Representing and Computing Uncertainty in Phonological Reconstruction

Johann-Mattis List
MCL Chair / DLCE
University of Passau / MPI-EVA
Passau / Leipzig, Germany

Nathan W. Hill
Trinity Centre for Asian Studies
University of Dublin
Dublin, Ireland

Robert Forkel
DLCE
MPI-EVA
Leipzig, Germany

Frederic Blum
DLCE
MPI-EVA
Leipzig

Abstract

Despite the inherently fuzzy nature of reconstructions in historical linguistics, most scholars do not represent their uncertainty when proposing proto-forms. With the increasing success of recently proposed approaches to automating certain aspects of the traditional comparative method, the formal representation of proto-forms has also improved. This formalization makes it possible to address both the representation and the computation of uncertainty. Building on recent advances in supervised phonological reconstruction, during which an algorithm learns how to reconstruct words in a given proto-language relying on previously annotated data, and inspired by improved methods for automated word prediction from cognate sets, we present a new framework that allows for the representation of uncertainty in linguistic reconstruction and also includes a workflow for the computation of fuzzy reconstructions from linguistic data.

1 Introduction

Phonological reconstruction refers to the techniques that linguists use to reconstruct the phonological and phonetic shape of word forms or morphemes in unattested ancestral languages. Although the results are inherently provisional (as witnessed by the changes in the fable by Schleicher 1868 over the last decades, cf. Lühr 2008), linguists typically present their results in the form of discrete phonological units, giving the impression of exactitude and rigor. Thus, although phonological reconstructions change with time as the knowledge or assumptions about a language family change, scholars typically provide the results as if they were final. By focusing on the uncertainty of phonological reconstructions, we aim to provide a new framework by which uncertainty in phonological reconstruction can be a) represented (in etymological databases or etymological dictionaries), and b) computed (from etymological datasets). Representation and computation have several benefits. On the one hand, improved representations allow for a more refined reconstruction practice that more directly and consistently indicates the weak points in a reconstruction. On the other hand, computing the uncertainty of a given reconstruction system allows scholars to refine their reconstructions by helping them to identify weak points and potential errors in their cognate judgments or correspondence patterns.

The traditional techniques for phonological reconstruction, by which ancestral word forms are reconstructed from observed words with the help of the comparative method, are of crucial importance for historical language comparison. Despite the inherently fuzzy nature of reconstructions, most scholars have so far hesitated to systematically represent their uncertainty when proposing proto-forms (for an exception see Baxter and Sagart 2014), and discussions of uncertainty are very spurious in the literature. With the increasing success of recently proposed techniques by which certain aspects of the traditional comparative method can be automated, the formal representation of words, morphemes, cognate sets, and proto-forms has also improved. This makes it possible to address the problems of both the representation and the computation of uncertainty. Supervised approaches have led to major advances in automated phonological reconstruction; scholars provide a partially annotated dataset in which a certain number of proto-forms are already provided, and an algorithm is then trained on the data in order to propose new proto-forms for so far unobserved cognate sets. This task is very similar to the task of cognate reflex prediction, in which the word forms which

Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change, pages 22–32
December 6, 2023 ©2023 Association for Computational Linguistics

have to be predicted are not proto-forms, but word forms from descendant languages, and algorithms have to predict the reflex of a cognate set in a given language based on the sound correspondence patterns and the reflexes of the cognate set in related languages. In the past decade, scholars have proposed quite a few new methods for both cognate reflex prediction and supervised phonological reconstruction.

Meloni et al. (2021) tested recurrent neural networks on a dataset of Romance languages originally compiled by Ciobanu and Dinu (2013), reporting very promising results on supervised approaches. This study was later extended by Kim et al. (2023), who used a Transformer architecture (Vaswani et al., 2017) and additionally tested the approach on a dataset of Chinese dialect varieties. List et al. (2022b) proposed a new framework based on support vector machines to predict proto-forms from phonetic alignments, which they tested on six different datasets covering several different language families. In a recently organized Shared Task on cognate reflex prediction (List et al., 2022c), Kirov et al. (2022) proposed two methods that outperformed alternative approaches, one originally designed for the handling of place name pronunciations in Japan (Jones et al., 2022), and one designed for the restoration of digital images in which pixels are missing (Liu et al., 2018). All in all, the most successful methods in the shared task all showed good performance: when retaining 90% of the data for training, the methods differed on average by one sound from the attested word forms.

