A New Framework for Fast Automated Phonological Reconstruction Using Trimmed Alignments and Sound Correspondence Patterns

Johann-Mattis List
DLCE
MPI-EVA
Leipzig
mattis_list@eva.mpg.de

Robert Forkel
DLCE
MPI-EVA
Leipzig
robert_forkel@eva.mpg.de

Nathan W. Hill
Trinity Centre for Asian Studies
University of Dublin
Dublin
nathan.hill@tcd.ie

Abstract

Computational approaches in historical linguistics have been increasingly applied during the past decade and many new methods that implement parts of the traditional comparative method have been proposed. Despite these increased efforts, there are not many easy-to-use and fast approaches for the task of phonological reconstruction. Here we present a new framework that combines state-of-the-art techniques for automated sequence comparison with novel techniques for phonetic alignment analysis and sound correspondence pattern detection to allow for the supervised reconstruction of word forms in ancestral languages. We test the method on a new dataset covering six groups from three different language families. The results show that our method yields promising results while at the same time being not only fast but also easy to apply and expand.

1 Introduction

Phonological reconstruction is a technique by which words in ancestral languages, which may not even be reflected in any sources, are restored through the systematic comparison of descendant words (*cognates*) in descendant languages (Fox, 1995). Traditionally, scholars apply the technique manually, but along with the recent quantitative turn in historical linguistics, scholars have increasingly tried to automate the procedure. Recent automatic approaches for linguistic reconstruction, be they supervised or unsupervised, show two major problems. First, the underlying code is rarely made publicly available, which means that they cannot be further tested by applying them to new datasets. Second, the methods have so far only been tested on a small amount of data from a limited number of language families. Thus, Bouchard-Côté et al. (2013) report remarkable results on the reconstruction of Oceanic languages, but the source code has never been published, and the method was never tested on additional datasets. Meloni et al. (2021) report very promising results for the automated reconstruction of Latin from Romance languages, using a new test set derived from a dataset originally provided by Dinu and Ciobanu (2014), but they could only share part of the data, due to restrictions underlying the data by Dinu and Ciobanu (2014). Bodt and List (2022) experiment with the prediction of so far unelicited words in a small group of Sino-Tibetan languages, which they registered prior to verification (Bodt and List, 2019), but they do not test the suitability of their approach for the reconstruction of ancestral languages. Jäger (2019) presents a complete pipeline by which words are clustered into cognate sets and ancestral word forms are reconstructed, but the method is only tested on a very small dataset of Romance languages.

With increasing efforts to unify and standardize lexical datasets from different sources (Forkel et al., 2018), more and more datasets that could be used to test methods for automated linguistic reconstruction have become available. Additionally, thanks to the huge progress which techniques for automated sequence comparison have made in the past decades (Kondrak, 2000; Steiner et al., 2011; List, 2014), it is much easier today to combine existing methods into new frameworks that tackle individual tasks in computational historical linguistics.

In this study, we present a new framework for automated linguistic reconstruction which combines state-of-the-art methods for automated sequence comparison with fast machine-learning techniques and test it on a newly compiled test set that covers multiple language families.

2 Materials

The number of cross-linguistic datasets amenable for automated processing has been constantly increasing during the past years, as reflected specifically also in the development of standards for data representation that are increasingly used by

Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change, pages 89 - 96
May 26-27, 2022 ©2022 Association for Computational Linguistics

