Towards a computer-assisted reconstruction of Proto-Burmish

Nathan W. Hill[1] and Johann-Mattis List[2]
[1]SOAS, London; [2]MPI-SHH, Jena

July, 2018

Although Lolo-Burmese benefited from a monograph length treatment already in 1979, the closely related Burmish language family has received much less attention. Burling (1967) did some preliminary work on the basis of three languages, but did not use Written or Old Burmese. Mann (1998) made an admirable efforts in his MA thesis, but the data he relied on is not consistent and his reconstructions are not easily validated. Nishi (1999) made an excellent assembly of cognates on the basis of Huáng (1992) and worked out tonal correspondences, but did not propose reconstructions.

Our reconstruction also relies, in the first instance, on Huáng (ibid.). By relying on fully digitized data we have been able to induce correspondence patterns using an algorithmic approach developed at the Max Plank Institute for the Study of Human History. This method allows us to identify inconsistencies and errors in Huang's transcription. In addition, our computer framework allows the reader to trace morphemes as they compound in different words in different languages and to explore the underlying data on which the reconstructions are based.

We hope that our work, both the methodology of computer aided reconstruction, and the style of presentation that takes full advantage of online display, will serve as an inspiration for more explicit etymological work in other areas of Sino-Tibetan.

1 Introduction

A lack of formalization in the comparative method

- Although scholars assume that the methodological framework underlying the comparative method is highly formalized, there is much discussion about the correct application of the comparative method among scholars, especially when it comes to less well-known language families such as Sino-Tibetan.

- Although most scholars agree that there is a "correct" way to apply the comparative method, many scholars also think that it is mostly impossible to formalize the method in such a way that at least parts of it could be carried out by computers.

- Scholars also often disagree with each other, and criticize one another for not properly understanding what it means to apply the comparative method.

→ We think that the problems of formalization and the lack of agreement among scholars reflect that the comparative method in its current form is more fragile than scholars typically realize. We are, however, confident, that the method can be properly formalized, especially within computer-assisted frameworks.

Embracing Neogrammarian rigor in reconstruction

- Every reconstruction will confront exceptions, but only if scholars provide transparent accounts of the exceptions they encounter (such as Grimm 1822 and Lottner 1862 did for the exceptions to Grimm's law), future generations of researchers can build on previous work and try to resolve exceptions by identifying regularity that has been previously overlooked (such as Grassmann 1863 and Verner 1877 did for Grimm's law) or by explaining word history by word history with the help of additional evidence.

- Current etymological frameworks are not sufficient to provide the rigor and transparency that is needed within a formal application of the comparative method. Too many arguments are provided in prose form, and often, major parts of the data from which scholars derive their conclusions, are not even provided.

- An ideal reconstruction would not only list all reflexes of the reconstructed form in the daughter languages, but also assess the certainty or the regularity of the form, by showing how often the underlying correspondence patterns are reflected in the given analysis, and how many alternative solutions could be proposed instead.

→ Without computational frameworks that assist in pre- and post-analyzing linguistic data, the comparative method cannot advance, due to the large amount of data and implicit decisions which cannot be readily processed even by experts in the field.

2 Preliminary Considerations

Available frameworks for computer-assisted language comparison

- Methods for automatic phonetic alignment of multiple words by now well-advanced (List 2012b; List forthcoming) and can be readily used, as implemented in the LingPy software package (List et al. 2017a) and illustrated in a recent tutorial (List et al. forthcoming).

- The same applies for methods for automatic cognate detection (List 2012a; List 2014; List et al. 2017b), which can now also handle partial cognates (List et al. 2016b).

- Methods for the analysis of language-internal cognate relations are less advanced, but initial approaches have been presented and can be readily used (Hill and List 2017). Additionally, annotation frameworks like EDICTOR (List 2017a) offer convenient ways to manually annotate language-internal relations between words, as described in Hill and List (2017).

Challenges in the reconstruction of Proto-Burmish

- Compounding is extremely frequent in the language family. As a result, scholars usually ignore compounds and do not reconstruct full lexemes back to the proto-language, but merely mono-syllabic morphemes (Mann 1998; Nishi 1999). Since it is clear that compound words were already present in the ancestral language, it is not realistic to restrict a linguistic reconstruction to morphemes. Furthermore, assembling morphemes across different languages to cognate sets is only straightforward if the original words of the compounds are traced back to their original compound word, due to secondary sound changes, like weakening or vowel shortening, which can occur frequently in compound words. As a result, current Burmish reconstructions (and many reconstructions for similar SEA languages) are not true lexical reconstructions, but rather pure phonological reconstructions with an opaque annotation of cognate sets.

