CRII: RI: Computational Historical Linguistics through Simulations

1 Introduction

The aim of this project is to employ simulations to understand the usefulness and limits of recent computational approaches: both Bayesian phylogenetic methods and machine learning classifier approaches applied to historical linguistics problems – cognate detection, phylogenetic inference, and ancestral language reconstruction.

Linguists classify languages into families based on genetic relatedness i.e., by determining if they have descended from a common ancestor. The Indo-European family is a group of related languages consisting of languages – Hindi, Russian, and English – spoken across a wide geographical range from India to Europe. The ancestral language called Proto-Indo-European has been reconstructed using the classical comparative method developed in the 19^{th} century [Fox, 1995, Anttila, 1972]. The comparative method relies on identifying cognates (words similar in sound, meaning and traceable to a common ancestor) to group languages together. English *ten* is cognate with German *Zehn* whose ancestral word (Proto-Germanic **tehun*; Ringe, 2017, 81) is not recorded but reconstructed using the comparative method. English, German, and Dutch belong to the Germanic family, whose ancestor is Proto-Germanic, which is a subgroup of the the larger Indo-European language family.

Until 2000s, the comparative method has been primarily a manual enterprise, when parts of the comparative method such as tree inference [Gray and Atkinson, 2003], automated cognate detection [List, 2012, Jäger et al., 2017, Rama, 2016, Hauer and Kondrak, 2011], proto-word reconstruction [Bouchard-Côté et al., 2013, Dekker and Zuidema, 2020], and reflex prediction [Cathcart and Rama, 2020] have been attempted through machine learning and Bayesian phylogenetic models. The new area of interest has come to known as *computational historical linguistics* [Jäger, 2019].

The Bayesian phylogenetic models [Ronquist et al., 2012, Bouckaert et al., 2014] from evolutionary biology, have been applied to linguistically well-studied (where the comparative method has been applied) different languages families of the world such as Indo-European [Bouckaert et al., 2012], Australian [Bowern and Atkinson, 2012], Bantu [Grollemund et al., 2015], and Austronesian [Gray et al., 2009] to infer the internal structure and time-scale. These Bayesian phylogenetic models consist of a cognate gain-loss model which is not originally designed for linguistics but for modeling presence or absence of a trait such as "number of antenna segments in ants" [Wright, 2019]. Apart from few studies that evaluated the performance at dating [Chang et al., 2015, Rama, 2018b], required data size [Rama and Wichmann, 2018], identification of the geographic origins of a family [Wichmann and Rama, 2021], mode of language evolution [Holman and Wichmann, 2017], **the performance limitations of the Bayesian phylogenetic models and machine learning models are not yet well understood**.

I propose to address the limitations of the different models through synthetic data (workflow given in figure 1). My research aims in this project are as follows:

- **Aim 1:** Generate synthetic data to perform extensive evaluation of different lexical models used in Bayesian phylogenetics. This evaluation would accompanied by development of linguistically interpretable models of cognate change.

- **Aim 2:** Evaluate, develop and interpret automated cognate detection models on synthetic data and test on real world data.
- **Aim 3:** Test *sequence to sequence* models at ancestral and reflex prediction tasks on synthetic and real world data.

Why synthetic data? The comparative reconstruction is supposed to have an upper limit of about 10,000 years known as time-depth problem [Trask, 2000, 344] beyond which the relationships between languages cannot be identified accurately since the amount of changes over such a long period obscured any evidence available to the comparative method [Harrison, 2003]. Therefore, synthetic data is a good option for understanding linguistic evolution beyond the recorded history. Second, many language families of the world do not have recorded history and as such it not directly possible to evaluate the performance of the computational methods at ancestral prediction tasks. Understanding the limitations of the computational models would allow future researchers to be well-informed regarding the choice of the model. The synthetic data seeded with existing language data can be used to train deep learning models for the tasks of automated cognate detection and ancestral word prediction.

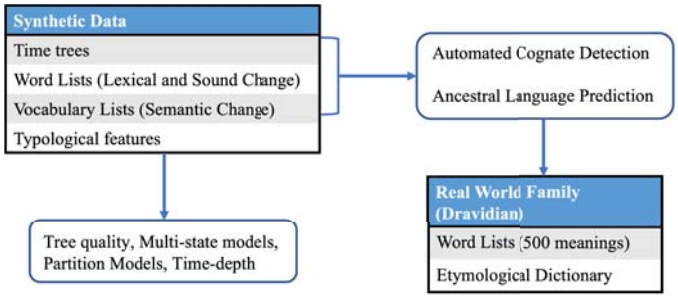

Figure 1: Testing computational models on synthetic and real world data.

Biological models designed for DNA or morphological data has been widely adopted in historical linguistics but the models themselves are not designed for linguistic data. Unlike biology where the size of alphabet is fixed, languages tend to exhibit a wide range of phoneme inventories ranging from eight consonants in Hawaai'an to more than a hundred consonants in the Taa language spoken in Botswana [Moran et al., 2014]. Discussions on parallels between biological evolution and linguistic evolution [Atkinson and Gray, 2005, List, 2016] suggest that the parallels are shared but not exact. There is a need for research in developing linguistically suitable simulations that model **linguistically unique changes** happening along phonology, lexicon, and meaning dimensions; and, to test the computational approaches exhaustively at different historical linguistic tasks using the synthetic data.

If the project **succeeds**, (1) it would better inform the practitioners of the phylogenetic models for informing the choice of models, (2) to train deep learning models to infer cognates and ancestral forms for less-studied and less-resourced language families, and (3) linguistically interpret the model parameters.