While the task of unsupervised phonological reconstruction, where algorithms would reconstruct a proto-language from cognate sets from scratch, has not been sufficiently investigated so far (an early approach by Bouchard-Côté et al. 2013 was only tested on Austronesian languages with the code never published), we can see that phonological reconstruction in a supervised setting has become a real option and could be integrated into computer-assisted workflows, in which scholars first annotate parts of their data, then compute new reconstructions automatically, and later refine them again.

With respect to the representation of uncertainty in reconstruction, linguists typically adopt ad-hoc solutions for individual language families or individual enterprises. In Indo-European studies, scholars express their uncertainty with respect to the three laryngeals (*h_1, *h_2 or h_3) by writing a capital *H. In their reconstruction of Old Chinese, Baxter and Sagart (2014) employ a complex notation system that puts uncertain parts of their reconstruction into brackets (with *-[n]* meaning, for example, that the reconstruction could be either the final *-n* or to *-r*). In other cases, scholars mention alternative reconstructions only in comments. While both manual and automated methods are inherently fuzzy with respect to phonological reconstruction, so far, few methods have explicitly embraced fuzziness, trying to present uncertainty in reconstructions explicitly. An exception was the method of List (2019), which offered degrees of uncertainty in the imputation of missing sounds in aligned cognate sets, but the fuzzy reconstructions were not further evaluated or inspected.

2 Materials

We work with three etymological datasets which are coded in Cross-Linguistic Data Formats (https://cldf.clld.org, Forkel et al. 2018; Forkel and List 2020), following the Lexibank workflow for the handling of multilingual wordlists (List et al., 2022a). The Burmish dataset consists of 8 Burmish languages, and 269 etymologies that are reflected in at least two descendant languages with a total of 1,442 reflexes. The data was originally compiled by Gong and Hill (2020) and later converted to CLDF formats by List and Forkel (2022) and further refined for the present study. It is accessible online at https://github.com/lexibank/hillburmish. The Karen dataset consists of 10 Karenic languages, and offers 365 etymologies originally proposed by Luangthongkum (2019), which are reflected in at least 2 languages with a total of 2,866 reflexes. The data was also compiled for the study by List and Forkel (2022) and slightly refined for this study. It is accessible online at https://github.com/lexibank/luangthongkumkaren. The Panoan dataset consists of 20 Panoan languages, and includes the reconstruction of 514 cognate sets across 470 concepts proposed by Oliveira (2014). In total, the dataset features 7,305 reflexes. During the digitization of this dataset, all cognate sets were manually aligned (Blum and Barrientos, 2023). It is accessible online at https://github.com/pano-tacanan-history/oliveiraprotopanoan.

Reconstruction	Initial	Nucleus	Coda	Tone
Predictor 1	d	u	-	²
Predictor 2	d	u	-	¹
Predictor 3	d	u	-	¹
Predictor 4	d	u	-	⁴
Predictor 5	d	u	-	⁴
Predictor 6	d	u	-	⁴
Predictor 7	d	u	-	⁴
Predictor 8	d	u	-	⁴
Predictor 9	d	u	-	⁴
Predictor 10	ˀt	u	-	⁴
Summary	d:90|ˀt:10	u:100	-:100	⁴:80|¹:20|²:10
Proto-Form	d	u		²

Table 1: Predictions for "belly" (cognate set 80) in Burmish. The table contrasts predicted word forms for all 10 different predictors, along with the aggregated representation (row Summary) and contrasted with the reconstruction provided by the experts (Proto-Form).

