Are Automatic Methods for Cognate Detection Good Enough for Phylogenetic Reconstruction in Historical Linguistics?

Taraka Rama[♠] Johann-Mattis List[◇] Johannes Wahle[♡] Gerhard Jäger[♡]

[♠]Department of Informatics, University of Oslo, Norway
[◇]Department of Linguistic and Cultural Evolution, MPI-SHH, Jena, Germany
[♡]Department of Linguistics, University of Tübingen, Germany
tarakark@ifi.uio.no, list@shh.mpg.de, {johannes.wahle,gerhard.jaeger}@uni-tuebingen.de

Abstract

We evaluate the performance of state-of-the-art algorithms for automatic cognate detection by comparing how useful automatically inferred cognates are for the task of phylogenetic inference compared to classical manually annotated cognate sets. Our findings suggest that phylogenies inferred from automated cognate sets come close to phylogenies inferred from expert-annotated ones, although on average, the latter are still superior. We conclude that future work on phylogenetic reconstruction can profit much from automatic cognate detection. Especially where scholars are merely interested in exploring the bigger picture of a language family's phylogeny, algorithms for automatic cognate detection are a useful complement for current research on language phylogenies.

1 Introduction

The task of *cognate detection*, i.e., the search for genetically related words in different languages, has traditionally been regarded as a task that is barely automatable. During the last decades, however, automatic cognate detection approaches since Covington (1996) have been constantly improved following the work of Kondrak (2002), both regarding the quality of the inferences (List et al., 2017b; Jäger et al., 2017), and the sophistication of the methods (Hauer and Kondrak, 2011; Rama, 2016; Jäger et al., 2017), which have been expanded to account for the detection of partial cognates (List et al., 2016b), language specific sound-transition weights (List, 2012) or the search of cognates in whole dictionaries (St Arnaud et al., 2017).

Despite the progress, none of the automated cognate detection methods have been used for the purpose of inferring phylogenetic trees using modern Bayesian phylogenetic methods (Yang and Rannala, 1997) from computational biology. Phylogenetic trees are hypotheses of how sets of related languages evolved in time. They can in turn be used for testing additional hypotheses of language evolution, such as the age of language families (Gray and Atkinson, 2003; Chang et al., 2015), their spread (Bouckaert et al., 2012; Gray et al., 2009), the rates of lexical change (Greenhill et al., 2017), or as a proxy for tasks like cognate detection and linguistic reconstruction (Bouchard-Côté et al., 2013). By plotting shared traits on a tree and testing how they could have evolved, trees can even be used to test hypotheses independent from language evolution, such as the universality of typological statements (Dunn et al., 2011), or the ancestry of cultural traits (Jordan et al., 2009).

In the majority of these approaches, scholars infer phylogenetic trees with help of *expert-annotated cognate sets* which serve as input to the phylogenetic software which usually follows a Bayesian likelihood framework. Unfortunately, expert cognate judgments are only available for a small number of language families which look back on a long tradition of classical comparative linguistic research (Campbell and Poser, 2008). Despite the claims that automatic cognate detection is useful for linguists working on less well studied language families, none of the papers actually tested, if automated cognates can be used instead as well for the important downstream task of Bayesian phylogenetic inference. So far, scholars have only tested distance-based approaches to phylogenetic reconstruction (Wichmann et al., 2010; Rama and Borin, 2015; Jäger, 2013), which employ aggregated linguistic distances computed from string similarity algorithms to infer phylogenetic trees.

In order to test whether automatic cognate detection is useful for phylogenetic inference, we collected multilingual wordlists for five different language families (230 languages, cf. section 2.1) and then applied different cognate detection methods (cf. section 2.2) to infer cognate sets. We then applied the Bayesian phylogenetic inference procedure (cf. section 3) to the automated and the expert-annotated cognate sets in order to infer phylogenetic trees. These trees were then evaluated against the *family gold standard trees*, based on external linguistic knowledge (Hammarström et al., 2017), using the *Generalized Quartet Distance* (cf. section 4.1). The results are provided in table 3 and the paper is concluded in section 5.