- Sound correspondences are acknowledged by scholars, but are not presented in a rigorous manner. Scholars show sets of striking similarities in structure (Mann 1998; Nishi 1999), but they fail to provide us with a full account on the different correspondence *patterns* that can be found across the data.

- Data are presented in an informal way that makes it difficult to evaluate them with computer-based approaches. Often, scholars pick data from different sources, but fail to note exactly which source they picked each form from. This is even more problematic in those cases where scholars split compounds into monosyllables and use very loose elicitation glosses, which make any attempt to understand a given analysis in depth a quixotic voyage of reverse-back-engineering of the origins of the data.

Cognates, alignments, sound correspondence patterns, and proto-forms

- By cognates, we refer to *partial cognates* as presented in Hill and List (2017). That is, we assume that the current tools allow us to give a very detailed account on which part of a given word in a given language is cognate with other parts inside that very language or another language.

- By alignments we mean matrices in which the words that reflect the same cognate set are assembled in such a way that all sounds which stem from the same proto-sound are placed in the same column, with gap symbols (a dash "-") representing instances where a sound has been lost or another sound was introduced later (like epenthesis, vowel breaking, etc.).

- Correspondence patterns are defined following an approach to be discussed in detail elsewhere: All columns (also called *sites*) in the alignments of the words in a given dataset can be clustered into correspondence patterns, which reflect basic sound correspondences across a given set of languages. In contrast to traditional accounts on sound correspondences, however, our analysis does not place multiple values in the cell for a given language (e.g., due to secondary split of some proto-sound into multiple sounds in the same language), but repeat these patterns explicitly for each environment encountered. A given proto-sound may thus correspond to multiple correspondence patterns, while each correspondence pattern may

only reflect a single proto-sound. In contrast, due to gaps in our data, a given alignment column or site may well reflect more than one proto-sound, as it is sometimes difficult to infer to which pattern a given alignment site belongs. This practice of annotating uncertainties in correspondence patterns is occasionally used in historical linguistics (Baxter and Sagart 2014; Jacques 2017), but has so far never been quantified.

- Proto-forms are an attempt to explain for a given cognate set all alignment sites by proposing one proto-sound per column. This is in fact trivial, but it is not often explicitly dealt with as a subtask of an alignment in historical linguistics.

Phonological reconstruction

- Note that the reconstructions we propose at this stage resemble those of Mann (1998), Nishi (1999), and also most of the reconstructions given in the STEDT project (Matisoff 2015). In contrast to these projects, however, we trace the forms not only explicitly back to the original compound forms in the data (as it is not done by Mann and Nishi, but as it is done in STEDT), but also provide a full account on the sounds in the cognate sets by providing alignments. Our reconstruction is thus a *phonological reconstruction*, and no *lexical reconstruction*. [1]

3 Materials and Methods

Materials

Language	Source	Word Forms	Concepts	Morphemes
Atsi	Huáng et al. 1992	489	484	717
Bola	Huáng et al. 1992	496	489	702
Lashi	Huáng et al. 1992	458	453	636
Maru	Huáng et al. 1992	483	477	647
Old Burmese	Okell 1971, Luce 1985, Nishi 1999	513	481	694
Rangoon	Huáng et al. 1992	529	505	750
Xiandao	Huáng et al. 1992	459	449	634
Total		4281	506	5819

Computer-assisted workflow

1. Data cleaning. Link the data to Concepticon (List et al. 2016a) and Glottolog (Hammarström et al. 2017). Convert the phonetic transcriptions to plain IPA, following the framework by the Cross-Linguistic Transcription Systems initiative (List 2017b).

2. Morpheme annotation. Annotate and identify language-internal cognates by following the annotation framework described in (Hill and List 2017).

[1] If we wanted to carry out a lexical reconstruction, we would need methods for ancestral state reconstruction, or *onomasiological reconstruction* (see Jäger and List 2018; List 2016) applied to the partial cognates in our data, in order to infer the most likely expression for a given concept in the data. We intend to produce a lexical reconstruction of Proto-Burmish once our phonological reconstruction is fully mature.

3. Partial cognate annotation. Use the method by List et al. (2016b) to search automatically for partial cognates in the data and manually correct them with the help of the EDICTOR tool (List 2017a).

4. Structure annotation of the phonetic entries. Add basic information on the prosodic structure of the phonetic entries, following the classical division of morphemes/syllables into *initial*, *medial*, *nucleus*, *coda*, and *tone*, often used in Chinese dialectology. This annotation can be automated, but needs to be refined manually.

5. Phonetic alignment. Use a new custom method for phonetic alignment based on the previously assigned structure of each monosyllable, implemented in Python.

6. Cross-semantic cognate annotation. By searching for identical *alignments* across the data, provide first hints to the expert for cross-semantic cognates, using a custom method for this search. While the pre-processing automates some of the search, most of this analysis needs to be done manually, although the EDICTOR tool provides useful help in semi-automating parts of this task on the level of an individual language.