3 Methods

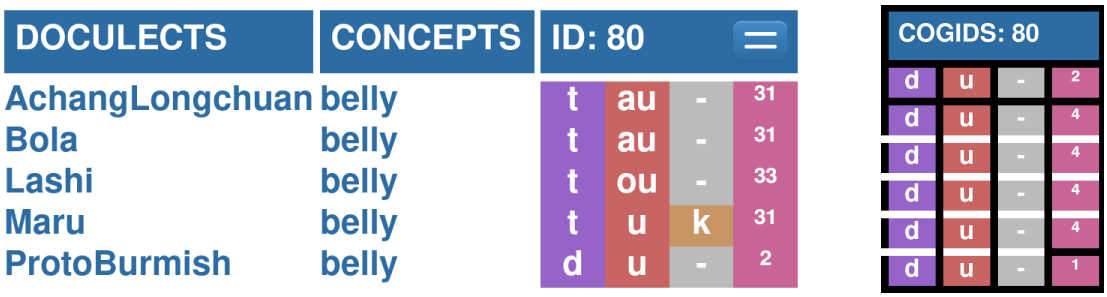
3.1 Representing Fuzzy Reconstructions

We follow Bodt and List (2022) who represent multiple options for the prediction of an individual sound by using the pipe symbol | as a separator for the different options. The symbol is often used in the meaning of "or" in regular expressions, which makes it particularly apt to represent uncertainty, since we can interpret a fictitious proto-form like [p a|i t] as a kind of a regular expression that matches both the form [p a t] and [p i t]. Note that this notation needs to be used with some care when more than one sound is treated as uncertain, since the resulting expression will always match the Cartesian product of the uncertain sounds. Thus, a fictitious proto-form [p a|i t|d] would match four distinct proto-forms, namely the forms [p a t], [p i t], [p a d], and [p i d]. If scholars want to explicitly propose two different proto-forms only, e.g. [p a t] vs. [p i d], our notation cannot be used. We recommend instead to assume two distinct forms, which can both be proposed as possible proto-forms for a given cognate set. Our fuzzy notation is thus only reserved for cases where the uncertainty is independent of contextual information that could be derived from the proto-form.

3.2 Computing Fuzzy Reconstructions

Our method for the creation of fuzzy reconstructions is straightforward. We expand the framework for supervised phonological reconstruction proposed by List et al. (2022b), by drawing several samples from the same data, in which different parts of the forms are intentionally ignored. While the framework of List et al. starts from a training set in which proto-forms are provided and then a model is trained that can be used to predict proto-forms for data that has not been seen before, we draw multiple samples, drop a certain number of words from each sample, and use the method by List et al. to train the "classifier" that can be used to predict proto-forms from aligned cognate sets. Since we drop data in each of the samples, each sample will produce slightly different proto-forms, depending on the data which has been randomly ignored. The different proto-forms offered may point to problems in the original data, or reveal cognate sets that in fact underspecify the proto-form.

While fuzziness could of course also be directly computed from the direct output of most approaches to supervised phonological reconstruction (since most of them work in a probabilistic manner that allows one to return not only the one and only best result, but also a certain range of candidates, see also Fourrier et al. 2021), our approach of using subsets of the original data has the clear advantage that it does not take the correctness of the original data for granted. When taking all data at once, it is difficult to spot irregularities in the data itself. When taking subsets, however, we test the robustness of the reconstructions for individual cognate sets. If the reconstruction, for example, depends on only one reflex, but this reflex is then discarded due to the subsampling in one particular



(a) Phonetic alignment of all word forms (including the proto-form). (b) Quintile representation.

Figure 1: Contrasting the alignment representation with the quintile representation of the fuzzy reconstruction in the EDICTOR tool.

run, the resulting reconstruction may turn out to be different, and this particular difference would then be accounted for in this trial and surface as an uncertainty.

In the default settings of our method, we create 10 proto-form predictors from the annotated data and remove 10% of the word forms in each of the samples. When creating an individual reconstruction, we feed our method with a concrete cognates set and then use all 10 predictors to predict proto-forms. The predictions are then summarized, and we count for each position in the original alignment how often which proto-sound occurs. These fuzzy reconstructions are then represented in the form of a sequence in which a column of the alignment is represented by at least one sound, and each possible sound is provided with the frequency in which it occurs in our 10 samples. Table 1 provides an example from the Burmish data for the fuzzy prediction procedure and the specific output produced by our method.

Since certain irregularities in the input data may be filtered from the different samples, irregular patterns which could lead an algorithm to propose erroneous proto-forms will be filtered out, and in this way the overall robustness of individual reconstruction can be tested.

3.3 Visualizing Fuzzy Reconstructions

Apart from the technical representation shown above, we have experimented with different ways to represent uncertainty or "fuzziness" in the tools we use to annotate etymological data. Since the manual curation of the cognate sets was carried out with the help of EDICTOR (List 2023, https://digling.org/edictor), a web-based tool for the creation and curation of etymological datasets, we extended the EDICTOR representation of phonetic alignments by adding a representation which we call quintile-representation. In this representation, we represent the frequencies observed in the ten predictions with the help of a table with five rows, in which each row represents the attested symbols (converted from 10 to 5, to keep the table representation neat). This is shown in Figure 1.