To the best of our knowledge, this is the first study in which the performance of several automatic cognate detection methods on the downstream task of phylogenetic inference is compared. While we find that on average the trees inferred from the expert-annotated cognate sets come closer to the gold standard trees, the trees inferred from automated cognate sets come surprisingly close to the trees inferred from the expert-annotated ones.

Dataset	**Mngs.**	**Lngs.**	**AMC**
Austronesian	210	45	0.79
Austro-Asiatic	200	58	0.90
Indo-European	208	42	0.95
Pama-Nyungan	183	67	0.89
Sino-Tibetan	110	64	0.91

Table 1: Datasets used in our study. The second, third, and fourth columns show the number of number of meanings, languages and average mutual coverage for each language family respectively.

2 Materials and Methods

2.1 Datasets

Our wordlists were extracted from publicly available datasets from five different language families: Austronesian (Greenhill et al., 2008), Austro-Asiatic (Sidwell, 2015), Indo-European (Dunn, 2012), Pama-Nyungan (Bowern and Atkinson, 2012), and Sino-Tibetan (Peiros, 2004). In order to make sure that the datasets were amenable for automatic cognate detection, we had to make sure that the transcriptions employed are readily recognized, and that the data is sufficient for those methods which rely on the identification of regular sound correspondences. The problem of transcriptions was solved by applying intensive semi-automatic cleaning. In order to guarantee an optimal data size, we selected a subset of languages from each dataset, which would guarantee a high *average mutual coverage* (AMC). AMC is calculated as the average proportion of words shared by all language pairs in a given dataset. All analyses were carried out with version 2.6.2 of LingPy (List et al., 2017a). Table 1 gives an overview on the number of languages, concepts, and the AMC score for all datasets.[1]

2.2 Automatic Cognate Detection

The basic workflow for automatic cognate detection methods applied to multilingual wordlists has been extensively described in the literature (Hauer and Kondrak, 2011; List, 2014). The workflow can be divided into two major steps: (a) word similarity calculation, and (b) cognate set partitioning. In the first step, similarity or distance scores for all word pairs in the same concept slot in the data are computed. In the second step, these scores are used to partition the words into sets of presumably related words. Since the second step is a mere clustering task for which many solutions exist, the most crucial differences among algorithms can be noted for step (a).

For our analysis, we tested six different methods for cognate detection: The Consonant-Class-Matching (CCM) Method (Turchin et al., 2010), the Normalized Edit Distance (NED) approach (Levenshtein, 1965), the Sound-Class-Based Alignment (SCA) method (List, 2014), the LexStat-Infomap method (List et al., 2017b), the SVM method (Jäger et al., 2017), and the Online PMI approach (Rama et al., 2017).

The **CCM** approach first reduces the size of the alphabets in the phonetic transcriptions by mapping consonants to *consonant classes* and discarding vowels. Assuming that different sounds which share the same sound class are likely to go back to the same ancestral sound, words which share the

[1] In order to allow for an easy re-use of our datasets, we linked all language varieties to Glottolog (Hammarström et al., 2017) and all concepts to Concepticon (List et al., 2016a). In addition to the tabular data formats required to run the analyses with our software tools, we also provide the data in form of the format specifications suggested by the Cross-Linguistic Data Formats initiative (Forkel et al., 2017). Data and source code are provided along with the supplementary material accompanying this paper.

first two consonant classes are judged to be cognate, while words which differ regarding their first two classes are regarded as non-cognate.

The **NED** approach first computes the *normalized edit distance* (Nerbonne and Heeringa, 1997) for all word pairs in given semantic slot and then clusters the words into cognate sets using a flat version of the UPGMA algorithm (Sokal and Michener, 1958) and a user-defined threshold of maximal distance among the words. We follow List et al. (2017b) in setting this threshold to 0.75.

The **SCA** approach is very similar to NED, but the pairwise distances are computed with help of the Sound-Class-Based Phonetic Alignment algorithm (List, 2014) which employs an extended sound-class model and a linguistically informed scoring function. Following List et al. (2017b), we set the threshold for this approach to 0.45.