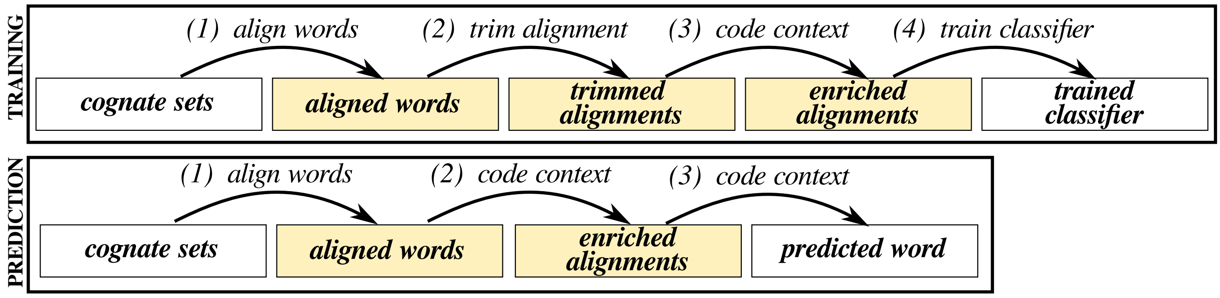

Figure 1: Workflow for the new framework for word prediction and linguistic reconstruction based on gap-free alignments and sound correspondence patterns.

Name	Source	Subgroup	L	C	W
`Bai`	Wang (2004)	Bai	10	459	3866
`*Burmish`	Gong and Hill (2020)	Burmish	9	269	1711
`*Karen`	Luangthongkum (2020)	Karen	11	365	3231
`Lalo`	Yang (2011)	Lalo (Yi)	8	1251	7815
`Purus`	Carvalho (2020)	Purus	4	199	693
`Romance`	Meloni et al. (2021)	Romance	6	4147	18806

Table 1: Datasets used in this study (L=Languages, C=Cognate Sets, W=Word Forms *=new data prepared for this study).

scholars (see Forkel et al. 2018 as well as List et al. 2021b for recent initiatives to make standardized cross-linguistic wordlists available in the form of open repositories). Unfortunately, the number of datasets in which proto-languages are provided along with descendant languages is still rather small. For the experiments reported here, a new cross-linguistic collection of six datasets from three language families (Sino-Tibetan, Purus, and Indo-European) was created. Datasets were all taken from published studies and then converted to Cross-Linguistic Data Formats (CLDF) (Forkel et al., 2018) using the CLDFBench Python package (Forkel and List, 2020) with the `PyLexibank` plugin (Forkel et al., 2021).

CLDF allows for a consistent handling of data when using software like Python or R. In addition, CLDF offers several levels of standardization by allowing to link the data to existing reference catalogs, such as Glottolog (Hammarström et al., 2021) for languages, Concepticon for concepts (List et al., 2021c), or Cross-Linguistic Transcription Systems (Anderson et al., 2018; List et al., 2021a) for speech sounds.

While three of the datasets (Bai, Lalo, and Purus) had been previously included into the Lexibank collection, a repository of lexical datasets in Cross-Linguistic Data Formats (List et al., 2021b), we converted the open part of the Latin dataset by Meloni et al. (2021) to CLDF. Additionally, we converted a selection of a smaller part of the data by Gong and Hill (2020) to CLDF and retro-standardized the data by Luangthongkum (2019). While all datasets provided forms for ancestral languages, not all datasets provided the direct links between these proto-forms and the reflexes in the descendant languages in the form of annotations indicating cognacy. While these were added manually for the Karen data, using the EDICTOR tool for etymological data curation (List, 2017, 2021), we used the automated method for partial cognate detection by List et al. (2016) to cluster proto-forms and reflexes into cognate sets for the data on Bai, Lalo, and Purus.

The datasets, along with their sources and some basic information regarding the number of languages (L), cognate sets (C), and word forms (W) are listed in Table 1. The collection offers a rather diverse selection, in which the amount of data varies both with respect to the number of word forms, cognate sets, and languages.

3 Methods

3.1 Workflow

The new framework can be divided into a training and a prediction stage. The training consists of four steps. In step (1), the cognate sets in the training data are *aligned* with a multiple phonetic alignment algorithm. In step (2), the alignments are *trimmed* by merging sounds in the ancestral language into clusters which would leave no trace in the descendant languages (§ 3.2). In step (3), the alignments of the descendant languages are enriched by *coding for context* that might condition sound changes (§ 3.3). In step (4) the enriched alignment sites are assembled and fed to a *classifier* for training.