3.4 Implementation

The method of fuzzy reconstruction is implemented as Version 1.4.1 of the LingRex software package (https://pypi.org/project/lingrex, List and Forkel 2023b, which is itself an extension of the LingPy software package for quantitative tasks in historical linguistics (https://pypi.org/project/lingpy, List and Forkel 2023a). The quintile visualization is implemented as part of Version 2.2 of the EDICTOR tool (https://digling.org/edictor, List 2023). The supplementary material shows how the package can be used and applied to the data, it is curated on GitHub (https://github.com/lingpy/fuzzy/releases/tag/v1.1) and has been archived with Zenodo (https://doi.org/10.5281/zenodo.10007475).

4 Evaluation

Since we do not have a clear account on what constitutes a good "fuzzy reconstruction" and what constitutes a bad one, we closely analyzed the fuzzy reconstructions proposed for the three datasets and further investigated the results both quantitatively and qualitatively. In the following, we will thus report on the proportion of fuzzy reconstructions per datasets, the most frequently confused sounds in fuzzy reconstructions, and then report on major problems in the original data revealed through a

Dataset	Prediction	Count	Proportion	Alignment Size
Burmish	correct	154	0.57	4.13
	false	115	0.43	4.29
	certain	199	0.74	4.13
	uncertain	70	0.26	4.39
Karen	correct	246	0.65	4.03
	false	133	0.35	4.27
	certain	310	0.82	4.05
	uncertain	69	0.18	4.41
Panoan	correct	405	0.79	4.25
	false	109	0.21	5.14
	certain	465	0.90	4.37
	uncertain	49	0.10	5.14

Table 2: Summary scores for the Burmish, Karenic, and Panoan predictions. Correct predictions refer to all those predictions which are identical with the reconstruction in the gold standard and which show no uncertainty. False predictions are those which show uncertainty or which are not identical with the proposed predictions in the gold standard. Certain predictions are those in which all ten trials on differently distorted data show the same results for a given proto-form. Uncertain predictions are those, accordingly, in which we observe differences. Alignment size refers to the size of the alignment of the cognate sets (conducted automatically).

close inspection of the fuzzy reconstructions proposed for the Burmish dataset.

4.1 Proportion of Fuzzy Reconstructions

As a first test of our approach, we computed fuzzy reconstructions from the three datasets and then compared whether (a) the reconstructions were fuzzy at all, and (b) to what extent they diverged from the proposed reconstructions. We explicitly chose a setting where we train and test the method on the same dataset, since we were not interested in the evaluation of the reconstruction method (which performs fairly well, but not perfect) but in the degree to which conclusions were based on the data in its entirety or different parts of it. For each proto-form in the three datasets, we computed fuzzy reconstructions, from which we created consensus reconstructions using the notation shown in Table 1. For each proposed reconstruction we tested (a) if the reconstruction had conflicts (i.e. if it was "fuzzy"), and (b) if it was not fuzzy, if it coincided with the reconstruction proposed by the linguist.

For the Burmish data, consisting of a total of 269 reconstructions, we arrived at the results reported in Table 2 (top). As can be seen from the table, the proportion of words reconstructed correctly by the approach and proportion of words that were reconstructed as "certain" (with no variation) is much larger than the proportion of false or uncertain reconstructions. Since a correct reconstruction has to be certain, it is not surprising that these numbers are similar, but the small difference of 57% vs. 74% shows that only a small part of the reconstructions identified as "certain" are also wrong. We find a small difference with respect to the alignment size (the number of words of which alignments for individual proto-forms are reconstructed) between correctly and falsely reconstructed proto-form, but due to the restricted syllable structure of Burmish languages, we do not find huge differences here. Additional studies are needed to find out what influences the certainty of automated reconstructions.

The results for the Karen data, consisting of 365 cognate sets, are shown in Table 2 (middle). As can be seen here, the number of correctly reconstructed proto-forms as well as the number for certain proto-forms are both higher than in the case of the Burmish data. One factor which may have contributed to this is that exceptional reflexes in this dataset have been manually identified and marked as such (as part of ongoing research), which means that certain irregularities in the data did not negatively impact the predictions. In contrast to the Burmish data, the differences in alignment size for correct vs. false proto-forms and certain vs. uncertain ones are more pronounced in this dataset.

