Beyond Cognacy: Challenges of Representing and Analyzing Etymological Data of South-East Asian Languages

Nathan Hill and Johann-Mattis List

This paper explores a dataset of Burmish languages as a case study in improving the methodology of computational reconstruction. The use of computational methods in comparative linguistics is ever increasing in popularity. The increasing deployment of such methods draws into focus those areas in which they are still inadequate. In particular, effectively all previous research treats cognacy as a binary phenomenon, but this heuristic is not realistic and leads to over simplified analyses. For example, Daniel Klein, says to the BBC regarding the system of computational reconstruction described in Bouchard-Côté et al. (2013), which he is a co-author of, the "system still has shortcomings", citing as an example that "it can't handle morphological changes or re-duplications - how a word like 'cat' becomes 'kitty-cat'" (Morelle 2013). In addition to morphological compounding, other areas that computational reconstruction models have so far not tackled include the modeling of tonogenesis and the origin of other supersegmentals and the stratification of borrowings and inherited forms.

For the purpose of this abstract a single example well illustrates some of the issues at play. The Burmese word for 'heart' is *nha-lu* and the Atsi word for 'heart' is *ṇik*. The first syllable of the Burmese word is cognate to the Atsi word, but the second syllable has no equivalent in the Atsi cognate. Any binary model of cognacy will fail to capture this relationship. We have developed an approach to cognacy that operates at the morpheme rather than at the word level. The orthographic aspiration of the nasal nh- (realized as [n] in the Rangoon dialect of today) corresponds to the creaky phonation of Atsi. This correspondences poses both an obstacle of data representation and an analytic obstacle. Reconstruction algorithms typically (and understandably) operate on IPA characters rather than orthographic transcriptions, but the phonetic interpretation of ancient scripts is insecure and frequently contested. We must decide whether to represent nh- as [n], [n̥], or some other option. No matter how it is represented, currently available algorithms fail to pick up on such relationships between initial manner and vowel phonation type. It is only by testing reconstruction systems on different language families that computational system can be made more robust.

Our project relies on a database of 250 concepts (liked to the *Concepticon*, cf. List et al. 2016) as expressed in a dozen Burmish languages. The primary data comes from Huáng et al. (1992.), as digitized by the STEDT project, but we supplement this with other relevant sources. We employ an iterative workflow combining the absolute rigor of a computer with the insightful intuitions of trained historical linguistics. After providing all of the data with unambiguous phonetic interpretations, including the explicit encoding of underdetermined segments, the computer provides a preliminary alignment and reconstruction. These reconstructions are then adjusted with an eye to the relevant literature on proto-Burmish (Nishi 1999, Dempsey 2003, Hill 2013). The adjustments are made inside of the workflow system so that the algorithm and general methodology will be enhanced and made more robust.

References

Bouchard-Côté, Alexandre, 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 110.11: 4224-4229.

Dempsey, Jakob (2003). 'Analysis of rime-groups in Northern-Burmish.' *Linguistics of the Tibeto-Burman Area* 26.1: 63-124.

Hill, Nathan W. (2013) 'The merger of Proto-Burmish *ts and *č in Burmese.' *SOAS Working Papers in Linguistics* 16: 334-345.

Huáng Bùfán 黄 布 凡 et al. eds. (1992). Zàng-Miǎn yǔzú yǔyán cíhuì 藏缅语族语言词汇. Běijīng: Zhōngyāng mínzú xuéyuàn chūbǎnshè 中央民族学院出版社.

List, Johann-Mattis & Cysouw, Michael & Forkel, Robert (eds.) 2015. Concepticon. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://concepticon.clld.org, Accessed on 2016-03-15.)

Morelle, Rebecca (2013). "Ancient languages reconstructed by computer program." BBC News. 12 February 2013. http://www.bbc.com/news/science-environment-21427896

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