January 20th, 2023

Automating Comparative Reconstructions: Case Study in Austronesian and Ongan

Although comparative reconstruction has always been one of the key endeavors of linguistics, there exists no widely accepted method for evaluating its applications (Michalove, 1998). Instead, evaluation is conducted through debate, often spanning decades, as in the case of Altaic, Nostratic, and, more recently, Dene-Yeniseian. Previous attempts to introduce quantitative measures for genetic relatedness are heuristics for estimating similarity, usually either by calculating the average phonetic distance for each putative word-pair (Downey et al., 2008; Kondrak, 2003) or by computing the proportion of cognates between the two wordlists (Chang et al., 2015; Atkinson & Gray, 2003). Since none of these previous attempts engage with diachronic change directly, most researchers agree that, while they are useful when manual reconstruction is not feasible, traditional methods are still the gold standard (Kiparsky, 2015).

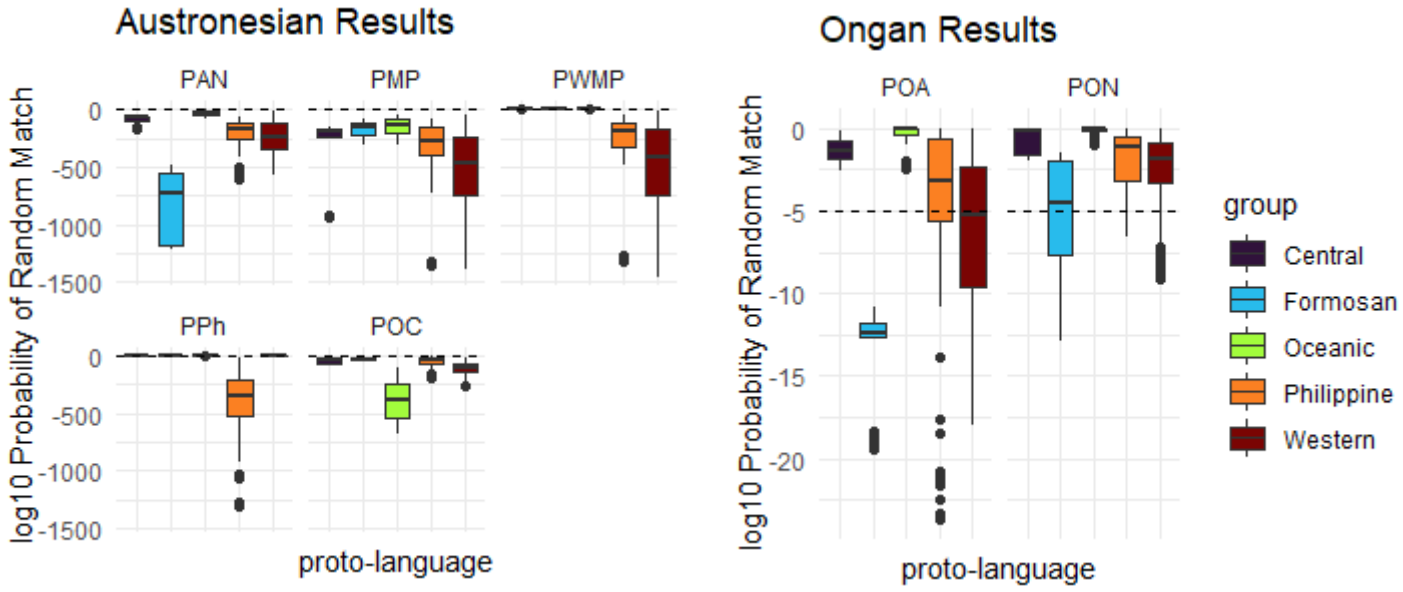
I present a probabilistic framework for evaluating comparative reconstruction attempts. The series of transformations – sound changes, borrowing, semantic change, etc – serves as the input to the framework's evaluation function. The output is the estimated probability that a randomly generated wordlist merits a reconstruction from the mother language using the same number of transformations or fewer than required by the daughter language. Thus, the framework evaluates reconstruction attempts themselves rather than the original dataset, setting it apart from previous quantitative measures.