2 Background and Pilot Work

In computational linguistics, synthetic data has been used to train sequence to sequence models at the task of morphological inflection [Bergmanis et al., 2017, Anastasopoulos and Neubig, 2019] for low resource languages and understanding the limitations of neural networks at structural prediction tasks [Wang and Eisner, 2016]. In computational historical linguistics, **word list simulations** were originally developed in Holman and Wichmann [2017], extended to incorporate geographic simulations [Kapur and Rogers, 2020, Wichmann, 2017]. The programs used the existing ASJP

code (Automated Similarity Judgment Program; Brown et al., 2008), a set of 41 symbols (including modifiers for representing aspiration) used for transcribing words in more than 75% of the world's languages.[1] The simulations used weighted correspondences (prominent alignments; Brown et al., 2013) to edit the words to simulate sound changes such as consonant and vowel insertion or deletion in addition to metathesis (transpositions) to generate synthetic word lists *but do not incorporate context-dependent changes*. Lexical change (replacement of an existing word with a new word: could be through borrowing or morphology; the word "pakṣi" in the Kolami language (table 1) is a Sanskrit word not a native Dravidian word) is implemented as relative to the amount of phonological change. In this project, our simulations are along the dimensions of lexicon and borrowing.

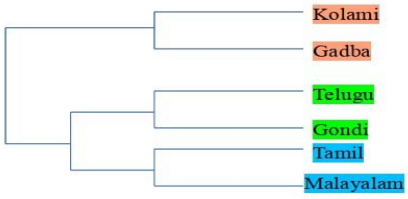

Figure 2: Phylogenetic Tree

Language	BIRD	CLASSES	Multi	Binary
Kolami	p i ʈ: a	PITA	3	0 0 1 0
Tamil	p a ɾ a ʋ a i	PARAWAI	2	0 1 0 0
Telugu	p i ʈ: a	PITA	3	0 0 1 0
Gondi	p i ʈ: ə	PITE	3	0 0 1 0
Gadba	t i t ə	TITE	4	0 0 0 1
Malayalam	p a k ʂi	PAKSI	1	0 0 0 0

Table 1: Different groups of Dravidian language family: South, South-Central, Central

Research in Bayesian linguistic phylogenetics employed binary encoding of cognate sets in multilingual word lists (column **Binary** in table 1), typically Swadesh word lists consisting of 100-200 items [Swadesh, 1955] to infer phylogenies. There hasn't been much work on testing 'morphological models' from biology that handle **multistate data** [Wright, 2019, Wright et al., 2016] shown in column **Multi** in table 1. In a multistate model, the transitions between the cognate sets would be handled using a instantaneous rate model called *Mk* model [Lewis, 2001]. An example of non-lexical multi-state data is the `number of consonants in languages` [Maddieson, 2013] in the WALS database[2] that is continuous in nature but is split into five categories ranging from *small* to *large* which are binarized for phylogenetic inference [Greenhill et al., 2017]. Apart from the early work of Pagel and Meade [2006] and Ringe et al. [2002], there is **not much work on using multistate models** for inferring phylogenies based on lexical cognate data in Bayesian framework.

Simulations [Atkinson et al., 2005, Barbançon et al., 2013] have been used to test the effect of **borrowing** on the inferred root age of the Indo-European tree using Bayesian phylogenetic methods, where the simulations generate *characters*[3] and *not the words* themselves.

Among-site Rates Models are used to model the reality that not all characters evolve at the same rate. The most common model used in linguistic phylogenetics is the Gamma site rates model [Yang, 1994] assumes that the rates are drawn from a discretized Gamma distribution (with four categories with the mean rate for each category) where the value of α (shape parameter) determines if the sites evolve at equal rates ($\alpha \to \infty$) or with very few sites ($\alpha \to 0$) having low rates of evolution. Non-parametric model based on the Dirichlet process where sites belonging to the same cluster share the same rate [Lartillot and Philippe, 2004] and **pilot work** using RevBayes [Höhna et al., 2016] on a dataset of the Bantu languages suggested that the number of clusters is three which is similar to the result found by Greenhill et al. [2017]; and, that lexical cognates evolve at a much slower pace than typological features (such as consonant inventory sizes) where

[1] https://asjp.clld.org/languages

[2] https://wals.info/

[3] Here *character or site or trait* shows values for languages across a cognate set or a nucleotide value and is not a character as in character language models used in Computational Linguistics or Natural Language Processing.

typological features tend to change faster. **Earlier work by the PI** [Rama and Wichmann, 2018] provided *support to the linguistic hypothesis that meanings that yield fewer cognate sets are more stable than meanings with larger number of cognate sets* [Oswalt, 1970].

Partition Models In addition to lexical cognates, morphological rules and sound change rules have been used to infer the Indo-European phylogeny [Ringe et al., 2002]. The sound change rules and morphological rules were selected based on the linguistic expertise in the Indo-European language family. For instance a sound change such as Grimm's law (English *father* $\sim$ Sanskrit *pita*; *p > f) is coded as present in the Germanic languages but not in other Indo-European languages. There has *not been any other work on testing the usefulness of sound change rules for phylogenetic inference* which could mainly be due to the amount of scholarly work (Indo-European has a scholarship of 200 years) required to determine such sound change rules manually.

Incorporating sounds into phylogenetics Jäger and List [2015] coded the presence of a sound within an aligned column as a binary site which is then used to infer phylogenetic trees which were not as good a lexical cognate based phylogeny. Jäger [2018] found that binary characters encoding presence or absence of a sound class within a meaning improves tree quality in comparison to the lexical data alone. A study by Macklin-Cordes et al. [2021] showed that although phonemes have phylogenetic signal (in Australian families) it is not as strong as lexical cognates. Our **pilot work** [Rama and List, Submitted] involving sound class transition models that model the probabilities of change between sound class characters such as $A \rightleftharpoons E$ and $P \rightleftharpoons W$ of five different language families yielded trees that are close to the phylogenies (for 3/5 families) given in the openly available Glottolog [Hammarström et al., 2019] database. This work follows up on the conclusions from Hruschka et al. [2015] who suggest that alignments from the Turkic family dataset are useful for predicting ages. The conflicting results suggest that *incorporating sounds into the current phylogenetic models is an open problem.*

Telugu	p i ʈː a
Gondi	p i ʈː ə
*Tamil	p i ʈː a

Table 2: Sound Alignment for the meaning `BIRD`.