The prediction consists of three steps. Given a cognate set as input, the word forms are aligned with the help of the same algorithm for multiple

	1	2	3	4	5	6	7
Latin	k	-	eː	n	aː	r	ɛ
	↑	↑	↑	↑	↑	↑	↑
Romanian	tʃ	-	i	n	a	-	-
Spanish	θ	-	e	n	a	ɾ	-
Portuguese	s	j	-	-	a	ɹ	-

Figure 2: Prediction problems when ancestral segments in multiple alignments do not show reflexes in the descendant languages.

alignment used in the training phase in step (1). In step (2), the alignment is enriched using the same method applied in the training phase and then passed to the classifier to predict the word form in the ancestral language in step (3).

Figure 1 illustrates the workflow, which is flexible with respect to individual methods used for individual steps. For phonetic alignment, we use the Sound-Class-Based Phonetic Alignment (SCA) algorithm (List, 2012), which is the current state-of-the-art method, but any other method that yields multiple alignments could be used. The same holds for the trimming procedure, (see § 3.2), the enrichment procedure, (see § 3.3), or the classifier (see § 3.4).

3.2 Trimming Alignments

Using multiple alignments to predict ancestral or new words is nothing new and has essentially been practised by classical historical linguists for a long time (Grimm, 1822). That multiple alignments can also be used in computational frameworks has been demonstrated by List (2019a), who inferred correspondence patterns from phonetic alignments and later used these correspondence patterns to predict words missing from the data. One problem not considered in this approach, however, is that correspondence patterns can only be inferred for those cases in which descendant languages have a reflex for a given sound in the ancestral language. In those cases where the sound has been lost, a prediction is not possible.

This problem is illustrated in Figure 2, where the Latin ending [ɛ] has no reflex sound in either of the descendant languages in the sample, yielding an alignment column that is completely filled with gap symbols. Our solution to deal with this problem is to post-process the multiple alignments in the training procedure by merging those columns which show only gaps in the descendant languages with the preceding alignment column. This is illustrated in Figure 3, where the Latin ending is now represented as a single sound unit [r.ɛ]. This trimming procedure, which was introduced for by Ciobanu and Dinu (2018) for pairwise alignments and is here extended to multiple alignments, is justified by the fact that correspondence patterns preceding lost sounds usually convey enough information to be distinguished from those patterns in which no sound has been lost.

3.3 Coding Context

Previous alignment-based approaches to automated word prediction have made exclusive use of the information provided by individual correspondence patterns derived from phonetic alignments (List, 2019a). While this has shown to yield already surprisingly good results, we know well that sound change often happens in certain phonetic environments. For example, we know that the initial position of a word is typically much stronger and less prone to change than the final position (Geisler, 1992). Similarly, consonants in the syllable onset position (preceding a vowel) also tend to show different types of sound change compared to consonants in the syllable offset (List, 2014). Last but not least, certain sound changes may be due to "long-range dependencies", or supra-segmental features like tone, which is typically marked in the end of a morpheme in the phonetic transcription of South-East Asian languages. In order to allow a classifier to make use of this information, our framework allows to enrich the phonetic alignments further, by deriving contextual information from individual phonetic alignments and adding it to the correspondence patterns that are then used to train the classifier. An example for this procedure

	1	2	3	4	5	6
Latin	k	-	eː	n	aː	r.ɛ
	↑	↑	↑	↑	↑	↑
Romanian	tʃ	-	i	n	a	-
Spanish	θ	-	e	n	a	ɾ
Portuguese	s	j	-	-	a	ɹ

Figure 3: Trimming alignments by merging sounds in the ancestral languages in those cases where an alignment column does not have sound reflexes in the descendant languages.