The framework was incorporated into a simulated annealing learning algorithm, where reconstructions from a mother wordlist to a daughter wordlist were suggested stochastically with a bias toward decreasing the probability of a random match. The algorithm was tested on a genetically diverse sample of Austronesian languages and 5 Austronesian proto-languages. Figure 1 presents the probability of a random match in automated reconstructions from the proto-languages to the 5 Austronesian groups tested, as defined in the Comparative Austronesian Dictionary (Trussel & Blust, 2010). The results are in line with general knowledge in the Austronesian field. For the 237 comparisons between an Austronesian proto-language and a direct descendant, the algorithm always found a reconstruction with a probability of a random below the chosen cut-off of reliability at .0001. The probability of a random match appears to be strongly correlated with the time depth of the reconstruction.

The case study was further extended to evaluate the putative Ongan-Austronesian connection (Blevins, 2007), a hypothesis not generally accepted in the field (Blust, 2014). Figure 2 presents the probability of a random match in reconstructions from proto-Ongan and proto-Ongan-Austronesian to the 5 Austronesian groups. In reconstructions from proto-Ongan-Austronesian to the Austronesian languages, the results are mixed with the algorithm finding probabilistically non-arbitrary reconstructions to 26 of the 74 of the Austronesian languages tested. Reconstructions from proto-Ongan-Austronesian to the Ongan languages are similarly mixed, with some extremely convincing and others not at all. In general, the results with respect to the Ongan-Austronesian hypothesis appear promising, but not conclusive.

This research is meant to introduce a framework for objective debate surrounding comparative reconstructions and controversial language groupings. The framework can also be used to reason about the comparative method more broadly. For example, the results of the case study reveal that the probability of a random match is mostly determined by the number of borrowings posited, as well the phonotactic complexity of the daughter language. The effect of individual sound changes on reconstruction arbitrariness is measurable but comparatively minor. Future implementations of the framework can be extended to other types of diachronic transformation, e.g. semantic change, morphological change, etc.

January 20th, 2023



Figures 1 & 2 : The log probability that a randomly generated wordlist merits a reconstruction of the same size or smaller than the one generated automatically by a simulated annealing algorithm for 74 Austronesian languages and 5 widely accepted Austronesian proto-languages (Figure 1) and 2 putative Ongan proto-languages (Figure 2). PAN = proto-Austronesian; PMP = proto-Malayo-Polynesian; PWMP = proto-West-Malayo-Polynesian; PPh = proto-Philippine; POC = proto-Oceanic; POA = proto-Ongan-Austronesian; PON = proto-Ongan.

References:

Atkinson, Q., & Gray, R. (2005). Curious Parallels and Curious Connections—Phylogenetic Thinking in Biology and Historical Linguistics. Systematic Biology, 54(4), 513-526.

Blevins, Juliette (2007). A long lost sister of proto-Austronesian? Proto-Ongan, mother of Jarawa and Onge of the Andaman Islands. *Oceanic Linguistics*, 46(1), 155-198.

Blust, Robert (2014). Some recent proposals concerning the classification of the Austronesian languages. *Oceanic Linguistics*, 53(2), 300-391.

Chang, W., Cathcart, C., Hall, D., & Garrett, A. (2015). Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. Language, 91(1), 194-244.

Downey, S., Hallmark, B., Cox, M., Norquest, P., & Lansing, J. (2008). Computational Feature-Sensitive Reconstruction of Language Relationships: Developing the ALINE Distance for Comparative Historical Linguistic Reconstruction. Journal of Quantitative Linguistics, 15(4), 340-369.

Kiparsky, P. (2015). New perspectives in historical linguistics. The Routledge Handbook of Historical Linguistics.

Kondrak, G. (2003). Phonetic Alignment and Similarity. Computers and the Humanities, 37(3), 273-291.

Michalove, P., Georg, S., & Ramer, A. (1998). Current Issues In Linguistic Taxonomy. Annual Review of Anthropology, 27(1), 451-472.

Trussel, Stepehen & Robert Blust (2010). *Austronesian Comparative Dictionary*. https://www.trussel2.com/ACD/introduction.htm