Automated Cognate Detection Automated cognate identification systems – binary Support Vector Machine (SVM) based classifiers [Jäger et al., 2017, Hauer and Kondrak, 2011] – are trained on string similarity measures such as edit distance or a weighted edit distance that differentially weight alignments either using a data-driven method such as pointwise mutual information [Wieling et al., 2009] or hand-crafted segment similarity [List, 2012]. There *has not been any investigation* if classifiers trained on sound correspondences (extracted as alignments) can actually perform better than features based on string similarities. The **PI's work** [Rama et al., 2018a] showed that different automated cognate detection methods output cognate sets that can be used to infer accurate phylogenies for five different well-studied language families. The automatically detected cognate sets were useful for dating the phylogenies [Rama and Wichmann, 2020] and investigating mode of language evolution [Jäger, 2018]. However, the limitations of these methods are not well-understood regarding the examples where they fail and succeed. **Pilot study** indicated that the cognate detection methods **fail at detecting highly divergent word pair** such as Armenian *erku* $\sim$ English 'two' and *are good at identifying loans that masquerade as cognates.*

Word prediction Sequence-to-sequence models from neural machine translation [Bahdanau et al., 2015, Luong et al., 2015] and morphological prediction tasks [Cotterell et al., 2018, Wu

and Cotterell, 2019] have been applied to reflex prediction [Cathcart and Rama, 2020], ancestral language reconstruction tasks [Meloni et al., 2021] and word prediction tasks in sister languages [Fourrier et al., 2021, Dekker and Zuidema, 2020] using language ID, meaning (BERT; [Devlin et al., 2018]), and syntax embeddings as cues. The performance of these models is worse in terms of word error rates (about 50%) but have a better phoneme error rate (about 25%; normalized edit distance) in different reconstruction methods suggesting a *hard problem*. Unlike morphological inflection prediction that uses morphological information to learn and predict a target word form, ancestral language prediction on multilingual word lists and vocabulary lists has access to meaning and language ID to predict a target word form.

Pilot work on Data Collection The PI and an hourly research assistant has expanded the 100 item list [Kolipakam et al., 2018] of 20 Dravidian languages to 40 languages and annotated them for cognate judgments (excerpt provided in table 1). The Dravidian family – 4 languages are literary and the rest are non-literary with a few endangered – is primarily spoken in South India. Phylogenetic inference suggested that the 100 items are not sufficient for inferring accurate phylogeny that places all the languages into their major subgroups accurately. One outcome of the project would be cognacy annotated expanded word lists of the Dravidian language family that will be a testbed for phylogenetic inference and automated cognate detection.

3 Proposed Research Plan

3.1 Aim I: Simulations and Phylogenetic Inference

The project team **will develop simulation programs to generate synthetic data and test various existing models of lexical and phonological change for the purpose of inferring phylogenies, evaluate the inferred phylogenies against the true synthetic phylogenies, and develop linguistically motivated models to handle sound change and lexical change.**

The `CLASSES` column in table 1 shows the mapping between phonetic alphabet and sound classes. An example of a sound class is the reduction of {ʈ, t} to `T`. We will *develop simulations to generate synthetic data in different sound classes* using sound class transition rates in LingPy library [List et al., 2019b]. **Words in International Phonetic Alphabet (IPA)** will be generated through **three** different strategies: (1) randomly map the sound class symbols into IPA symbols, (2) start with a IPA string from a word list for a language such as North EuraLex database [Dellert et al., 2020] and then modify the word through feature switching based on the features given in `PyCLTS` package [List et al., 2019a]. An example is given in figure 3. We will experiment with other feature sets such as PanPhon [Mortensen et al., 2016] to simulate word lists in IPA. (3) Train a LSTM (Long-Short term Memory Network; Hochreiter and Schmidhuber, 1997) sound class language model on the real world lists (data in table 1) to generate a starting word at the root of the tree that evolves along the branches of the tree. The outcome of the simulations will be multilingual word lists (shown in table 1) for language families with different sizes. Lexical change involving borrowing will be incorporated in our simulations. Borrowing both family-internal and family-external will be controlled through the lexical change probability parameter and new words are generated using a character-based LSTM model. By varying the amount of lexical replacement against phonological change we can synthesize families like Indo-European or a family with large number of phonological changes where the forms are cognate with one another but highly divergent

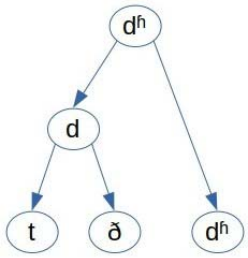

Figure 3: Sound change down a tree.

such as Hindi *chakra* and English *wheel* traceable to *$*k^wk^welo$-. The Hindi-English cognate can be traced back to the original Proto-Indo-European root where the English word went through a number of sound changes making the pair not so obviously cognate like *tongue* and German *Zunge*.

Planned experiments We will infer phylogenies on the synthesized datasets using Bayesian analyses and evaluate the phylogenies in the following settings: (i) Effect of number of meanings and cognate sets. (ii) Performance on datasets with low to high rates of lexical replacement. (iii) Effect of inclusion and exclusion of borrowings and homoplastic characters on phylogeny. (iv) Effect of *polymorphism*, when a meaning shows multiple forms. In these experiments, we will work with binary data and assume a binary continuous-time Markov chain model of evolution [Bouckaert et al., 2012] with Gamma site rates model that assumes that sites evolve at varying rates and will evaluate the inferred phylogenies against the true phylogeny using quartet distance [Christiansen et al., 2006]. All the simulations will be performed using `pylogeny` package that is built on the **software released by the PI** [Rama and List, 2019].

