

Combined lexical and phonotactic data resolve uncertainties in the evolutionary diversification of the Japonic language family

John L.A. Huisman^{1,*} , Bonnie McLean¹, Chieh-Hsi Wu²

¹Department of Linguistics and Philology, Uppsala University, Box 635, 751 26 Uppsala, Sweden

²School of Mathematical Sciences, University of Southampton, Highfield Campus, University Road, Southampton, SO17 1BJ, United Kingdom

*Corresponding author. John L.A. Huisman, Department of Linguistics and Philology, Uppsala University, Box 635, 751 26 Uppsala, Sweden. E-mail: john.huisman@lingfil.uu.se

Associate Editor: Dan Dediu

The use of phylogenetic methods in linguistics has provided new insights into the structure, age, and spread of language families. Despite increasing recognition of Japonic as one of the world's primary language families, research on the family's phylogeny remains limited. This study presents a new reconstruction of Japonic language history based on *NichiRyuuLex*, a new lexical dataset comprising data from 48 Japanese and 33 Ryukyuan lects for 256 concepts. The study combines lexical and phonotactic data to increase precision in the phylogenetic parameter estimates, providing a more informative reconstruction. The analyses presented here confirm previous findings on the age of the family as a whole, estimating a Japanese-Ryukyuan split at around 400–500 BCE, supporting that the time of diversification coincides with the influx of Bronze Age rice agriculturists during the Yayoi Period. The topology of the mainland Japanese clade uncovered in the analyses unifies two divisions recognized in traditional Japanese dialectology: the East-West division, and the center-periphery division. The topology of the Ryukyuan clade largely followed geographical segmentation—Northern Ryukyuan (Amami; Okinawa) vs. Southern Ryukyuan (Miyako; Macro-Yaeyama). The ancestor of Ryukyuan was dated to around the ninth century, coinciding with the first traces of cereal farming in the Northern Ryukyu Islands. In sum, the study provides greater certainty about the linguistic history and internal structure of mainland Japanese, as well as a more detailed perspective on the history of the Ryukyuan languages.

Keywords: Japonic languages; linguistic phylogenetics; historical linguistics; lexical change; phonotactics.

1. Background

The use of phylogenetic methods to analyze linguistic data has provided new insights into the structure, age, and spread of language families (e.g. Gray et al. 2009; Honkola et al. 2013; Chang et al. 2015; Bouckaert et al. 2018; Koile et al. 2022; Ferraz Gerardi et al. 2023; Hegarty et al. 2023), which in turn allows us to study general patterns of language evolution (e.g. Jordan et al. 2009; Dunn et al. 2011; Guillon and Mace 2016; Haynie and Bower 2016; Greenhill et al. 2017; Beyer et al. 2019; Huisman et al. 2019; Shcherbakova and Allasonnière-Tang 2024).¹

¹Recent years have seen a steady increase in the range of language families and topics studied using these methods. For a full up to date overview, we recommend the online database curated by Simon Greenhill: <https://simon.net.nz/phylogenies/>.

Research on the phylogeny of Japonic languages is limited (Lee and Hasegawa 2011), as Japanese was long seen as an isolate. It is now recognized as one member of the small Japonic language family, which has generated renewed and broader interest in questions on its history (Pellard 2015, 2021; de Boer et al. 2020; Jarosz et al. 2022; Igarashi 2023a, 2023b; Takahashi et al. 2023; de Boer 2024). This study combines lexical and phonotactic data to produce a new reconstruction of the Japonic language phylogeny, which provides estimates with greater precision and sheds light on hitherto unresolved facets of the history of Japonic languages.

1.1 The Japonic language family

The Japonic languages spoken across the 3,000 km long Japanese archipelago form a fairly small family

not demonstrably related to any others—see [Tian et al. \(2022\)](#) contra [Robbeets et al. \(2021\)](#) but also [Vovin \(2011, 2017\)](#), and [Janhunen \(2023\)](#). The Japonic languages are thought to have been brought to the archipelago by Bronze Age rice agriculturists who migrated there in the first millennium BCE (the Yayoi period; [Hanihara 1991](#); [Hudson et al. 2020](#)). The exact number of ‘languages’ is not established, but reports on mutual intelligibility suggest at least a dozen ([Smith 1960](#); [Yamagiwa 1967](#); [Tominaga 1988](#); [Yamada et al. 2020](#)), and further work might even uncover a number closer to 50 (see [Takubo 2018](#)). The Ryukyuan and Hachijo lects are recognized as endangered *languages* ([Moseley 2010](#)), but many other ‘dialects’ are equally under pressure from Standard Japanese.