7. Correspondence pattern inference. Pre-analyse the data and search automatically for the most frequent correspondence patterns in the data. Once this is done, manually refine the results and make a pre-selection of all correspondence patterns which seem to be reliable, reflecting regular sound change.

8. Semi-automatic phonological reconstruction. Annotate manually all correspondence patterns with one proto-sound and then search through all alignments to assign (in decreasing order by frequency) all potential proto-sounds to a given alignment site, if this site is compatible with a given pattern that was previously selected manually. Note that this procedure can also be carried out on a small part of the data, which is annotated by the expert: when re-applying the algorithm, it can use the manually annotated data to impute the missing values for the rest of the dataset.

9. Evaluating reconstruction plausibility. Use the frequency of the inferred correspondence patterns as a proxy to assess the regularity of each semi-automatically reconstructed proto-form. For this, we set a threshold of three occurrences as a minimal requirement for a proto-sound to be regular and then check for all proto-forms, whether the reconstructed entry is indeed regarded as regular or not. We can annotate uncertainties by providing the frequency of a given proto-form in superscript numbers, and we can also annotate uncertainties in the form of the correct proto-sound for a given alignment site, if we encounter cases where a given alignment site can reflect more than one proto-form.

10. Use the language-specific morpheme annotation to provide a semantic reconstruction shortcut. Our semantic reconstructions are not to be taken at face value, but rather reflect the information we extracted from language-internal and cross-semantic inspection of the data. They should be regarded as preliminary, but they reflect in some sense the same proto-glosses which scholars provide in reconstructions like the one by Mann (1998), Nishi (1999), or the proto-forms of the STEDT project (Matisoff 2015).

4 Results

Basic statistics

- We assigned all morphemes in the data to 1685 different cognate sets, of which 1073 occur in more than one language.
- We provide 648 manually checked and aligned proto-forms for these cognate sets.
- The remaining 425 proto-forms will need to be checked further, and are currently only reflected in the form of automated reconstructions.
- We can assess the overall regularity of the different positions (initial, medial, nucleus, coda, and tone) that we reconstruct. Our findings indicate that the nucleus position is the weakest (83% of "regular" forms with patterns recurring more than twice), followed by medial (84%), initial (85%), coda (87%), and tone (93%).
- Unlike the classical reconstructions which provide fixed proto-forms with often no ways to test them against the original data, our approach allows us to display uncertainty, but also to annotate exceptions with regard to expected reflexes in the descendant languages.

Data and Code

Data and code is currently still much in flux. It can, however, be inspected on our GitHub repository at `https://github.com/digling/burmish`. Our database is available via `https://dighl.github.io/burmish`. If you want to use the data or the code described for your work, we would appreciate it if you could contact us before submitting any papers. In such a case, we would make sure to discuss the publication of a preliminary version and suggest how we prefer to be quoted.

Acknowledgments

This research would not have been possible without the LFK Young Scholars Symposium (University of Washington, Seattle, 2013), generously hosted by the Li Fang-Kuei Society for Chinese Linguistics (`http://lfksociety.org/`), during which both authors made first acquaintance and began their collaboration. We would like to acknowledge the generous support of the European Research Council for supporting this research under the auspices of 'Beyond Boundaries: Religion, Region, Language and the State' (ERC Synergy Pro- ject 609823 ASIA, NWH) and 'Computer-Assisted Language Comparison' (ERC Starting Grant, JML), and the German Research Foundation (DFG) for supporting JML from 2015 to 2016 with a research scholarship on 'Vertical and Lateral Aspects of Chinese Dialect History' (Grant No. 261553824). We would further like to thank Doug Cooper and Mark Miyake for providing invaluable help with linguistic data.

References

Baxter, W. H. and L. Sagart (2014). *Old Chinese. A new reconstruction*. Oxford: Oxford University Press.

Burling, R. (1967). *Proto-Lolo-Burmese*. Bloomington: Indiana University.

Grassmann, H. (1863). "Ueber die aspiraten und ihr gleichzeitiges vorhandensein im an- und auslaute der wurzeln". In: *Zeitschrift für vergleichende Sprachforschung auf dem Gebiete des Deutschen, Griechischen und Lateinischen* 12 (2), pp. 81–110.

Grimm, J. (1822). *Deutsche Grammatik*. 2nd ed. Vol. 1. Göttingen: Dieterichsche Buchhandlung. Google Books: MnsKAAAAIAAJ.

Hammarström, H., R. Forkel, and M. Haspelmath (2017). *Glottolog*. Version 3.0. URL: http://glottolog.org.

Hill, N. W. and J.-M. List (2017). "Challenges of annotation and analysis in computer-assisted language comparison: A case study on Burmish languages". In: *Yearbook of the Poznań Linguistic Meeting* 3.1, 47–76.