Multistate models for lexical change The morphological models used to model cognate gain-loss are from biology, this project will *develop new lexical models that are linguistically interpretable*. We will (1) develop **Reversible Jump models** [Green, 2003] that allow for some of the state transitions to be zero allowing us to determine the ordering of the states that might have evolved in a particular order, (2) **evaluate the preferred order of the state changes against the recorded lexical replacement** using synthetic data, (3) evaluate **the different models by comparing the inferred trees against the true trees from simulations and linguistically interpret the parameters**. A complimentary test that accompanies the phylogenetic inference is **ancestral state reconstruction test (ASR)** to verify if the phylogenetic algorithms can infer the right ancestral states for a meaning in terms of cognate sets. Based on the cognacy states in table 1 and the tree in figure 2, the original state at the root of the tree using *marginal or joint* reconstruction methods [Yang, 2014, Sec 4.4]. We will **evaluate the performance of the ASR algorithms on the synthetic cognacy data generated under varying levels of phonological change and lexical change**.

Effect of unobserved patterns In the context of morphological models, not all possible patterns are observed in the data. For instance, a site with all **0**s would not be found in linguistic data since a cognate set should be attested at least once to be included in the data. **0**s are accounted through a correction to the likelihood called *ascertainment bias correction* [Felsenstein, 1992]. The ascertainment bias correction makes a difference in the likelihood for datasets when the number of sites is less than 100. The correction used in the phylogenetic software assumes that the number of sites are fixed over the tree. However, the likelihood can be severely underestimated in the cases when the number of constant sites is quite high when compared to the number of observed sites [Tamuri and Goldman, 2017]. We will **test the effect of the number of unobserved sites on the quality of inferred phylogenies**.

Sound based phylogenetic inference We will investigate if sound-based alignments given in table 2 can yield accurate phylogenies. In the case of Tamil, the reflex was predicted (the sister language words prediction tasks given in section 3.3). As mentioned earlier, many of the language families' phylogenies (including Turkic family) are not fully resolved and it is **necessary to test the sound transition models** on synthetic datasets. In addition to testing the quality of the trees, we will develop and test the multi-state models (similar to the lexical change models) automatically

infer sound changes that are recorded in the simulations. Through partition models, **characters extracted from recorded sound changes to infer and evaluate phylogenetic trees**.

Meaning stability The stability indices for 1460 concepts from Loanword Typology project [Haspelmath and Tadmor, 2009] will be used to synthesize datasets that have realistic rates of lexical change. We will **test if the among-site rate and Reverse-jump models yield rates that correlate with the original rates.**

Language	Form	Meaning
Tamil	para	to fly, hover, flutter
Telugu	paracu	to run away, flee, flow
Malayalam	parakku	to fly, flee
Proto-Dravidian	*paːt	to run, flee

Table 3: Dravidian etymological dictionary [Burrow and Emeneau, 1984, 4020]

Typological features for phylogenetic inference that can be shared between any two languages of the world through chance or diffusion and not solely through *common descent* have been claimed to be useful for phylogenetic inference [Dunn et al., 2005] and doubted in Donohue et al., 2008 but supported later in a joint phylogenetic analysis of lexicon and structural features [Greenhill et al., 2017]. We will use the WALS database [Dryer and Haspelmath, 2013] to simulate such features down the tree [Wichmann and Holman, 2009] where a feature value is retained or changed according to a prior probability. We will **test if combinations of typological and lexical data help produce better phylogenies.**

Time limit test We will use the simulated phylogenies to test **if lexical cognate datasets are accurate at inferring the true phylogenies and the right dates** using dating methods [Zhang et al., 2016, Ronquist et al., 2016] from biology. These methods require fossil ages (attested languages such as Gothic which went extinct without leaving a descendant) or node calibration points to infer ages for the rest of the tree's nodes using Bayesian MCMC simulations. The simulation software has timesteps as another aspect which allow us to run the simulations for a large number of time steps. A time step can represent fixed number of geological years [Kapur and Rogers, 2020], a random number of years chosen according to a probability distribution, estimated from a paper that gamma regressed sub-treelengths to ages [Rama and Wichmann, 2020]. We **will evaluate the predicted age estimates for each internal node against the true ages.**

3.2 Aim II: Automated Cognate detection

We will **test the performance limitations of machine learning models for cognate detection through synthetic data**. The motivation is to *understand the limits of automated cognate detection methods and also to verify if automated cognate detection methods are useful* at inferring phylogenies when the true phylogeny is known before-hand which is not the case for most of the language families of the world. Next, we **will test if we can train such machine learning models trained on synthetic data (seeded from real word language family data) can be used for the purpose of testing on real world data.**

Through simulations, we will **investigate if alignments are useful as features for successful cognate detection in non-neural classifiers**. Neural network based classifiers, both convolutional and attention-coupled recurrent networks (LSTM) were trained on words where a phoneme is represented as a binary vector of articulatory features or through sound embeddings [Kumar et al., 2017, Rama, 2016] in a Siamese network that would share a layer of the neural network architecture to predict binary cognacy. We **will probe if the attention networks actually learn patterns similar to the sound correspondences present in the simulated data**. All the pair-wise scores will be supplied to a clustering algorithm (such as Dirichlet Process,

Rama, 2018a) to infer cognate sets. We will **test meaning embeddings such as BERT for computing meaning similarity** (table 3) for automated cognate detection.

We will **test different strategies to incorporate language distances from multiple sources such as geography, language distance through a weighted string similarity measure** [Wichmann et al., 2010] into the cognate classifiers. The SVM classifiers (**training**: Chinese dialects and Uralic languages; **testing**: Indo-European) yielded high F-scores that do not translate into accurate phylogenetic trees [Rama et al., 2018b]. In addition to the attention models [Wu and Cotterell, 2019, Luong et al., 2015], we will investigate the use of transformer models [Vaswani et al., 2017] for cognate detection task. We will investigate the use of different phonetic features from PanPhon and CLTS for initializing the phoneme embeddings.

We will test the **applicability** of simulations to real world data by training cognate classifiers on word lists synthetized by seeding the simulations with recorded languages words such as Telugu **to test on the multilingual word lists of the Dravidian language family collected as part of the project**.