The **LexStat-Infomap** method builds on the SCA method by employing the same sound-class model, but individual scoring functions are inferred from the data for each language pair by applying a permutation method and computing the *log-odds scores* (Eddy, 2004) from the expected and the attested distribution of sound matches (List, 2014). While SCA and NED employ flat UGPMA clustering for step 2 of the workflow, LexStat-Infomap further uses the Infomap community detection algorithm (Rosvall and Bergstrom, 2008) to partition the words into cognate set. Following List et al. (2017b), we set the threshold for LexStat-Infomap to 0.55.

The **OnlinePMI** approach (Rama et al., 2017) estimates the sound-pair PMI matrix using the online procedure described in Liang and Klein (2009). The approach starts with an empty PMI matrix and a list of synonymous word pairs from all the language pairs. The approach proceeds by calculating the PMI matrix from alignments calculated for each minibatch of word pairs using the current PMI matrix. Then the calculated PMI matrix for the latest minibatch is combined with the current PMI matrix. This procedure is repeated for a fixed number of iterations. We employ the final PMI matrix to calculate pairwise word similarity matrix for each meaning. In an additional step, the similarity score was transformed into a distance score using the sigmoid transformation: $1.0-(1+\exp(-x))^{-1}$ The word distance matrix is then supplied as an input to the Label Propagation algorithm (Raghavan et al., 2007) to infer cognate clusters. We set the threshold for the algorithm to be 0.5.

For the **SVM** approach (Jäger et al., 2017) a linear SVM classifier was trained with PMI similarity (Jäger, 2013), LexStat distance, mean word length, distance between the languages as features on cognate and non-cognate pairs extracted from word lists from Wichmann and Holman (2013) and List (2014). The details of the training dataset are given in table 1 in Jäger et al. (2017). We used the same training settings as reported in the paper to train our SVM model. The trained SVM model is then employed to compute the probability that a word pair is cognate or not. The word pair probability matrix is then given as input to InfoMap algorithm for inferring word clusters. The threshold for InfoMap algorithm is set to 0.57 after cross-validation experiments on the training data.

We evaluate the quality of the inferred cognate sets using the above described methods using B-cubed F-score (Amigó et al., 2009) which is widely used in evaluating the quality of automatically inferred cognate clusters (Hauer and Kondrak, 2011). We present the cognate evaluation results in table 2. The SVM system is the best in the case of Austro-Asiatic and Pama-Nyungan whereas LexStat algorithm performs the best in the case of rest of the datasets. This is surprising since LexStat scores are used as features for SVM and we expect the SVM system to perform better than LexStat in all the language families. On the other hand, both OnlinePMI and SCA systems perform better than the algorithmically simpler systems such as CCM and NED. Given these F-scores, we hypothesize that the cognate sets output from the best cognate identification systems would also yield the high quality phylogenetic trees. However, we find the opposite in our phylogenetic experiments.

3 Bayesian Phylogenetic Inference

The objective of Bayesian phylogenetic inference is based on the Bayes rule in 1.

$$f(\tau, v, \theta|X) = \frac{f(X|\tau, v, \theta) f(\tau, v, \theta)}{f(X)} \qquad (1)$$

where X is the data matrix, τ is the topology of the tree, v is the vector of branch lengths, and θ is the substitution model parameters. The data matrix X is a binary matrix of dimensions $N \times C$

Method	Austro-Asiatic	Austronesian	Indo-European	Pama-Nyungan	Sino-Tibetan
CCM	0.71	0.7	0.75	0.74	0.48
NED	0.73	0.77	0.69	0.53	0.49
SCA	0.76	0.78	0.81	0.71	0.56
LexStat	0.76	0.84	0.83	0.84	0.6
OnlinePMI	0.76	0.81	0.82	0.72	0.56
SVM	0.82	0.81	0.79	0.86	0.5

Table 2: B-cubed F-scores for different cognate detection methods across the language families.

where N is the number of languages and C is the number of cognate clusters in a language family. The posterior distribution $f(\tau, v, \theta|X)$ is difficult to calculate analytically since one has to sum over all the possible topologies $(\frac{(2N-3)!}{2^{N-2}(N-2)!})$ to compute the marginal in the denominator. However, posterior probability of all the parameters of interest (here, $\Psi = \{\tau, v, \theta\}$) can be computed from samples drawn using a Markov chain Monte Carlo (MCMC) method. Typically, Metropolis-Hastings (MH) algorithm is the MCMC algorithm used to sample phylogenies from the posterior distribution (Huelsenbeck et al., 2001).