Dialect classifications based on geographical variation abound ([Tojo 1927, 1951](#); [Kindaichi 1955](#); [Fujiwara 1962](#)), with general consensus on a division between a Japanese branch (spoken across the four main islands), and a Ryukyuan branch (spoken across the Ryukyu Islands in the south of Japan). The historical relationships between the dialect subgroups have traditionally received considerably less attention—although see e.g. [Hattori \(1978–1979\)](#) and [Peng and Peng \(1990\)](#) for examples of traditional historical linguistics studies. Currently, there are only two extensive dated phylogenies of Japonic, both presented by [Lee and Hasegawa \(2011\)](#). Their main analysis shows a primary split between Japanese and Ryukyuan at around 200 BCE. A second tree, based on a smaller but independent dataset, presented in their supplementary materials, estimates this split at the start of the first century CE. As their aim was to investigate whether Japonic spread together with agriculture, they mainly focused on the root age without extensively discussing the internal structure of the family.

1.2 Contribution of this study

We use lexical and phonotactic data to provide a more informative reconstruction of the Japonic phylogeny, while accommodating the differences in the evolutionary process between the two data types. In evaluating the results reported by [Lee and Hasegawa \(2011\)](#), we focus on three main aspects: 1, the main split at the root, and its timing; 2, the timing and topology of the mainland Japanese branch; and 3, more extensive representation of Ryukyuan branch. In addition, our new dataset also remedies minor issues with data coding and availability.

First, while [Lee and Hasegawa \(2011\)](#) find a split between Japanese and Ryukyuan during the Yayoi period (1,000 BCE—third century CE), the reported 95 per cent HPD-intervals range from 2,000 BCE to 800 CE. This covers 1,000 years of the preceding non-Japonic

hunter-gatherer Jomon period (13,500–300 BCE) and does not rule out the competing hypothesis that Japanese and Ryukyuan split after the rise of centralized states in the Kofun period (third to sixth centuries CE)—as suggested by earlier glottochronological work ([Hattori 1954, 1976](#)) and the recent historical linguistics work by [Pellard \(2015, 2021\)](#). Both cases are problematic for the language-with-farming hypothesis. In addition, another stream of recent research argues that the main split is instead between a Ryukyu-Kyushu clade versus the remaining mainland lects, based on shared non-basic vocabulary, and a shared sound change in one specific group of verbs ([Jarosz 2019](#); [De Boer 2020](#); [Igarashi 2023b, 2023a](#); [Jarosz and Orlandi 2023](#)).

Second, the main tree in [Lee and Hasegawa \(2011\)](#) estimates the ancestor of the contemporary mainland Japanese lects to date from the 17th century, whereas the supplementary tree estimates this ancestor as dating from the 12th century. This discrepancy requires further investigation, as does the timing in general, given the attestation of linguistic variation in the earliest writings—Old Japanese (eighth century); see e.g. [Vovin and Ishisaki-Vovon \(2021\)](#) and [Kupchik \(2023\)](#). Moreover, the internal structure of the Japanese branch shows higher levels of uncertainty than generally reported in phylolinguistic studies, with posterior probabilities of 0.50 for the mainland Japanese clade and 0.55 for the—broadly recognized—Eastern Japanese subgroup, for example. Many lower internal nodes showed posterior probabilities lower than 0.50. This likely results from the limited lexical differentiation inherent to its fairly shallow time-depth.

Finally, the Ryukyuan clade is underrepresented in the [Lee and Hasegawa \(2011\)](#) study, comprising only 10 Ryukyuan lects in total, with just one each from the Amami- and Okinawa subgroups. While the Ryukyuan lects are generally understudied, there are compendia of comparative data with sufficient basic vocabulary coverage for computational work.

This study aims to rectify the issues raised above through analyses of a new comparative dataset, comprising both lexical and phonotactic data, with expanded coverage of Ryukyuan. Therefore, the new dataset is expected to provide estimates with greater precision and address some of the unresolved questions in the [Lee and Hasegawa \(2011\)](#) study regarding the competing hypotheses.