3.3 Aim III: Word prediction

We will **perform a systematic evaluation of the different prediction models both LSTM and transformer models with geographical coordinates and meaning embeddings for predicting the sister language reflexes and ancestral word forms.**

Ancestral word prediction We will develop and evaluate several *seq2seq* models and variants of variational autoencoder [Kingma and Welling, 2013, VAE] models for ancestral word prediction tasks. **First** model would be based on multiple sequence alignment (table 2), to predict the ancestral word for a cognate set where a model inspired by HMMs using a transition model (for instance from the LingPy library) and language model probabilities (LSTM based) would be used to generate the ancestral word. **Second**, in a MCMC based sampling of ancestral words, a *seq2seq* model trained to predict cognates in sister languages with target language ID (and meaning representation) as input will be used to propose a candidate string at each internal node. The phylogenetic structure will be factored into the likelihood of the cognate set, computed through the same *seq2seq* model, given the sampled internal nodes' words. New words at each internal node can be proposed through the same *seq2seq* model by adding a k-hot vector (representing the k leaves under a node) as input. A VAE trained on the leaf words will generate and sample a latent representation ($z = \mu + \sigma\epsilon$, ϵ is Gaussian noise) for a cognate set at each internal node where z will be decoded (from the VAE's decoder) to yield an internal node word. z would be sampled at each internal node to yield a new word that can be used to estimate the likelihood of the observed cognate set using the *seq2seq* model trained on sister languages. The *seq2seq* model can be fine-tuned in both the cases in a cycle consisting of internal word sampling followed by a training step using SGD techniques. **Third**, the phylogenetic tree for the languages under study will be used to train a variational recursive autoencoder [Socher et al., 2011] where the decoder will be predict word forms at internal nodes of the tree for a cognate set. We will **test these models on the real world etymological dictionary of Dravidian language family** (see table 3).

Application to borrowing detection Automated methods using character based LSTM model on monolingual word lists tend to fail is at the task of detecting borrowings that are very close in terms of form but have different histories. For instance, the native word for *fire* in Telugu is *nippu* which occurs in the other Dravidian languages. Another word 'aggi' (related to Latin *igneous*) is borrowed from Indo-Aryan languages which can be detected through word prediction methods

(Burrow and Emeneau, 1984, nr. 2929). *Seq2seq* models (based on LSTMS and transformer models) can be used to **predict the reflex in a target language and compare against the attested form to detect if the candidate word is a borrowing or not**.

4 Work Plan

The work plan is given in table 4. The project will span over six semesters including two summer semesters. The simulations and the tests (**Part I**) involving different Bayesian phylogenetic models will be completed in the first semester. The development and testing of **multistate models** will be done in the second semester. The models testing at site rates, ASR, and time depth tasks will be completed by the end of first summer semester. **Part II** involving cognate detection will take up about 1 semester where development and testing of various classifiers on the simulated data. The cognate detection step will be taken up in the first semester itself once the synthetic data is ready. We will complete the experiments on ancestral word prediction (**Part III**) in the fifth semester and submit papers in the final semester.

5 Project Team

The project team consists of the PI and a graduate student with a Masters degree in computational linguistics. The computational linguistics program at UNT is inter-disciplinary focusing on courses in both linguistics and computer science departments. The linguistics department has NSF funded research collaboration with the Information Science department where both departments are affiliated with the College of Information. An hourly graduate student assistant will work on expansion of the word lists from 100 items to 500 items for 40 Dravidian languages which will be used for real world testing in parts II and III. The PI's doctoral dissertation and subsequent post-doctoral work (Tübingen) has been in topics related to language evolution: Bayesian phylogenetic inference [Smith and Rama, Forthcoming] and neural networks [Rama, 2016], resource creation for Indian languages [Rama and Vajjala, 2018, Borin et al., 2014]. The proposed project builds on the past strands of the PI's work: 1) understanding the limitations of Bayesian phylogenetic methods, 2) automated cognate detection, and 3) resource creation.

	Activity	S1	S2	S3	S4	S5	S6
Simulations and Tests	Model Testing	■	■				
	Multistate models			■	■		
	Partition models		■	■			
	Time Depth, ASR			■	■		
Cognate Detection	Classifiers Testing		■	■	■		
	Real world		■	■	■		
Word Prediction	Reflex pred.				■	■	
	Ancestral pred.					■	■

Table 4: Most of the activities span over the two years with the grayed semesters showing most activity.

6 Justification for Funding Requested

The PI has no resources for a PhD student apart from an hourly graduate assistant (20 hours a week) provided by the Department of Linguistics. The funds requested through this grant will allow the PI to hire a PhD student for two years to (1) facilitate the development of the Bayesian linguistic

phylogenetic package (`Pylogeny`) (2) an hourly research assistant to expand the multilingual word lists for the Dravidian family from 100 to 500 meanings for 40 Dravidian languages (listed in the Data Management Plan). The outcome of the project (if successful) will help the submission of a `CAREER` proposal in 2023 extending the proposed simulations to model morphological change, meaning change, and syntactic change. The CAREER proposal will consist of scaling up the phylogenetic algorithms for large language families such as Austronesian consisting of more than 1000 languages.

7 Intellectual Merit

The project is the first to develop linguistically motivated simulations for testing the limits of Bayesian phylogenetic methods and deep learning methods to a range of historical linguistics problems. The project addresses important questions about the accuracy and confidence of the findings from the Bayesian linguistic phylogenetics literature using simulations. Results from the tests of the synthetic data will guide the choice of Bayesian linguistic phylogenetic models in future projects. The project tests if synthetic data can be used to train machine learning models for cognate detection and word prediction tasks which, if successful, will allow the application of such machine learning models to severely under-studied languages of the world.

8 Broader Impacts

Training impact In the proposed project, a PhD student will be trained in both machine learning and historical linguistics. The University of North Texas (UNT) is a Hispanic-serving institution and a Minority-serving institution and the project is committed to recruit a minority student. The Texas Academy of Mathematics and Sciences (TAMS) is part of UNT featuring advanced high school students who attend courses along the undergraduates and graduate with transferable undergraduate credits (two years). The project will involve and train an undergraduate student from TAMS. The PI will develop teaching modules based on the outcomes of the project in the computational linguistics curriculum at UNT.