The results for the Panoan data in Table 2 show some interesting differences. Of the total of 514 cognate sets, 405 (79%) are reconstructed correctly, a considerably higher number than for the other

Burmish			Karen			Panoan		
Sound A	**Sound B**	**Freq.**	**Sound A**	**Sound B**	**Freq.**	**Sound A**	**Sound B**	**Freq.**
^{4}	^{1}	14	n	n̥	18	n	r^{n}	24
^{4}	^{3}	9	n	ɴ	14	k	-	13
i	e	8	ɴ	ŋ	10	r^{n}	~	10
ŋ	-	7	^{55}	^{0}	9	-	t^{r}	10
^{2}	^{3}	7	l	l̥	8	n	-	9
-	ʔ	7	m̥	m	6	r^{n}	-	6
^{ʔ}s	s	6	-	ʔ˥	6	k	t^{r}	6
^{ʔ}k	g	6	^{1}	^{0}	5	t	t^{r}	5
^{2}	^{4}	6	k	g	5	t	-	5
r	j	6	ʔ˥	ʔ	4	r^{n}	r	5

Table 3: Frequently confused sounds in the three datasets. Frequency refers to the number of cognate sets in which the automated reconstruction proposed alternative proto-sounds.

datasets. The number of reconstructions that are provided as "certain" is also higher (90%) than for the other datasets. There is also a considerable difference in alignment size: The alignment size for correct (4.25) and "certain" (4.37) reconstructions is much lower than for false (5.14) and "uncertain" (5.14) reconstructions. Here, a larger alignment size arises as a possible source influencing the correctness and certainty of automated reconstructions. This illustrates that it may be worthwhile to investigate more closely how the reconstructions differ between different language families and between alignments of different sizes within the language families.

4.2 Frequently Confused Sounds

Each fuzzy reconstruction proposes at least two alternative sounds for one proto-segment in a given proto-form. Investigating these more closely in order to understand which sounds are frequently confused by the analysis, allows us to gain insights into those sounds in the proto-languages which are particularly difficult to reconstruct. Table 3 provides the 10 most frequently confused sound pairs in both datasets (our workflow reports all of them).

As can be seen from the individual results for the Karen and Burmish data, the particular problems are quite different across both datasets and cannot be directly compared with each other. A major difficulty in the Karenic data is the reconstruction of voiceless sonorants ([n̥], [l̥], [m̥], etc.), which the author proposes on the basis of the tonal development in some of the descendant languages (Luangthongkum, 2019). Since there are quite a few exceptions with respect to the tonal development, we find that the original reconstruction itself cannot always indicate clearly whether a proto-sound should be voiced or voiceless, which is at times marked by putting the h, which is used to mark a sonorant as voiceless in parentheses (resulting in forms like (h)n-, ibid.). The confusion of the tone marked as [^{0}] with other tones results from our annotation practice of certain weak syllables, in which originally no tone was reconstructed. Since we wanted to indicate a tone nevertheless, to fill the slot in our alignment, the [^{0}] thus marks an underspecified value, which – as the fuzzy reconstructions show – might just as well be given a more concrete reconstruction.

In the Burmish data, on the other hand, we find three major types of confusion. The first relates to the reconstruction of tones. The reconstruction here is often predicted by the nature of the final consonants, which are not actively used in the automated reconstruction method. This may explain the confusion in this case. The second case relates to the reconstruction of gaps (marked by the symbol [-]), which are often confused with sounds occurring in coda position, such as [ŋ], [r], or [ʔ]. The confusion of pre-glottalized initials like [^{ʔ}s] and [^{ʔ}k] and their non-glottalized counterparts also results from the fact that the reconstruction of pre-glottalized initials depends on the vowels that appear as reflexes in certain Burmish languages. Since this information was not taken into account by our automated method, it is not surprising that results may vary here. The confusion resulting from information that is not represented in the individual column of an alignment but in other parts shows

Achang	m̥	z̥	a	ŋ	31
Old Burmese	k^h	j	i	j	5
Fuzzy Proto-Form	ʔm:70|ʔk:30	r:50|j:30|-:20	i:80¦a:20	-:100	2:100
Proto Burmish	ʔk	j	i		2

(a) Erroneous cognate judgments in cognate set #288 “dung (horse)” for Achang.

Achang	s	ɔ	ʔ	55
Atsi	p	a	n	21
Bola	x	a	ʔ	55
Lashi	ʃ	ɐ̲	ʔ	55
Maru	s	o	ʔ	55
Rangoon	tθ	ɑ	-	53
Xiandao	s	ɔ	ʔ	55
Fuzzy Proto-Form	ʔs:70|s:30	a:100	k:100	4:100
Proto Burmish	ʔs	a	k	4

(b) Context-dependency of individual patterns in cognate set #536 “shy, be / bashful”.