	Ro	Sp	Pt	P	S	Ini		Lt
1	tʃ	θ	s	1	C	^	→	k
2	-	-	j	2	C	−	→	-
3	i	e	-	3	V	−	→	eː
4	n	n	-	4	C	−	→	n
5	a	a	a	5	V	−	→	aː
6	-	ɾ	ɹ	6	c	$	→	r.ɛ

Figure 4: Enriching a phonetic alignment by coding various forms of context.

is given in Figure 4, where the phonetic alignment is given in transposed form (switching columns and rows), with each row corresponding to one correspondence pattern. While the information from correspondence patterns alone would only account for the first three columns of the matrix, three additional types of phonetic context have been added. Thus, column *P* indicates the position of a pattern in the form of an index. Column *S* provides information on the syllable structure following List (2014), and column *Ini* indicates, whether a pattern occurs in the beginning (^), the end ($) or the middle (−) of a word form. Enriching alignments should be done in a careful way, in order to avoid over-fitting the classifier. In our experiments, we contrast all eight possible combinations, ranging from the full coding shown in Figure 4, up to a coding of the alignment without additional enrichment.

3.4 Classifiers

Our approach is very flexible with respect to the choice of the classifier. In order to keep the approach *fast*, we decided to restrict our experiments to the use of a Support Vector Machine (SVM) with a linear kernel, since SVMs have been successfully applied in recent approaches in computational historical linguistics dealing with different classification tasks (Jäger et al., 2017; Cristea et al., 2021). We compare this approach with the graph-based method based on correspondence patterns (henceford called CorPaR) presented by List (2019a), which we modified slightly. While the original method uses a greedy algorithm to identify the largest cliques in the network, we now compute all cliques and rank them by counting the number of nodes they cover. An alignment site in an alignment is now compared against the consensus patterns extracted from the cliques in the graph and the prediction for the pattern with the largest number of reflexes is taken as the prediction. When no compatible pattern can be found, a search for the best candidates among patterns that are only partially compatible with the alignment site is invoked. This increases the chances too find a suitable reconstruction in those cases where the correspondence patterns are not fully regular.

3.5 Evaluation

Most scholars tend to report only the edit distance – also called Levenshtein distance (Levenshtein, 1965) – between the predicted and the attested string, both normalized by the length of the longer string and in unnormalized form. However, reporting the edit distance alone has the disadvantage that systematic differences between predicted and attested forms may be penalized too high, which is why we follow List (2019b) in computing the *B-Cubed F-scores* (Amigó et al., 2009) of the alignments of source and target sequences. B-Cubed F-Scores measure the difference between two classifications, ranging from 0 to 1, with 1 indicating complete similarity with respect to the structure of the classifications. Since the prediction of words can be seen as a classification task in which a certain number of sound slots should be classified by rendering them as identical or different from each other, B-Cubed F-Scores do not measure whether automated reconstructions are identical with attested reconstructions in the gold standard, but rather whether automated reconstructions approximate the structure of the reconstructions in the gold standard. As a result, B-Cubed F-Scores can show to which degree an automated reconstruction comes structurally close to the gold standard, even if individual reconstructed sounds differ. Given that B-Cubed F-Scores measure consistency across a set of reconstructed word forms, they should not be applied to individual items.

3.6 Implementation

The new framework is implemented as part of the LingRex Python package (List and Forkel, 2022) and allows the use of classifiers from the Scikit-Learn Python package (Pedregosa et al., 2011).