The MH algorithm constructs a Markov chain of the parameters' states by proposing change to a single parameter or a block of parameters in Ψ. The current state Ψ in the Markov chain has a parameter θ and a new value θ^* is proposed from a distribution $q(\theta^*|\theta)$, then θ^* is accepted with a probability

$$r = \frac{f(X|\tau, v, \theta^*)}{f(X|\tau, v, \theta)} \frac{f(\theta^*)}{f(\theta)} \frac{q(\theta|\theta^*)}{q(\theta^*|\theta)} \quad (2)$$

The likelihood of the data $f(X|\Psi)$ is computed using the Felsenstein's pruning algorithm (Felsenstein, 1981) also known as sum-product algorithm (Jordan et al., 2004). We assume that τ, θ, v are independent of each other.

4 Experiments

In this section, we report the experimental settings, the evaluation measure, and the results of our experiments.

All our Bayesian analyses use binary datasets with states 0 and 1. We employ the Generalized Time Reversible Model (Yang, 2014, chapter 1) for computing the transition probabilities between individual states. The rate variation across sites is modeled using a four category discrete Γ distribution (Yang, 1994). We follow Lewis (2001) and Felsenstein (1992) in correcting the likelihood calculation for ascertainment bias resulting from unobserved 0 patterns. We used a uniform tree prior (Ronquist et al., 2012) in all our analyses which constructs a rooted tree and draws internal node heights from uniform distribution. In our analysis, we assumes a Independent Gamma Rates relaxed clock model (Lepage et al., 2007) where the rate for a branch j of length b_j in the tree is drawn from a Gamma distribution with mean 1 and variance σ^2_{IG}/b_j where σ^2_{IG} is a parameter sampled in the MCMC analysis.

We infer τ, v, θ from two independent random starting points and sample every 1000th state in the chain until the phylogenies from the two independent runs do not differ beyond 0.01. For each dataset, we ran the chains for 15 million generations and threw away the initial 50% of the chain's states as part of burnin. After that we computed the generalized quartet distance from each of the posterior trees to the gold standard tree described in subsection 4.1. All our experiments are performed using MrBayes 3.2.6 (Zhang et al., 2015).

4.1 GQD

Pompei et al. (2011) introduced Generalized Quartet Distance (GQD) as an extension to Quartet Distance (QD) in order to compare binary trees with a polytomous tree, since gold standard trees can have non-binary internal nodes. It was widely used for comparing inferred language phylogenies with gold standard phylogenies (Greenhill et al., 2010; Wichmann et al., 2011; Jäger, 2013).

QD measures the distance between two trees in terms of the number of different quartets (Estabrook et al., 1985). A quartet is defined as a set of four leaves selected from a set of leaves without replacement. A tree with n leaves has $\binom{n}{4}$ quartets in total. A quartet defined on four leaves a, b, c, d can have four different topologies: $ab|cd$, $ac|bd$, $ad|bc$, and $ab \times cd$. The first three topologies have an internal edge separating two pairs of leaves. Such quartets are called as *butterflies*. The fourth quartet has no internal edge and as such is known as star quartet. Given a