Tools and data Both the synthetic and real world data will be generated in cross-linguistic data format [Forkel et al., 2018] that is ready to analyze and visible to the community. We will develop an easy to use tool, integrated into the LingPy project, for historical linguistic researchers without computational linguistics or computer science expertise to analyze lexical data (word lists) through Bayesian phylogenetic inference and infer cognate judgments. The tool would be tested on word lists of less-resourced languages spoken in India studied by my linguistic colleagues (Dr. Shobhana Chelliah and Dr. Sadaf Munshi) at UNT. A collaborative paper by the PI [Smith and Rama, Forthcoming] with Dr. Alex Smith, a linguist colleague working on 87 Bornean island languages, showed that using automated cognate judgments helped the linguist colleague at expediting the cognate judgment procedure within a week. The Dravidian family data collected during the project period will be released to the research community for further research.

9 Results from Prior NSF Support

None. The PI has started the tenure-track position in 2019.

References

Antonios Anastasopoulos and Graham Neubig. Pushing the limits of low-resource morphological inflection. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 984–996, 2019.

Raimo Anttila. *An Introduction to Historical and Comparative Linguistics*. Macmillan, New York, 1972. ISBN 90-272-3556-2.

Quentin Atkinson, Geoff Nicholls, David Welch, and Russell Gray. From words to dates: water into wine, mathemagic or phylogenetic inference? *Transactions of the Philological Society*, 103 (2):193–219, 2005.

Quentin D Atkinson and Russell D Gray. Curious parallels and curious connections—phylogenetic thinking in biology and historical linguistics. *Systematic Biology*, 54(4):513–526, 2005.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL `http://arxiv.org/abs/1409.0473`.

François Barbançon, Steven N Evans, Luay Nakhleh, Don Ringe, and Tandy Warnow. An experimental study comparing linguistic phylogenetic reconstruction methods. *Diachronica*, 30(2): 143–170, 2013.

Toms Bergmanis, Katharina Kann, Hinrich Schütze, and Sharon Goldwater. Training data augmentation for low-resource morphological inflection. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 31–39, 2017.

Lars Borin, Anju Saxena, Taraka Rama, and Bernard Comrie. Linguistic landscaping of South Asia using digital language resources: Genetic vs. areal linguistics. In *Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3137–3144, 2014.

Alexandre Bouchard-Côté, David Hall, Thomas L Griffiths, and Dan Klein. Automated reconstruction of ancient languages using probabilistic models of sound change. *Proceedings of the National Academy of Sciences*, 110(11):4224–4229, 2013.

Remco Bouckaert, Philippe Lemey, Michael Dunn, Simon J Greenhill, Alexander V Alekseyenko, Alexei J Drummond, Russell D Gray, Marc A Suchard, and Quentin D Atkinson. Mapping the origins and expansion of the Indo-European language family. *Science*, 337(6097):957–960, 2012.

Remco Bouckaert, Joseph Heled, Denise Kühnert, Tim Vaughan, Chieh-Hsi Wu, Dong Xie, Marc A Suchard, Andrew Rambaut, and Alexei J Drummond. BEAST 2: A software platform for Bayesian evolutionary analysis. *PLoS computational biology*, 10(4):e1003537, 2014.

Claire Bowern and Quentin Atkinson. Computational phylogenetics and the internal structure of Pama-Nyungan. *Language*, 88(4):817–845, 2012.

Cecil H Brown, Eric W Holman, Søren Wichmann, and Viveka Velupillai. Automated classification of the world's languages: A description of the method and preliminary results. *Language Typology and Universals*, 61(4):285–308, 2008.

Cecil H Brown, Eric W Holman, and Søren Wichmann. Sound correspondences in the world's languages. *Language*, pages 4–29, 2013.

Thomas Burrow and Murray B. Emeneau. *A Dravidian Etymological Dictionary*. Clarendon Press, Oxford, 1984. URL `https://dsalsrv04.uchicago.edu/dictionaries/burrow/frontmatter.html`.

Chundra Cathcart and Taraka Rama. Disentangling dialects: a neural approach to indo-aryan historical phonology and subgrouping. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 620–630, 2020.

Will Chang, Chundra Cathcart, David Hall, and Andrew Garrett. Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language*, 91(1):194–244, 2015.

Chris Christiansen, Thomas Mailund, Christian NS Pedersen, Martin Randers, and Martin Stig Stissing. Fast calculation of the quartet distance between trees of arbitrary degrees. *Algorithms for Molecular Biology*, 1(1), 2006.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D. McCarthy, Katharina Kann, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, David Yarowsky, Jason Eisner, and Mans Hulden. The CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 1–27, Brussels, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/K18-3001. URL `https://aclanthology.org/K18-3001`.

Peter Dekker and Willem Zuidema. Word prediction in computational historical linguistics. *Journal of Language Modelling*, 8(2):295–336, 2020.

Johannes Dellert, Thora Daneyko, Alla Münch, Alina Ladygina, Armin Buch, Natalie Clarius, Ilja Grigorjew, Mohamed Balabel, Hizniye Isabella Boga, and Zalina Baysarova. NorthEuraLex: A wide-coverage lexical database of Northern Eurasia. *Language resources and evaluation*, 54(1): 273–301, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Mark Donohue, Søren Wichmann, and Mihai Albu. Typology, areality, and diffusion. *Oceanic Linguistics*, pages 223–232, 2008.

Matthew S Dryer and Martin Haspelmath. The World Atlas of Language Structures Online. 2013.

Michael Dunn, Angela Terrill, Ger Reesink, Robert A Foley, and Stephen C Levinson. Structural phylogenetics and the reconstruction of ancient language history. *Science*, 309(5743):2072–2075, 2005.

Joseph Felsenstein. Phylogenies from restriction sites: A maximum-likelihood approach. *Evolution*, 46(1):159–173, 1992.