2. Materials and methods

2.1 Lexical data

We compiled a new lexical dataset, *NichiRyuuLex*, using an expanded concept list that consolidates the

Austronesian Basic Vocabulary Database list (Greenhill et al. 2008, as used by Lee and Hasegawa 2011), the 100-item and 200-item Swadesh lists (Swadesh 1952, 1955), and the Leipzig-Jakarta List (Tadmor 2009). The combined list comprised 266 concepts once cleaned, but was further adapted to better fit the Japonic lexicon, resulting in a total of 256 concepts—see Supplementary Materials for details. We collated data for 48 Japanese (including Old Japanese) and 33 Ryukyuan lects using several compendia (Hirayama 1966, 1967, 1992–1994; Arakaki 2000). This geographically balanced sample is agnostic towards ‘language’ status, and instead represents each of the commonly recognized major dialect subgroups through multiple lects. Lects appearing in multiple sources helped cross-referencing transcription systems and unify all data into a single IPA-based transcription. All entries were coded for cognacy based on regular sound correspondences and reconstructed sound changes discussed in Hattori (1978–1979); Hirayama (1966; 1967; 1992–1994), Itoyo et al. (1982–1984), Nakamoto (1976; 1981), and Thorpe (1983). Clearly identifiable intra-family borrowings were excluded (e.g. Standard Japanese-like forms not showing any of the expected sound correspondences). The coding was binarised for the phylogenetic analyses.

2.2 Phonotactic data

The complexity of the Bayesian approach, estimating many parameters in a single analysis, requires larger amounts of data, and so the prevalent approach in phylolinguistics has been to use cognacy data, as these amounts have been argued to be ‘only really available in the lexicon’ (Greenhill et al. 2020, p. 236). However, the fairly shallow time-depth of Japonic entails limited lexical differentiation (e.g. Hattori 1973), which complicates unravelling the relations between closely-related lects. To address this, we supplemented the lexical data with phonotactic data, which has been used as a source of phylogenetic signal recently (Dockum 2017; Macklin-Cordes et al. 2021). Macklin-Cordes et al. (2021) was the first to combine lexical and phonotactic data in a single phylogenetic analysis, but did not find that adding the phonotactic characters improved the model. However, his study was on the Western Pama-Nyungan family, which has a greater time-depth (and thus more lexical differentiation), but is more homogenous in terms of phonotactics. In contrast, the lexically more similar Japonic lects still show phonotactic differentiation through e.g. differing obstruent mergers and monophthongisation patterns, making the addition of phonotactic data worthwhile (see also below).

For all contemporary lects,² we extracted biphones from the 256 lexical items, following the procedure described in Macklin-Cordes et al. (2021), which takes all two-segment sequences from each lexical item (including word boundaries)—e.g. Tokyo *atama* ‘head’ has the following biphones: #a, at, ta, am, ma, a#. Doing this for the entire word list creates a biphone inventory for each lect. Doing this for all lects creates an inventory of biphones across the entire family. We then coded binary presence/absence of each biphone in every lect.

2.2.1 The added value of phonotactic data

Following Macklin-Cordes et al. (2021), we measured the phylogenetic signal in our data using the *D* statistic (Fritz and Purvis 2010). Each independent binary character was tested against two null hypotheses: 1, Its values are distributed randomly with regards to phylogeny; and 2, its values are distributed as expected from a Brownian evolution model. To account for this, we performed a Bonferroni correcting, dividing the threshold for statistical significance by two (0.025). For the reference phylogeny, we matched our lects to the standard classification of the Japonic languages (adapted from Fig. 4.7 in De Boer 2020). The *D* statistic and associated *P*-values were calculated in R, using the *phylo.d* function in the *caper* package (Orme et al. 2012). Table 1 shows how many characters contain phylogenetic signal—for the lexical data by itself, versus for the combined lexical and phonotactic data.

The phonotactic data added an additional 355 characters with phylogenetic signals. A chi-squared test also showed a significant difference between the two datasets, $\chi^2(2) = 6.41$, $P = 0.041$. Standardized residuals revealed that the combined dataset contained significantly more characters with phylogenetic signal (81 per cent vs. 76 per cent, std. res. = 1.98), significantly fewer characters of indeterminate nature (17 per cent vs. 22 per cent, std. res. = -2.43), and no differences in the number of characters consistent with randomness (3 per cent vs. 2 per cent, std. res. = 0.90). Together, this shows the added value of phonotactic data for Japonic.