Method	Austro-Asiatic	Austronesian	Indo-European	Pama-Nyungan	Sino-Tibetan
Expert cognate sets	**0.0081 ± 0.001**	0.1056 ± 0.0118	**0.0249 ± 0.0079**	**0.1384 ± 0.0225**	**0.0561 ± 0.0123**
CCM	0.0243 ± 0.018	0.0854 ± 0.0176	0.0369 ± 0.0148	0.1617 ± 0.0162	0.1424 ± 0.027
NED	0.0265 ± 0.007	0.0458 ± 0.0152	0.046 ± 0.0132	0.196 ± 0.0166	0.1614 ± 0.0282
SCA	0.0152 ± 0.0035	0.0514 ± 0.013	0.0256 ± 0.009	0.166 ± 0.0153	0.0704 ± 0.0206
LexStat	0.0267 ± 0.0085	0.0848 ± 0.0226	0.0314 ± 0.0091	0.1507 ± 0.0143	0.0786 ± 0.0209
OnlinePMI	0.0158 ± 0.0048	0.1056 ± 0.0198	0.0457 ± 0.0135	0.1717 ± 0.0185	0.1184 ± 0.031
SVM	0.0146 ± 0.0039	0.0989 ± 0.0224	0.0452 ± 0.011	0.1827 ± 0.0237	0.1199 ± 0.0269

Table 3: The mean and standard deviation for each method and family is computed from 7500 posterior trees. The automatic methods which comes closest to the gold standard phylogeny is shaded in gray, and where the expert cognate sets perform best, this is indicated with a **bold** font.

tree τ with n leaves, the quartets can be partitioned into sets of butterflies, $B(\tau)$, and sets of stars, $S(\tau)$. Then, the QD between τ and τ_g is defined as $1 - \frac{|S(\tau)\cap S(\tau_g)|+|B(\tau)\cap B(\tau_g)|}{\binom{n}{4}}$. The QD formulation counts the butterflies in an inferred tree τ as errors. The tree τ should not be penalized if an internal node in the gold standard tree τ_g is m-ary. To this end, Pompei et al. (2011) defined a new measure known as GQD to discount the presence of star quartets in τ_g. GQD is defined as $DB(\tau,\tau_g)/B(\tau_g)$ where $DB(.)$ is the number of butterflies between τ, τ_g.

We extracted gold standard trees from Glottolog (Hammarström et al., 2017) for the purpose of evaluating the inferred posterior trees from each automated cognate identification system. We note that the Bayesian inference procedure produces rooted trees with branch lengths whereas the gold standard trees do not have any branch lengths. Although there are other linguistic phylogenetic inference algorithms such as those of Ringe et al. (2002) we do not test the algorithms due to the non-availability and scalability of the software to datasets with more than twenty languages.

4.2 Results

The results of our experiments are given in table 3. A average lower GQD score implies that the inferred trees are closer to the gold standard phylogeny than a higher average GQD score. Except for Austronesian, Bayesian inference based on expert cognate sets yields trees that are very close to the gold standard tree. Surprisingly, algorithmically simple systems such as NED and CCM show better performance than the machine-learned SVM model except from Sino-Tibetan. SCA is a subsystem of LexStat but emerges as the winner in two language families (Indo-European and Sino-Tibetan). Given that SCA is outperformed by SVM and LexStat in automatic cognate detection, this is very surprising, and further research is needed to find out, why the simpler models perform well on phylogenetic reconstruction. Although our results indicate that expert-coded cognate sets are generally more suitable for phylogenetic reconstruction, we can also see that the difference to trees inferred from automated cognate sets is not very large.

5 Conclusion

In this paper, we carried out a preliminary evaluation of the usefulness of automated cognate detection methods for phylogenetic inference. Although the cognate sets predicted by automated cognate detection methods yield phylogenetic trees that come close to expert trees, there is still room for improvement, and future research is needed to further enhance automatic cognate detection methods. However, as our experiments show, expert-annotated cognate sets are also not free from errors, and it seems likewise useful to investigate, how the consistency of cognate coding by experts could be further improved.

As future work, we intend to create a cognate identification system that combines the output of different algorithms in a more systematic way. We intend to infer cognate sets from the combined system and use them to infer phylogenies and evaluate the inferred phylogenies against the gold standard trees.

Acknowledgments

This research was supported by the ERC Advanced Grant 324246 EVOLAEMP (GJ, JW), the DFG-KFG 2237 Words, Bones, Genes, Tools (GJ), the ERC Starting Grant 715618 CALC (JML), and the BIGMED project (TR). We thank our anonymous reviewers for helpful comments, Mei-Shin Wu for helping with Sino-Tibetan data, Claire Bowern and Tiago Tresoldi for helping with Pama-Nyungan data, Paul Sidwell for helping with Austro-Asiatic data, as well as the audience at the CESC 2017 conference (MPI-SHH, Jena) for their helpful comments on an earlier version of the paper.

