PHYLOGENETIC RECONSTRUCTION FOR JAPONIC: NEW DATA TYPES AND METHODOLOGICAL INSIGHTS

JOHN L. A. HUISMAN[*1], BONNIE MCLEAN[1], and CHIEH-HSI WU[2]

[*]Corresponding Author: john.huisman@lingfil.uu.se
[1]Department of Linguistics and Philology, Uppsala University, Uppsala, Sweden
[2]Mathematical Sciences, University of Southampton, Southampton, UK

1. Background

While synchronic variation in the Japonic languages is well understood from the meticulous work of Japanese dialectology, the relationships between the various linguistic groups from a historical perspective have received considerably less attention. Existing studies tend to overrepresent either of the two main branches of the family—i.e. Japanese (Lee & Hasegawa, 2011) and Ryukyuan (Pellard, 2009, 2015)—leaving some questions about the full family history unanswered. Bayesian phylogenetic analyses have the potential to recover these deeper-level branching events, but the complexity of such analyses requires large amounts of data, which have been argued to only be "really available from cognacy in the lexicon" (Greenhill et al., 2020). This can pose a challenge in contexts of recent diversification, areal features, and regular contact where lexical diversity is limited—e.g., the situation found on the Japanese mainland (Hattori, 1973).

2. The current study

To address these issues, we analysed a sample of languages that more effectively captures linguistic diversity of both the Japanese mainland and the Ryukyuan islands, and experiment with analyses that—unlike the prevalent approach in computational historical linguistics—include features other than cognate sets as phylogenetic characters. Specifically, we use phonotactic traits as phylogenetic characters using the method described in Macklin-Cordes et al. (2021), and assess their informativeness for language families with a relatively limited time-depth.

2.1. Data and methods

We started by collating basic vocabulary data for 65 Japanese and 28 Ryukyuan varieties using a 250-item list compiled from other commonly used concept lists (Greenhill et al., 2008; Swadesh, 1952; Tadmor, 2009) and coded the lexemes for cognacy. Subsequently, we used the vocabulary data to extract phonotactic traits

using the approach set out by Macklin-Cordes and colleagues (2021), i.e., the presence/absence of sequences of sound segments, the relative transition frequencies between sound segments, and the relative transition frequencies between sound classes. We inferred phylogenetic trees using Bayesian approaches (Bouckaert et al., 2019) with different tree priors and models of evolution in three analyses (lexical traits only, phonotactic traits only, and lexical and phonotactic traits combined), and compare our findings to previously suggested phylogenies using different methods.

2.2. Results

As expected, we found that reduced levels of lexical diversity resulted in low clade support, particularly among the Japanese mainland varieties, in the analysis based on lexical traits alone. In the analysis on phonotactic traits alone, we found several (albeit minor) conflicts with generally accepted phylogenies for the Ryukyuan branch—which, we discuss, likely result from parallel changes in the phoneme inventory that reverberate in phonotactics. However, when combining the lexical and phonotactic traits into a single holistic analysis, the model was able to both capture language relationships accurately, while also considerably decreasing the overall uncertainty in the model. We provide an overview of the strengths of the method used here (e.g., "free" algorithmically inferred data), as well as the challenges that remain with potential solutions, e.g., testing against simulated data (Wichmann & Rama 2021). We also discuss how our results shed light on ongoing questions in Japanese historical linguistics (e.g., the overall tree topology; the timing of initial diversification), and previously raised concerns about how the spatial structure of the Japonic language family complicates phylogenetic inference and its interpretation (e.g., Murawaki, 2015).

3. Summary

We present a case study of the use of phonotactic traits in computational historical linguistics (following Macklin-Cordes et al. 2021). Our findings suggest that non-lexical data can be a valuable addition to analyses that aim to untangle fine-grained phylogenetic structures in contexts of recent diversification and language contact. The reconstructed dated phylogeny of the Japonic language family provides new insights into the linguistic history of Japan, which can e.g., further our understanding of geographic factors in language diversification (cf. Huisman et al., 2019), or facilitate assessing the validity of predictive mathematical models of language change (cf. Takahashi & Ihara, 2020).

References

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., ... & Drummond, A. J. (2019). BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Computational Biology, 15*(4), e1006650.

Greenhill, S. J., Blust, R., & Gray, R. D. (2008). The Austronesian basic vocabulary database: From bioinformatics to lexomics. *Evolutionary Bioinformatics*, *4*, EBO–S893.

Greenhill, S. J., Heggarty, P., & Gray, R. D. (2020). Bayesian phylolinguistics. *The Handbook of Historical Linguistics*, *2*, 226–253.

Hattori, S. (1973). Japanese dialects. In H. Hoenigswald (Ed.), *Diachronic, areal and typological linguistics* (pp. 368–400). De Gruyter Mouton.

Huisman, J. L. A., Majid, A., & Van Hout, R. (2019). The geographical configuration of a language area influences linguistic diversity. *PLOS ONE*, *14*(6), e0217363.

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.

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

Murawaki, Y. (2015). Spatial structure of evolutionary models of dialects in contact. *Plos One*, *10*(7), e0134335.

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

Pellard, T. (2015). The linguistic archeology of the Ryukyu Islands. In P. Heinrich, S. Miyara, & M. Shimoji (Eds.), *Handbook of the Ryukyuan Languages. History, Structure, and Use* (pp. 13–38). De Gruyter Mouton.

Swadesh, M. (1952). Lexico-Statistic Dating of Prehistoric Ethnic Contacts: With Special Reference to North American Indians and Eskimos. *Proceedings of the American Philosophical Society*, *96*(4), 452–463.

Tadmor, U. (2009). Loanwords in the world's languages: Findings and results. *Loanwords in the World's Languages: A Comparative Handbook*, *55*, 75.

Takahashi, T., & Ihara, Y. (2020). Quantifying the spatial pattern of dialect words spreading from a central population. *Journal of The Royal Society Interface*, *17*(168), 20200335.

Wichmann, S., & Rama, T. (2021). Testing methods of linguistic homeland detection using synthetic data. *Philosophical Transactions of the Royal Society B: Biological Sciences, 376*(1824), 20200202.