4 Results

In order to evaluate the framework, we tested two classifiers, a Support Vector Machine, and the Cor-

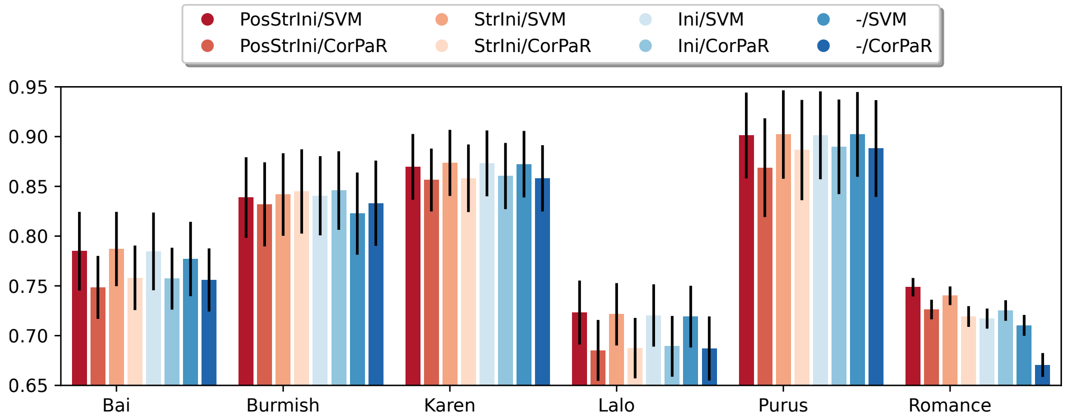

Figure 5: Comparing the results for selected coding techniques and classifiers on individual datasets.

PaR classifier (see § 3.4). Furthermore, we tested three different forms of alignment enrichment by coding individual positions (`Pos`), prosodic structure (`Str`), as well as whether a sound appears in the beginning or the end (`Ini`). For each test, we ran 100 trials in which 90% of the data were used for training and 10% for evaluation.

Classifier	Analysis	ED	NED	BC
SVM	PosStrIni	0.7491	0.1598	0.8110
SVM	PosStr	0.7478	0.1594	0.8115
SVM	PosIni	0.7701	0.1624	0.8077
SVM	StrIni	0.7578	0.1601	0.8110
SVM	Pos	0.7685	0.1618	0.8084
SVM	Str	0.7681	0.1614	0.8086
SVM	Ini	0.7895	0.1641	0.8061
SVM	none	0.8059	0.1673	0.8006
CorPaR	PosStrIni	0.8503	0.1816	0.7862
CorPaR	PosStr	0.8655	0.1826	0.7854
CorPaR	PosIni	0.8425	0.1802	0.7882
CorPaR	StrIni	0.8402	0.1771	0.7924
CorPaR	Pos	0.8836	0.1847	0.7840
CorPaR	Str	0.9048	0.1851	0.7848
CorPaR	Ini	0.8342	0.1763	0.7946
CorPaR	none	0.9379	0.1898	0.7821

Table 2: Results for edit distance, normalized edit distance, and B-Cubed F-Scores on all datasets.

Table 2 shows the results for all eight combinations between the three techniques for alignment enrichment. As can be seen, the SVM classifier outperforms the CorPaR method, although the differences are not very large. While the impact of the alignment enrichment techniques on the results is not very large, we still find that they enhance the results in all SVM trials, while the raw coding of the position (`Pos`) leads to lower scores for the CorPaR classifier in our test set. For the SVM classifier, coding for prosodic structure (`Str`) and information on whether a segment occurs at the beginning, in the middle, or the end of a sequence (`StrIni`) yields the best results with respect to all measures, while `Ini` coding outperforms the other techniques for the CorPaR classifier. From these results, we can see that alignment enrichment is a promising technique that deserves further exploration, but we do not think that the current codings are the last word on the topic.

Figure 5 compares the results for four coding techniques on individual datasets. As can be seem from the figure, the impact of the coding techniques varies quite drastically across datasets. This shows that it would be premature to rule out any of the techniques tested here directly, but rather calls for a careful selection of alignment enrichment techniques dependent on the language family one wants to investigate.

5 Conclusion

In this study, we have presented a new framework for supervised phonological reconstruction, which is implemented in the form of a small Python package. The new framework has the advantage of being easy to use, easy to extend, and fast to apply, while at the same time yielding promising results on a newly compiled collection of datasets from three different languages families. Given that our framework can be easily extended, by varying the individual components of the worfklow, we hope that it will provide a solid basis for future work on phonological reconstruction, as well as the prediction of words from cognate reflexes (Bodt and List, 2022; Dekker and Zuidema, 2021; Beinborn et al., 2013; Fourrier et al., 2021) in computational historical linguistics.