References

Enrique Amigó, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. 2009. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information retrieval*, 12(4):461–486.

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

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. 2012. Mapping the origins and expansion of the Indo-European language family. *Science*, 337(6097):957–960.

Claire Bowern and Quentin D. Atkinson. 2012. Computational phylogenetics of the internal structure of Pama-Nguyan. *Language*, 88:817–845.

Lyle Campbell and William J. Poser. 2008. *Language classification: History and Method*. Cambridge University Press.

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

Michael A. Covington. 1996. An algorithm to align words for historical comparison. *Computational Linguistics*, 22(4):481–496.

Michael Dunn. 2012. Indo-European lexical cognacy database (IELex).

Michael Dunn, Simon J. Greenhill, Stephen C. Levinson, and Russell D. Gray. 2011. Evolved structure of language shows lineage-specific trends in word-order universals. *Nature*, 473(7345):79–82.

Sean R. Eddy. 2004. Where did the BLOSUM62 alignment score matrix come from? *Nature Biotechnology*, 22(8):1035–1036.

George F Estabrook, FR McMorris, and Christopher A Meacham. 1985. Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units. *Systematic Biology*, 34(2):193–200.

Joseph Felsenstein. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376.

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

Robert Forkel, Johann-Mattis List, Michael Cysouw, and Simon J. Greenhill. 2017. *CLDF. Cross-Linguistic Data Formats. Version 1.0*. Max Planck Institute for the Science of Human History, Jena.

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

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

Simon J. Greenhill, Robert Blust, and Russell D. Gray. 2008. The Austronesian Basic Vocabulary Database: From bioinformatics to lexomics. *Evolutionary Bioinformatics*, 4:271–283.

Simon J. Greenhill, Alexei J. Drummond, and Russell D. Gray. 2010. How accurate and robust are the phylogenetic estimates of Austronesian language relationships? *PloS one*, 5(3):e9573.

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

Harald Hammarström, Robert Forkel, and Martin Haspelmath. 2017. *Glottolog*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Bradley Hauer and Grzegorz Kondrak. 2011. Clustering semantically equivalent words into cognate sets in multilin In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 865–873, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

John P Huelsenbeck, Fredrik Ronquist, Rasmus Nielsen, and Jonathan P Bollback. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. *science*, 294(5550):2310–2314.

Gerhard Jäger. 2013. Phylogenetic inference from word lists using weighted alignment with empirically determined weights. *Language Dynamics and Change*, 3(2):245–291.

Gerhard Jäger, Johann-Mattis List, and Pavel Sofroniev. 2017. 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. Long Papers*, pages 1204–1215, Valencia. Association for Computational Linguistics.

Fiona M. Jordan, Russell D. Gray, Simon J. Greenhill, and Ruth Mace. 2009. Matrilocal residence is ancestral in Austronesian societies. *Proceedings of the Royal Society B: Biological Sciences*, 276(1664):1957–1964.

Michael I Jordan et al. 2004. Graphical models. *Statistical Science*, 19(1):140–155.

Grzegorz Kondrak. 2002. *Algorithms for language reconstruction*. Ph.D. thesis, University of Toronto, Ontario, Canada.

Thomas Lepage, David Bryant, Hervé Philippe, and Nicolas Lartillot. 2007. A general comparison of relaxed molecular clock models. *Molecular biology and evolution*, 24(12):2669–2680.

VI Levenshtein. 1965. Binary codes capable of correcting spurious insertions and reversals. *Cybernetics and Control Theory*, 10:707–710.

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

Percy Liang and Dan Klein. 2009. Online em for unsupervised models. In *Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics*, pages 611–619. Association for Computational Linguistics.

Johann-Mattis List. 2012. LexStat: Automatic detection of cognates in multilingual wordlists. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pages 117–125, Avignon, France. Association for Computational Linguistics.

Johann-Mattis List. 2014. *Sequence comparison in historical linguistics*. Düsseldorf University Press, Düsseldorf.