Achang	ts	ɔ	-	31
Atsi	s	i	k	55
Bola	t	a	-	31
Lashi	l	ə	ŋ	55
Maru	ts	ɔ	-	
Old Burmese	c	a	-	55
Rangoon	s^h	ɑ̃	-	53
Xiandao	s	ɔ	ŋ	55
Fuzzy Proto-Form	dz:100	a:100	-:90|ŋ:10	2:100
Proto Burmish	dz	a	-	2

(c) Deep and unclear etymologies in cognate set #93 “granddaughter”.

Atsi	m	j	i̲	-	55
Bola	m	-	ə̲i	-	31
Lashi	m	j	e̲:i	-	53
Old Burmese	m	-	i	j.ʔ	3
Rangoon	m	-	i	-	53
Xiandao	n	-	i	-	35
Fuzzy Proto-Form	m:100	-:100	e:90|i:10	-:100	3:100
Proto Burmish	m	-	i	-	3

(d) Systematic ambiguities in particular languages in cognate set #414 “forget”.

Table 4: Examples for causes of fuzziness in Burmish reconstructions.

that additional analyses in which we take the vowel information in the Burmish languages and the tonal information in the Karenic languages into account would be useful in the future.

The confused sounds for the Panoan data fall mainly into two groups, (a) word-final stops [k], [t], and [t^r], and (b) word-final nasal and liquid consonants [n], [r^n], and [r]. Interestingly, those cases are either described as uncertain due to missing data by the original author (word-final nasals), or are the most debated feature of the reconstruction (word-final stops instead of three-syllabic words with an open syllable). The word-final sonorants are described by the author of the dataset as being uncertain due to the lack of reflexes in Kaxararí, a nearly undescribed Panoan language which retains the contrast of word-final [r] and [n]. This is the main source of confusion for [n], [r^n], and [r], but also for some of the word-final stops. Here, the confusion primarily arises because the reconstructions are proposed based on reflexes of only a few languages, which often do not provide sufficient evidence for identifying the phonemic nature of the reflex in the proto-language. Our method thus correctly identifies the cases in which the provided reconstructions should be considered "fuzzy", given their uncertain nature. It also validates the large part of correct reconstructions.

4.3 Detailed Examples for Burmish

A closer inspection of discrepancies in the Burmish data reveals four major kinds of problems, namely (1) problems resulting from problematic cognate judgments in our data, (2) problems resulting from the context-dependency of reconstructions which our automated reconstruction method does not (yet) account for, (3) problems resulting from deep etymologies which are not (yet) well understood, and (4) problems resulting from some systematic and so far not clearly understood ambiguities in particular languages.

4.3.1 Problematic Cognate Judgments

The method allows us to identify quite a few cases where individual cognate judgments turned out to be erroneous and should be modified in future versions of our data. As an example, consider cognate set #288 "dung (horse)" in our Burmish data, shown in Table 4 (a). That erroneous cognate judgments occur in larger etymological projects is inevitable to some degree. Here, our method for the reconstruction of "fuzzy" proto-forms directly helps us to identify and eliminate these problems in future releases of our data.

4.3.2 Context-Dependency of Reconstructions

While phonological reconstruction can, in the majority of the cases, be successfully carried out by considering individual correspondence patterns alone, there are certain cases where it is not enough to look at a pattern in isolation. What needs to be done instead is to evaluate the pattern in combination with other patterns from the same alignment. Although our method for automatic phonological reconstruction was designed in such a way that it can in theory account for this context-dependency of individual reconstructions, we did not take specific and known processes of sound change in the Burmish and the Karenic data into account, when applying our method to the data. This was done intentionally, since we wanted to see how far we can get with a unified approach. Individual reconstruction errors and cases of uncertainty in the automated reconstruction, however, show that context-dependency should be accounted for in future applications of our approach.

As an example for the problems resulting from ignoring context-dependencies, Table 4 (b) shows the reconstruction for the cognate set #536 "shy, be / bashful" in the Burmish data. As we can see, Lashi has a tense vowel (indicated by the bar under the vowel, shaded in gray in the table). Tense vowels are taken as evidence for the reconstruction of pre-glottalized initials in Proto-Burmish, while the correspondence pattern of the initial itself does not provide concrete evidence for the presence or absence of pre-glottalization. As a result, we can see that the automated method is uncertain, proposing a pre-glottalized initial in 70% of the cases, and a plain initial in 30%.