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.

Cormac Anderson, Tiago Tresoldi, Thiago Costa Chacon, Anne-Maria Fehn, Mary Walworth, Robert Forkel, and Johann-Mattis List. 2018. A Cross-Linguistic Database of Phonetic Transcription Systems. *Yearbook of the Poznań Linguistic Meeting*, 4(1):21–53.

Lisa Beinborn, Torsten Zesch, and Iryna Gurevych. 2013. Cognate production using character-based machine translation. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 883–891.

Timotheus A. Bodt and Johann-Mattis List. 2019. Testing the predictive strength of the comparative method: An ongoing experiment on unattested words in Western Kho-Bwa languages. *Papers in Historical Phonology*, 4(1):22–44.

Timotheus Adrianus Bodt and Johann-Mattis List. 2022. Reflex prediction. A case study of Western Kho-Bwa. *Diachronica*, 39(1):1–38.

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 of the United States of America*, 110(11):4224–4229.

Alina Maria Ciobanu and Liviu P. Dinu. 2018. Simulating language evolution: A tool for historical linguistics. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 68–72. Association of Computational Linguistics.

Alina Maria Cristea, Liviu P. Dinu, Simona Georgescu, Mihnea-Lucian Mihai, and Ana Sabina Uban. 2021. Automatic discrimination between inherited and borrowed Latin words in Romance languages. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2845–2855, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Fernando O. de Carvalho. 2021. A comparative reconstruction of proto-purus (arawakan) segmental phonology. *International Journal of American Linguistics*, 87(1):49–108.

Peter Dekker and Willem Zuidema. 2021. Word prediction in computational historical linguistics. *Journal of Language Modelling*, 8(2):295–336.

Liviu Dinu and Alina Maria Ciobanu. 2014. Building a dataset of multilingual cognates for the Romanian lexicon. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1038–1043, Reykjavik, Iceland. European Language Resources Association (ELRA).

Robert Forkel, Simon J Greenhill, Hans-Jörg Bibiko, Christoph Rzymski, Tiago Tresoldi, and Johann-Mattis List. 2021. *PyLexibank. The python curation library for lexibank [Software Library, Version 2.8.2]*. Zenodo, Geneva.

Robert Forkel and Johann-Mattis List. 2020. Cldfbench. give your cross-linguistic data a lift. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation*, pages 6997–7004, Luxembourg. European Language Resources Association (ELRA).

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. 2018. Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data*, 5(180205):1–10.

Clémentine Fourrier, Rachel Bawden, and Benoît Sagot. 2021. Can cognate prediction be modelled as a low-resource machine translation task? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 847–861, Online. Association for Computational Linguistics.

Anthony Fox. 1995. *Linguistic reconstruction*. Oxford University Press, Oxford.

Hans Geisler. 1992. *Akzent und Lautwandel in der Romania*. Narr, Tübingen.

Xun Gong and Nathan Hill. 2020. *Materials for an Etymological Dictionary of Burmish*. Zenodo, Geneva.

Jacob Grimm. 1822. *Deutsche Grammatik*, 2 edition, volume 1. Dieterichsche Buchhandlung, Göttingen.

Harald Hammarström, Martin Haspelmath, Robert Forkel, and Sebastiaon Bank. 2021. *Glottolog [Dataset, Version 4.5]*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

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

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.

Grzegorz Kondrak. 2000. A new algorithm for the alignment of phonetic sequences. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 288–295.

Vladimir. I. Levenshtein. 1965. Dvoičnye kody s ispravleniem vypadenij, vstavok i zameščenij simvolov. *Doklady Akademij Nauk SSSR*, 163(4):845–848.