Robert Forkel, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping, and Russell D. Gray. Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data*, 5(180205):1–10, 2018.

Clémentine Fourrier, Rachel Bawden, and Benoît Sagot. Can Cognate Prediction Be Modelled as a Low-Resource Machine Translation Task? In *ACL-IJCNLP 2021-Findings of the Association for Computational Linguistics*, 2021.

Anthony Fox. *Linguistic reconstruction: An introduction to theory and method*. Oxford University Press, 1995.

Russell D Gray and Quentin D Atkinson. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, 426(6965):435–439, 2003.

Russell D Gray, Alexei J Drummond, and Simon J Greenhill. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science*, 323(5913):479–483, 2009.

Peter J Green. Trans-dimensional Markov chain Monte Carlo. *Oxford Statistical Science Series*, pages 179–198, 2003.

Simon J Greenhill, Chieh-Hsi Wu, Xia Hua, Michael Dunn, Stephen C Levinson, and Russell D Gray. Evolutionary dynamics of language systems. *Proceedings of the National Academy of Sciences*, 114(42):E8822–E8829, 2017.

Rebecca Grollemund, Simon Branford, Koen Bostoen, Andrew Meade, Chris Venditti, and Mark Pagel. Bantu expansion shows that habitat alters the route and pace of human dispersals. *Proceedings of the National Academy of Sciences*, 112(43):13296–13301, 2015.

Harald Hammarström, Martin Haspelmath, and Robert Forkel. *Glottolog. Version 4.1*. Max Planck Institute for the Science of Human History, Jena, 2019. URL `https://glottolog.org`.

Sheldon P Harrison. On the limits of the comparative method. In *The handbook of Historical Linguistics*, pages 213–243. Blackwell, MA, 2003.

Martin Haspelmath and Uri Tadmor, editors. *Loanwords in the world's languages. A comparative handbook*. de Gruyter, Berlin and New York, 2009.

Bradley Hauer and Grzegorz Kondrak. Clustering semantically equivalent words into cognate sets in multilingual lists. In *Proceedings of 5th international joint conference on natural language processing*, pages 865–873, 2011.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Sebastian Höhna, Michael J Landis, Tracy A Heath, Bastien Boussau, Nicolas Lartillot, Brian R Moore, John P Huelsenbeck, and Fredrik Ronquist. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Systematic biology*, 65 (4):726–736, 2016.

Eric W Holman and Søren Wichmann. New evidence from linguistic phylogenetics identifies limits to punctuational change. *Systematic biology*, 66(4):604–610, 2017.

D. J. Hruschka, S. Branford, E. D. Smith, J. Wilkins, A. Meade, M. Pagel, and T. Bhattacharya. Detecting regular sound changes in linguistics as events of concerted evolution. *Curr. Biol.*, 25 (1):1–9, 2015.

Gerhard Jäger. Global-scale phylogenetic linguistic inference from lexical resources. *Scientific data*, 5(1):1–16, 2018.

Gerhard Jäger. Computational historical linguistics. *Theoretical Linguistics*, 45(3-4):151–182, 2019.

Gerhard Jäger, Johann-Mattis List, and Pavel Sofroniev. Using support vector machines and state-of-the-art algorithms for phonetic alignment to identify cognates in multi-lingual wordlists. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1205–1216, 2017.

Gerhard Jäger and Johann-Mattis List. Factoring lexical and phonetic phylogenetic characters from word lists. In H. Baayen, G. Jäger, M. Köllner, J. Wahle, and A. Baayen-Oudshoorn, editors, *Proceedings of the 6th Conference on Quantitative Investigations in Theoretical Linguistics*, Tübingen, 2015.

Rhea Kapur and Phillip Rogers. Modeling language evolution and feature dynamics in a realistic geographic environment. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 788–798, 2020.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Vishnupriya Kolipakam, Fiona M Jordan, Michael Dunn, Simon J Greenhill, Remco Bouckaert, Russell D Gray, and Annemarie Verkerk. A Bayesian phylogenetic study of the Dravidian language family. *Royal Society Open Science*, 5(3):171504, 2018.

Shantanu Kumar, Ashwini Vaidya, and Sumeet Agarwal. Discovering cognates using lstm networks. In *Proceedings of the 4th Annual Conference of the Association for Cognitive Science, Hyderabad, India*, 2017.

Nicolas Lartillot and Hervé Philippe. A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular biology and evolution*, 21(6):1095–1109, 2004.

Paul O. Lewis. A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic Biology*, 50(6):913–925, 2001.

Johann-Mattis List. SCA. Phonetic alignment based on sound classes. In M. Slavkovik and D. Lassiter, editors, *New directions in logic, language, and computation*, pages 32–51. Springer, Berlin, Heidelberg, 2012.

Johann-Mattis List. Beyond cognacy: Historical relations between words and their implication for phylogenetic reconstruction. *Journal of Language Evolution*, 1(2):119–136, 2016. doi: 10.1093/jole/lzw006.

Johann-Mattis List, Cormac Anderson, Tiago Tresoldi, Christoph Rzymski, Simon Greenhill, and Robert Forkel. Cross-linguistic transcription systems. *Max Planck Institute for the Science of Human History, Jena*, 2019a.

Johann-Mattis List, Simon Greenhill, Tiago Tresoldi, and Robert Forkel. *LingPy. A Python library for quantitative tasks in historical linguistics*. Max Planck Institute for the Science of Human History, Jena, 2019b. URL `http://lingpy.org`.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

Jayden L Macklin-Cordes, Claire Bowern, and Erich R Round. Phylogenetic signal in phonotactics. *Diachronica*, 38(2):210–258, 2021.

Ian Maddieson. Consonant inventories. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013. URL `https://wals.info/chapter/1`.

Carlo Meloni, Shauli Ravfogel, and Yoav Goldberg. Ab Antiquo: Neural Proto-language Reconstruction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4460–4473, 2021.

Steven Moran, Daniel McCloy, and Richard Wright. Phoible online. 2014.