List and Forkel (2022) have described in detail, how context-dependency can be accounted for by means of "extended alignments" or "multi-tiered sequence representations". Future studies are needed to test how well these work to handle the Burmish (and also the Karenic) data.

4.3.3 Etymologies with Unclear Variation

There are a couple of cognate sets where we have in principle no doubts that the words in question are cognates, but we have problems to understand the etymological processes in full. These deep etymologies with unclear variation are usually of great importance when it comes to advancing ex-

isting reconstruction systems. However, since they may well reflect processes predating the history of the language family in question, the solution may only be achieved when taking more languages from higher clades of the language family in question into account.

As an example, consider Table 4 (c), showing alignments and reconstructions for cognate set #93 "granddaughter". While it is possible that all forms are cognate, it is hard to decide for sure, given that individual languages show reflexes which do not follow our expectations. Thus the initial [l-] in Lashi does not fit the pattern at all, and from the pattern, we have evidence for three different finals in the data. Future work may either show that we have to refine the cognate assignment of individual words in this pattern, or we may find solutions in certain etymological processes that counteract regular sound change.

4.3.4 Systematic Ambiguities in Languages

As a final type of difficulty, there are cases where we have clear ambiguities in individual languages which we cannot (yet) resolve and explain. As an example, Table 4 (d) shows ambiguities for the reconstruction of the vowel nucleus in the cognate set #414 "forget", where our reconstruction proposes **i*, while the automated method sees more evidence for **e* (90%) and less evidence for **i* (10%). The evidence from the correspondence pattern is difficult to interpret. While Old Burmese points to an **i*, Bola and Lashi point to an **e*. The fuzzy reconstruction approach thus correctly points to the ambiguity of the pattern in the light of our data.

5 Conclusion

In this study, we have introduced some novel ideas regarding the handling of uncertainty in phonological reconstruction in historical linguistics. We have tried to show that it may be useful to transparently record uncertainty not only in classical reconstructions but also in reconstructions proposed by automated approaches.

These considerations resulted in the proposal of a new framework for fuzzy reconstructions that allows one to compute fuzzy reconstructions from annotated comparative wordlists. Applying this framework to three datasets, two from the Sino-Tibetan language family (Burmish and Karenic), as well as on the Panoan language family, we have shown how the framework can be used to compute the degree of uncertainty in a given dataset, how frequently confused sounds can be computed, and how an individual inspection of the data reveals major classes of errors in the original data.

In the future, we hope to refine our current approach in three ways. First, we want to enhance the individual automatic reconstructions for the Burmish data and the Karenic data by taking the context of important sounds into account. Second, we want to enhance our data by correcting cases where we identified problematic cognate judgments. Third, we want to apply our method to more data from other language families in order to see how the approach performs there.

Supplementary Material

Supplementary material needed to replicate the experiments shown here, including data and code, has been curated on GitHub (https://github.com/lingpy/fuzzy/releases/tag/v1.1) and archived with Zenodo (https://doi.org/10.5281/zenodo.10007475).

Limitations

Our approach comes with some limitations. First, since the computation depends on the original data, fuzzy cognates do not only reflect true cases of uncertainty (where scholars would assess that the evidence is not enough to decide for one particular among several sounds) but can also be due to errors in the originally coded data. Second, since we use a specific procedure of grouping sounds in those cases where a proto-sound does not correspond to any sound in the descendant data,[1] our automated reconstruction approach currently may reconstruct phonotactically incorrect proto-forms. These forms may consist, for example, of two identical finals. Third, as also discussed in the study, context-dependencies which are not explicitly handled in the reconstruction procedure may yield ambiguities even in those cases, where we know they should not occur. Fourth, so far, our experiments have only been dealing with alignments that were computed automatically. Manually annotated alignments have not yet been tested.

[1] This is labelled *trimming* in List et al. 2022b, but the term does not seem a good choice, given that trimming in biology refers to cases where entire columns in an alignment are dropped, see Blum and List 2023.

Ethics Statement

Our data are taken from publicly available sources. There are no ethical issues or conflicts of interest in this work that we would know of at the time of writing.