Johann-Mattis List, Michael Cysouw, and Robert Forkel. 2016a. Concepticon. A resource for the linking of concept lists. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 2393–2400. European Language Resources Association (ELRA).

Johann-Mattis List, Simon Greenhill, and Robert Forkel. 2017a. *LingPy. A Python library for quantitative tasks in historical linguistics*. Max Planck Institute for the Science of Human History, Jena.

Johann-Mattis List, Simon J. Greenhill, and Russell D. Gray. 2017b. The potential of automatic word comparison for historical linguistics. *PLOS ONE*, 12(1):1–18.

Johann-Mattis List, Philippe Lopez, and Eric Bapteste. 2016b. Using sequence similarity networks to identify partial cognates in multilingual wordlists. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 599–605, Berlin, Germany. Association for Computational Linguistics.

J. Nerbonne and W. Heeringa. 1997. Measuring dialect distance phonetically. In *Proceedings of SIGPHON-97: 3rd Meeting of the ACL Special Interest Group in Computational Phonology*.

Ilia Peiros. 2004. *[Dataset on Sino-Tibetan languages encoded in STAR* Russian State University for the Humanities, Moscow.

Simone Pompei, Vittorio Loreto, and Francesca Tria. 2011. On the accuracy of language trees. *PloS one*, 6(6):e20109.

Usha Nandini Raghavan, Réka Albert, and Soundar Kumara. 2007. Near linear time algorithm to detect community structures in large-scale networks. *Physical review E*, 76(3):036106.

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

Taraka Rama and Lars Borin. 2015. Comparative evaluation of string similarity measures for automatic language classification. In Ján Mačutek and George K. Mikros, editors, *Sequences in Language and Text*, pages 203–231. Walter de Gruyter.

Taraka Rama, Johannes Wahle, Pavel Sofroniev, and Gerhard Jäger. 2017. Fast and unsupervised methods for multilingual cognate clustering. *arXiv preprint arXiv:1702.04938*.

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

Fredrik Ronquist, Seraina Klopfstein, Lars Vilhelmsen, Susanne Schulmeister, Debra L. Murray, and Alexandr P. Rasnitsyn. 2012. A total-evidence approach to dating with fossils, applied to the early radiation of the hymenoptera. *Systematic Biology*, 61(6):973–999.

Martin Rosvall and Carl T Bergstrom. 2008. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123.

Paul Sidwell. 2015. Austroasiatic dataset for phylogenetic analysis: 2015 version. *Mon-Khmer Studies (Notes, Reviews, Data-Papers)*, 44:lxviii–ccclvii.

Robert. R. Sokal and Charles. D. Michener. 1958. A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletin*, 28:1409–1438.

Adam St Arnaud, David Beck, and Grzegorz Kondrak. 2017. Identifying cognate sets across dictionaries of related languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2509–2518.

Peter Turchin, Ilja Peiros, and Murray Gell-Mann. 2010. Analyzing genetic connections between languages by matching consonant classes. *Journal of Language Relationship*, 3:117–126.

Søren Wichmann and Eric W Holman. 2013. Languages with longer words have more lexical change. In *Approaches to Measuring Linguistic Differences*, pages 249–281. Mouton de Gruyter.

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

Søren Wichmann, Eric W. Holman, Taraka Rama, and Robert S. Walker. 2011. Correlates of reticulation in linguistic phylogenies. *Language Dynamics and Change*, 1(2):205–240.

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

Ziheng Yang. 2014. *Molecular evolution: A statistical approach*. Oxford University Press, Oxford.

Ziheng Yang and Bruce Rannala. 1997. Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo method. *Molecular biology and evolution*, 14(7):717–724.

Chi Zhang, Tanja Stadler, Seraina Klopfstein, Tracy A Heath, and Fredrik Ronquist. 2015. Total-evidence dating under the fossilized birth–death process. *Systematic biology*, page syv080.

A Supplemental Material

The code and data used in this paper are uploaded as a zip file. In addition, they are available for download via Zenodo at `https://doi.org/10.5281/zenodo.1218060`.