Using sound correspondences in Bayesian phylolinguistics: a case study on Japonic

John Huisman[1], Bonnie McLean[1], Chieh-Hsi Wu[2], Tiago Tresoldi[1]

[1]Uppsala University, [2]University of Southampton

Keywords: sound correspondences, phylogenetic analysis, computational historical linguistics, Japonic

The Japonic language family has received considerable scholarship, but many questions surrounding its history remain unanswered. Throughout its long history, Japanese dialectology has meticulously mapped the geographical distribution of synchronic variation and while this has led to a well-established classification of dialects into various subgroups, the relationship between these subgroups from a historical perspective has received less attention. More recent work focusing on this diachronic dimension of the Japonic languages has produced more detailed phylogenetic classifications [1-3], but unraveling the relations between the closely related varieties that have developed from a dialect chain on the Japanese mainland continues to pose a challenge. For example, the application of Bayesian phylogenetic methods [2] produced results with lower overall support than what has been found in other language families.

The advances made by computational historical linguistics in recent decades are largely built on cognacy data [4-7]. The complexity of Bayesian analyses requires large amounts of data, which has been argued to be "only really available from cognacy in the lexicon" [4]. This approach works well for many major language families, as their time-depths reach into the thousands of years, providing ample opportunity for lexical innovations even in basic vocabulary. However, in contexts of recent diversification, areal features, and regular contact, such as the situation found for the Japanese mainland varieties, accurately inferring tree structure from cognate data alone can be difficult as there is less differentiation. To address this issue, we experiment with analyses that, unlike the prevalent approach in computational historical linguistics, use features other than cognate sets as phylogenetic characters.

Specifically, we examine the use of a new method based on synchronic patterns of sound correspondences, which are used as a proxy for sound changes. The identification of regular sound correspondences through the systematic comparison of cognate sets forms the cornerstone of the comparative method, and here we test their value in computational methods using the Japonic languages as a case study. In contrast to previous work based on pairwise comparisons [8-11], the method presented here uses all systematic correspondences found across sets of multiple languages—rather than selecting only a subset of changes, e.g., [12]. Given the time and expertise needed to code these correspondences in the ever-growing wealth of openly available linguistic data, we also test the accuracy of algorithmically inferred correspondences [13], as well as character n-grams found in lexemes as a proxy for phonotactics, following the approach of [14]. We analyse the sound correspondence data using unrooted phylogenetic networks [15], and infer trees using Maximum-Likelihood [16] and Bayesian [17] approaches, with different tree priors and models of evolution. While theoretically more liable to homoplasy than cognacy [18-20], our findings suggest that the use of sound correspondences provides enough phylogenetic signal to be recovered by statistical analyses. A comparison between our results and previously suggested phylogenies of the Japonic language family [1-3] shows that the method accurately captures language relationships. We also discuss potential improvement to the method, showing how the integration of sound correspondences into analyses can help untangle fine-grained phylogenetic structure in contexts such as that of Japonic.

References

[1] Pellard, T. (2009). *Ōgami: Éléments de description d'un parler du Sud des Ryūkyū* (Doctoral dissertation, Ecole des Hautes Etudes en Sciences Sociales (EHESS)).

[2] Lee, S., & Hasegawa, T. (2011). Bayesian phylogenetic analysis supports an agricultural origin of Japonic languages. *Proceedings of the Royal Society B: Biological Sciences, 278*(1725), 3662-3669.

[3] de Boer, E. The classification of the Japonic languages. In *The Oxford Guide to the Transeurasian Languages* (pp. 39-58). Oxford University Press.

[4] Greenhill, S. J., Heggarty, P., & Gray, R. D. (2020). Bayesian phylolinguistics. In *The handbook of historical linguistics*, 2, 226-253.

[5] List, J. M., Walworth, M., Greenhill, S. J., Tresoldi, T., & Forkel, R. (2018). Sequence comparison in computational historical linguistics. *Journal of Language Evolution, 3*(2), 130-144.

[6] Jäger, G. (2019). Computational historical linguistics. *Theoretical Linguistics, 45*(3-4), 151-182.

[7] Hoffmann, K., Bouckaert, R., Greenhill, S. J., & Kühnert, D. (2021). Bayesian phylogenetic analysis of linguistic data using BEAST. *Journal of Language Evolution, 6*(2), 119-135.

[8] Grimes, J. E., & Agard, F. B. (1959). Linguistic divergence in Romance. *Language, 35*(4), 598-604.

[9] Hoenigswald, Henry Max. 1960. *Language Change and Linguistic Reconstruction*, 4. aufl. 1966 edition. The University of Chicago Press, Chicago.

[10] Kondrak, G. (2009). Identification of Cognates and Recurrent Sound Correspondences in Word Lists. *Traitement Automatique des Langues, 50*(2), 201-235.

[11] List, J. M. (2013). *Sequence Comparison in Historical Linguistics* (Doctoral dissertation, Heinrich Heine Universität Düsseldorf).

[12] Ringe, D., Warnow, T., & Taylor, A. (2002). Indo-European and computational cladistics. *Transactions of the Philological Society, 100*(1), 59-129.

[13] List, J. M. (2019). Automatic inference of sound correspondence patterns across multiple languages. *Computational Linguistics, 45*(1), 137-161.

[14] Macklin-Cordes, J. L., Bowern, C., & Round, E. R. (2021). Phylogenetic signal in phonotactics. *Diachronica*.

[15] Huson, D. H., & Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution, 23*(2), 254-267.

[16] Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics, 30*(9), 1312-1313.

[17] Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C. H., Xie, D., Suchard M.A., Rambaut, A., & Drummond, A. J. (2014). BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Computational Biology, 10*(4), e1003537.

[18] Chacon, T. C., & List, J. M. (2016). Improved computational models of sound change shed light on the history of the Tukanoan languages. *Journal of Language Relationship, 13*(3-4), 177-204.

[19] Kassian, A. S. (2017). Linguistic homoplasy and phylogeny reconstruction. The cases of Lezgian and Tsezic languages (North Caucasus). *Folia Linguistica, 51*(s38), 217-262.

[20] Hruschka, D. J., Branford, S., Smith, E. D., Wilkins, J., Meade, A., Pagel, M., & Bhattacharya, T. (2015). Detecting regular sound changes in linguistics as events of concerted evolution. *Current Biology, 25*(1), 1-9.