Huáng, B. ◆., ed. (1992). *Zàngmiǎn yǔzú yǔyán cíhuì*. Běijīng 北京: Zhōngyāng Mínzú Dàxué 中央民族大学 [Central Institute of Minorities]. Digital Version: https://stedt.berkeley.edu/~stedt-cgi/rootcanal.pl/source/TBL.

Jacques, G. (2017). "A reconstruction of Proto-Kiranti verb roots". In: *Folia Linguistica Historica* 38.1, pp. 177–215.

Jäger, G. and J.-M. List (2018). "Using ancestral state reconstruction methods for onomasiological reconstruction in multilingual word lists". In: *Language Dynamics and Change* 8.1, pp. 22–54.

List, J.-M. (2012a). "LexStat. Automatic detection of cognates in multilingual wordlists". In: *Proceedings of the EACL 2012 Joint Workshop of Visualization of Linguistic Patterns and Uncovering Language History from Multilingual Resources*. LINGVIS & UNCLH 2012. (Avignon, 04/23–04/24/2012). Stroudsburg, pp. 117–125.

– (2012b). "Multiple sequence alignment in historical linguistics. A sound class based approach". In: *Proceedings of ConSOLE XIX*. The 19th Conference of the Student Organization of Linguistics in Europe. (Groningen, 01/05–01/08/2011). Ed. by E. Boone, K. Linke, and M. Schulpen, pp. 241–260. PDF: http://media.leidenuniv.nl/legacy/console19-proceedings-list.pdf.

– (2014). *Sequence comparison in historical linguistics*. Düsseldorf: Düsseldorf University Press.

– (2016). "Beyond cognacy: Historical relations between words and their implication for phylogenetic reconstruction". In: *Journal of Language Evolution* 1.2, pp. 119–136.

– (2017a). "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*. Valencia: Association for Computational Linguistics, pp. 9–12.

– (2017b). *Establishing a cross-linguistic database of phonetic notation systems*. paperworkshop.

– (forthcoming). "SCA: Phonetic alignment based on sound classes". In: *New directions in logic, language, and computation*. Ed. by M. Slavkovik and D. Lassiter. Berlin and Heidelberg: Springer, pp. 32–51.

List, J.-M., M. Cysouw, and R. Forkel (2016a). "Concepticon. A resource for the linking of concept lists". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation*. LREC 2016. (Portorož, 05/23–05/28/2016). Ed. by N. C. C. Chair), K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis. European Language Resources Association (ELRA), pp. 2393–2400.

List, J.-M., P. Lopez, and E. Bapteste (2016b). *Using sequence similarity networks to identify partial cognates in multilingual wordlists*. paperconference. Berlin: Association of Computational Linguistics.

List, J.-M., S. Greenhill, and R. Forkel (2017a). *LingPy. A Python library for quantitative tasks in historical linguistics*. Jena: Max Planck Institute for the Science of Human History.

List, J.-M., S. J. Greenhill, and R. D. Gray (01/2017b). "The potential of automatic word comparison for historical linguistics". In: *PLOS ONE* 12.1, pp. 1–18.

List, J.-M., M. Walworth, S. J. Greenhill, T. Tresoldi, and R. Forkel (forthcoming). "Sequence comparison in computational historical linguistics. Phonetic alignments and cognate detection with LingPy 2.6". In: *Journal of Language Evolution* ??, ??–??

Lottner, C. (1862). "Ausnahmen der ersten lautverschiebung". In: *Zeitschrift für vergleichende Sprachforschung auf dem Gebiete des Deutschen, Griechischen und Lateinischen* 11.3/4, pp. 161–205.

Luce, G. H. (1985). *Phases of Pre-Pagán Burma: Languages and history*. Oxford: Oxford University Press.

Mann, N. W. (1998). "A phonological reconstruction of Proto Northern Burmic". MA. Arlington: The University of Texas.

Matisoff, J. A. (2015). *The Sino-Tibetan Etymological Dictionary and Thesaurus project*. Berkeley: University of California.

Nishi, Y. (1999). *Four papers on Burmese: Toward the history of Burmese (the Myanmar language)*. Tokyo: Institute for the study of languages, cultures of Asia, and Africa, Tokyo University of Foreign Studies.

Okell, J. (1971). “K Clusters in Proto-Burmese”. In: *Papers presented at the Sino-Tibetan Conference*. (Bloomington, 10/08–10/09/1971). Bloomington.

Verner, K. A. (1877). “Eine Ausnahme der ersten Lautverschiebung”. In: *Zeitschrift für vergleichende Sprachforschung auf dem Gebiete der Indogermanischen Sprachen* 23.2, pp. 97–130.