Acknowledgements

This project was supported by the ERC Consolidator Grant ProduSemy (PI Johann-Mattis List, Grant No. 101044282, see https://doi.org/10.3030/101044282) and the Max Planck Research Grant CALC³ (https://calc.digling.org, PI Johann-Mattis List). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency (nor any other funding agencies involved). Neither the European Union nor the granting authority can be held responsible for them. We thank the anonymous reviewers for helpful comments and all people who share their data openly, so we can use it in our research.

References

William H. Baxter and Laurent Sagart. 2014. *Old Chinese. A new reconstruction*. Oxford University Press, Oxford.

Frederic Blum and Carlos Barrientos. 2023. A new dataset with phonological reconstructions in CLDF. *Computer-Assisted Language Comparison in Practice*, 6(6).

Frederic Blum and Johann-Mattis List. 2023. Trimming phonetic alignments improves the inference of sound correspondence patterns from multilingual wordlists. In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 52–64, Dubrovnik, Croatia. Association for Computational Linguistics.

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. 2013. Automatic detection of cognates using orthographic alignment. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers)*, pages 99–105.

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.

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

Llion Jones, Richard Sproat, and Haruko Ishikawa. 2022. Helpful neighbors: Leveraging geographic neighbors to aid in placename pronunciation. In preparation.

Young Min Kim, Kalvin Chang, Chenxuan Cui, and David R. Mortensen. 2023. Transformed protoform reconstruction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 24–38, Toronto, Canada. Association for Computational Linguistics.

Christo Kirov, Richard Sproat, and Alexander Gutkin. 2022. Mockingbird at the SIGTYP 2022 Shared Task: Two types of models for the prediction of cognate reflexes. In *The Fourth Workshop on Computational Typology and Multilingual NLP*, Online. Association for Computational Linguistics.

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

Johann-Mattis List. 2023. *EDICTOR. A web-based tool for creating, editing, and publishing etymological datasets [Software, Version 2.1.0]*. MCL Chair at the University of Passau, Passau.

Johann-Mattis List and Robert Forkel. 2022. Automated identification of borrowings in multilingual wordlists [version 3; peer review: 4 approved]. *Open Research Europe*, 1(79):1–11.

Johann-Mattis List and Robert Forkel. 2023a. *LingPy. A Python library for quantitative tasks in historical linguistics [Software Library, Version 2.6.11]*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Johann-Mattis List and Robert Forkel. 2023b. *LingRex: Linguistic reconstruction with LingPy [Software, Version 1.4.2]*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Johann-Mattis List, Robert Forkel, Simon J. Greenhill, Christoph Rzymski, Johannes Englisch, and Russell D. Gray. 2022a. Lexibank, a public repository of standardized wordlists with computed phonological and lexical features. *Scientific Data*, 9(316):1–31.

Johann-Mattis List, Nathan W. Hill, and Robert Forkel. 2022b. A new framework for fast automated phonological reconstruction using trimmed alignments and sound correspondence patterns. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 89–96, Dublin. Association for Computational Linguistics.

Johann-Mattis List, Ekatarina Vylomova, Robert Forkel, Nathan Hill, and Ryan D. Cotterell. 2022c. The SIGTYP shared task on the prediction of cognate reflexes. In *Proceedings of the 4th Workshop on Computational Typology and Multilingual NLP*, pages 52–62, Seattle. Association for Computational Linguistics, Max Planck Institute for Evolutionary Anthropology.

Guilin Liu, Fitsum A. Reda, Kevin J. Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. 2018. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the 15th European Conference on Computer Vision (ECCV 2018)*, pages 89–105, Munich, Germany. Springer International Publishing. Preprint.

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

Rosemarie Lühr. 2008. Von Berthold Delbrċk bis Ferdinand Sommer: Die Herausbildung der Indogermanistik in Jena [From Berthold Delbrück to Ferdinand Sommer: The Development of Indo-European Studies in Jena]. Vortrag im Rahmen einer Ringvorlesung zur Geschichte der Altertumswissenschaften (09.01.2008, FSU-Jena).

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.

Sanderson Castro Soares de Oliveira. 2014. *Contribuições para a reconstrução do Protopáno*. Ph.D. thesis, Universidade de Brasília, Brasília.

August Schleicher. 1868. Eine fabel in indogermanischer sprache. In A. Kuh and August Schleicher, editors, *Beiträge zur vergleichenden Sprachforschung auf dem Gebiete der arischen, celtischen und slawischen Sprachen*, 5, pages 206–208. Ferdinand Dümmler, Berlin.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 1–11. Curran Associates.