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

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

Johann-Mattis List. 2017. A web-based interactive tool for creating, inspecting, editing, and publishing etymological datasets. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. System Demonstrations*, pages 9–12, Valencia. Association for Computational Linguistics.

Johann-Mattis List. 2019a. Automatic inference of sound correspondence patterns across multiple languages. *Computational Linguistics*, 45(1):137–161.

Johann-Mattis List. 2019b. Beyond Edit Distances: Comparing linguistic reconstruction systems. *Theoretical Linguistics*, 45(3-4):1–10.

Johann-Mattis List. 2021. *EDICTOR. A web-based tool for creating, editing, and publishing etymological datasets [Software Tool, Version 2.0.0]*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Johann-Mattis List, Cormac Anderson, Tiago Tresoldi, and Robert Forkel. 2021a. *Cross-Linguistic Transcription Systems [Dataset, Version 2.1.0]*. Max Planck Institute for the Science of Human History, Jena.

Johann-Mattis List and Robert Forkel. 2022. *LingRex: Linguistic reconstruction with LingPy*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Johann-Mattis List, Robert Forkel, Simon J. Greenhill, Christoph Rzymski, Johannes Englisch, and Russell D. Gray. 2021b. Lexibank: A public repository of standardized wordlists with computed phonological and lexical features [Preprint, Version 1]. *Research Square*, pages 1–31.

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

Johann-Mattis List, Christoph Rzymski, Simon J. Greenhill, Nathanael E. Schweikhard, Kristina Pianykh, Annika Tjuka, Carolin Hundt, and Robert Forkel. 2021c. *Concepticon. A resource for the linking of concept lists [Dataset, Version 2.5.0]*. Max Planck Institute for the Science of Human History, Jena.

Theraphan Luangthongkum. 2019. A view on Proto-Karen phonology and lexicon. *Journal of the Southeast Asian Linguistics Society*, 12(1):i–lii.

Carlo Meloni, Shauli Ravfogel, and Yoav Goldberg. 2021. 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, Online. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Lydia Steiner, Peter F. Stadler, and Michael Cysouw. 2011. A pipeline for computational historical linguistics. *Language Dynamics and Change*, 1(1):89–127.

Feng Wang. 2004. *Language contact and language comparison. The case of Bai*. Phd, City University of Hong Kong, Hong Kong.

Cathryn Yang. 2011. *Lalo regional varieties: Phylogeny, dialectometry and sociolinguistics*. PhD dissertation, La Trobe University, Bundoora.

A Appendix

A.1 Source Code and Data

The new data collection along with the source code and the data needed to replicate the results reported in this study have been curated on GitHub at https://github.com/lingpy/supervised-reconstruction-paper (Version 1.0) and archived with Zenodo (DOI: https://doi.org/10.5281/zenodo.6426074).

A.2 Table of Results for Individual Datasets

A.2.1 SVM

DATASET	PosStrIni	StrIni	Str	Ini	none
Bai	0.7848	0.7870	0.7832	0.7846	0.7770
Burmish	0.8388	0.8418	0.8420	0.8405	0.8226
Karen	0.8696	0.8736	0.8734	0.8731	0.8723
Lalo	0.7232	0.7214	0.7204	0.7202	0.7191
Purus	0.9011	0.9021	0.9016	0.9013	0.9022
Romance	0.7487	0.7401	0.7310	0.7171	0.7103

A.2.2 CorPaR

DATASET	PosStrIni	StrIni	Str	Ini	none
Bai	0.7485	0.7581	0.7560	0.7572	0.7560
Burmish	0.8319	0.8449	0.8422	0.8458	0.8331
Karen	0.8564	0.8581	0.8614	0.8604	0.8581
Lalo	0.6852	0.6874	0.6890	0.6893	0.6871
Purus	0.8688	0.8865	0.8730	0.8897	0.8880
Romance	0.7262	0.7192	0.6871	0.7253	0.6705