Geman. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741.

R. D. Gray and Q. Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origins. *Nature*.

John P. Huelsenbeck, Fredrik Ronquist, Rasmus Nielsen, and Jonathan P. Bollback. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*.

P. Koehn. 2002. Europarl: A Multilingual Corpus for Evaluation of Machine Translation.

G. Kondrak. 2002. *Algorithms for Language Reconstruction*. Ph.D. thesis, University of Toronto.

S. Li, D. K. Pearl, and H. Doss. 2000. Phylogenetic tree construction using Markov chain Monte Carlo. *Journal of the American Statistical Association*.

Bob Mau and M.A. Newton. 1997. Phylogenetic inference for binary data on dendrograms using markov chain monte carlo. *Journal of Computational and Graphical Statistics*.

L. Nakhleh, D. Ringe, and T. Warnow. 2005. Perfect phylogenetic networks: A new methodology for reconstructing the evolutionary history of natural languages. *Language*, 81:382–420.

F. J. Och and H. Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29:19–51.

P. Resnik, Mari Broman Olsen, and Mona Diab. 1999. The bible as a parallel corpus: Annotating the "book of 2000 tongues". *Computers and the Humanities*, 33(1-2):129–153.

D. Ringe, T. Warnow, and A. Taylor. 2002. Indo-european and computational cladistics. *Transactions of the Philological Society*, 100:59–129.

M. Swadesh. 1955. Towards greater accuracy in lexicostatistic dating. *Journal of American Linguistics*, 21:121–137.

A. Venkataraman, J. Newman, and J.D. Patrick. 1997. A complexity measure for diachronic chinese phonology. In J. Coleman, editor, *Computational Phonology*. Association for Computational Linguistics.

G. Wu, J. A. You, and G. Lin. 2007. Quartet-based phylogeny reconstruction with answer set programming. *IEEE/ACM Transactions on computational biology*, 4:139–152.

Ziheng Yang and Bruce Rannala. 1997. Bayesian phylogenetic inference using dna sequences: A markov chain monte carlo method. *Molecular Biology and Evolution 14*.