David R Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori Levin. Panphon: A resource for mapping IPA segments to articulatory feature vectors. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3475–3484, 2016.

Robert L Oswalt. The detection of remote linguistic relationships. *Computer Studies in the Humanities and Verbal Behavior*, 3(3):117–129, 1970.

Mark Pagel and Andrew Meade. Estimating rates of lexical replacement on phylogenetic trees of languages. In P. Forster and C. Renfrew, editors, *Phylogenetic methods and the prehistory of languages*, pages 173–182. McDonald institute Monographs, 2006.

Taraka Rama. Siamese convolutional networks for cognate identification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1018–1027, 2016.

Taraka Rama. Similarity dependent Chinese restaurant process for cognate identification in multilingual wordlists. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 271–281, 2018a.

Taraka Rama. Three tree priors and five datasets: A study of Indo-European phylogenetics. *Language Dynamics and Change*, 8(2):182–218, 2018b.

Taraka Rama and Johann-Mattis List. An automated framework for fast cognate detection and Bayesian phylogenetic inference in computational historical linguistics. In *57th Annual Meeting of the Association for Computational Linguistics*, page 6225–6235. Association for Computational Linguistics, 2019.

Taraka Rama and Johann-Mattis List. Are sounds sound for the reconstruction of language trees? comparing lexical cognates and sound correspondences in bayesian phylogenetic inference, Submitted.

Taraka Rama and Sowmya Vajjala. A dependency treebank for Telugu. In *Proceedings of the 16th international workshop on treebanks and linguistics theories*, pages 119–128, 2018.

Taraka Rama and Søren Wichmann. Towards identifying the optimal datasize for lexically-based Bayesian inference of linguistic phylogenies. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1578–1590, 2018.

Taraka Rama and Søren Wichmann. A test of Generalized Bayesian dating: A new linguistic dating method. *PloS one*, 15(8):e0236522, 2020.

Taraka Rama, Johann-Mattis List, Johannes Wahle, and Gerhard Jäger. Are automatic methods for cognate detection good enough for phylogenetic reconstruction in historical linguistics? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 393–400, New Orleans, Louisiana, June 2018a. Association for Computational Linguistics. doi: 10.18653/v1/N18-2063. URL `https://aclanthology.org/N18-2063`.

Taraka Rama, Johann-Mattis List, Johannes Wahle, and Gerhard Jäger. Are Automatic Methods for Cognate Detection Good Enough for Phylogenetic Reconstruction in Historical Linguistics? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 393–400, 2018b.

Donald Ringe, Tandy Warnow, and Ann Taylor. Indo-European and computational cladistics. *Transactions of the Philological Society*, 100(1):59–129, 2002.

Donald A Ringe. *From Proto-Indo-European to Proto-Germanic*, volume 1. Oxford University Press, 2017.

Fredrik Ronquist, Maxim Teslenko, Paul van der Mark, Daniel L Ayres, Aaron Darling, Sebastian Höhna, Bret Larget, Liang Liu, Marc A Suchard, and John P Huelsenbeck. MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic biology*, 61(3):539–542, 2012.

Fredrik Ronquist, Nicolas Lartillot, and Matthew J Phillips. Closing the gap between rocks and clocks using total-evidence dating. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1699):20150136, 2016.

Alex Smith and Taraka Rama. Environmental factors affect the evolution of linguistic subgroups in Borneo. *Diachronica*, Forthcoming.

Richard Socher, Eric Huang, Jeffrey Pennin, Christopher D Manning, and Andrew Ng. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. *Advances in neural information processing systems*, 24, 2011.

Morris Swadesh. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics*, 21(2):121–137, 1955. ISSN 0020-7071.

Asif Tamuri and Nick Goldman. Avoiding ascertainment bias in the maximum likelihood inference of phylogenies based on truncated data. *BioRxiv*, page 186478, 2017.

Robert Lawrence Trask. *The dictionary of historical and comparative linguistics*. Psychology Press, 2000.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

Dingquan Wang and Jason Eisner. The galactic dependencies treebanks: Getting more data by synthesizing new languages. *Transactions of the Association for Computational Linguistics*, 4: 491–505, 2016.

Søren Wichmann. Modeling language family expansions. *Diachronica*, 34(1):79–101, 2017.

Søren Wichmann and Eric W Holman. Assessing temporal stability for linguistic typological features. *München: LINCOM Europa*, 2009.

Søren Wichmann and Taraka Rama. Testing methods of linguistic homeland detection using synthetic data. *Philosophical Transactions of the Royal Society B*, 376(1824):20200202, 2021.

Søren Wichmann, Eric W Holman, Dik Bakker, and Cecil H Brown. Evaluating linguistic distance measures. *Physica A: Statistical Mechanics and its Applications*, 389(17):3632–3639, 2010.

Martijn Wieling, Jelena Prokić, and John Nerbonne. Evaluating the pairwise string alignment of pronunciations. In *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education (LaTeCH–SHELT&R 2009)*, pages 26–34, 2009.

April M Wright. A systematist's guide to estimating bayesian phylogenies from morphological data. *Insect Systematics and Diversity*, 3(3):2, 2019.

April M Wright, Graeme T Lloyd, and David M Hillis. Modeling character change heterogeneity in phylogenetic analyses of morphology through the use of priors. *Systematic Biology*, 65(4): 602–611, 2016.

Shijie Wu and Ryan Cotterell. Exact Hard Monotonic Attention for Character-Level Transduction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1530–1537, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1148. URL `https://aclanthology.org/P19-1148`.

Ziheng Yang. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *Journal of Molecular evolution*, 39(3):306–314, 1994.

Ziheng Yang. *Molecular Evolution: A Statistical Approach*. Oxford University Press, 2014.

Chi Zhang, Tanja Stadler, Seraina Klopfstein, Tracy A Heath, and Fredrik Ronquist. Total-evidence dating under the fossilized birth–death process. *Systematic biology*, 65(2):228–249, 2016.