2.3 Phylogenetic analyses

Each model contained both the lexical- and phonotactic characters, i.e. the two data types were analysed jointly to infer a single phylogenetic tree. We fitted two-state continuous-time Markov chain (CTMC) models to

²No phonotactic characters were extracted for Old Japanese given the longstanding debate around its precise phonology—although see Miyake (2013) for an in-depth attempt at reconstruction.

Table 1. Comparison of the number of characters containing phylogenetic signals across the lexical and phonotactic datasets.

| | Phylogenetic signal | No phylogenetic signal | Indeterminate |
|----------------------------|---------------------|------------------------|---------------|
| Lexical data only | 435 | 11 | 123 |
| Lexical + phonotactic data | 790 | 26 | 163 |

$\chi^2(2) = 6.41, P = 0.041$

our binary data in BEAST 2 (Bouckaert *et al.* 2019). In addition to a strict clock model, we considered a log-normal relaxed clock model to account for potential rate variation across branches. Gamma-site rate models were employed to allow for rate variation across characters. As the properties of evolution for lexical features can be expected to be different and separate to that for phonotactic features, the CTMC-, gamma-site rate-, and clock models were fitted separately for the lexical and phonotactic data within each run. To estimate the trees in calendar time units, we applied two priors. For the Japanese branch, we applied a log-normal prior on the Old Japanese tip based on the attestation of Old Japanese ($M = 1,260$, $SD = 25.5$; as in Lee and Hasegawa 2011). For the Ryukyuan branch, we applied a normal prior ($M = 1,023$, $SD = 50$) on the time of the most recent common ancestor (tMRCA) of Southern Ryukyuan,³ which forms a coherent linguistic subgroup (Pellard 2015). Archaeology suggests that Japonic speakers settled the Southern Ryukyus a single wave, based on the emergence of the *Gusuku*-type pottery and agriculture in the 11th century (see e.g. Asato and Doi 1999; Asato *et al.* 2004; Takamiya 2005; Pearson 2013), which also lines up with the timing of the genetic differentiation of the Southern Ryukyuan population (Matsunami *et al.* 2021; Cooke *et al.* 2023; Koganebuchi *et al.* 2023). As a sensitivity analysis, we considered two different tree priors: the calibrated Yule (Heled and Drummond 2012) and the calibrated birth-death model (Heled and Drummond 2015), which shared the same prior density on the tMRCA of Southern Ryukyuan. In addition, we ran all models without Old Japanese to ensure the lack of phonotactic data did not majorly affect our results and found that the tree topology was robust to excluding Old Japanese.

3. Results

In the description below, we summarize the findings across all different models. The OSF repository linked

³We define the Southern Ryukyuan branch to include the lects Hateruma, Hatoma, Hirara, Ikema, Iriomote, Ishigaki, Kuroshima, Nagahama, Ohama, Oura, Tarama, Taketomi, Uechi, and Yonaguni.

in the *Data availability statement* contains the maximum clade credibility trees and full outcomes of each individual model, including those excluding Old Japanese.

3.1 Tree topology

A primary split between a mainland Japanese clade—including Kyushu—on the one hand, and a Ryukyuan clade on the other, was present with posterior probability > 0.9 across all tree priors and clock models. Figure 1 summarizes the topologies of the Japanese and Ryukyuan clades, with reference to their geographical patterning.

Within the Japanese clade, Old Japanese appeared as an outgroup to the rest of the clade in all analyses. The posterior probability for the clade comprising all contemporary lects was 1 across all tree priors and clock models. Across all analyses, a core Tohoku Japanese clade emerged with posterior probability 1, which was the first to separate from its ancestor. For the remaining mainland lects, the relaxed clock analyses suggested a large Central Japanese clade, comprising everything west of the Japanese Alps (posterior probabilities: Yule 0.88; Birth-Death 0.90). Within Central Japanese, we found a major split between the ‘core’ lects spoken around the historical capital area Kyoto-Nara (posterior probability 0.95 across both tree priors), and the remaining ‘peripheral’ lects (posterior probabilities: 0.93; 0.94). The Kyushu lects consistently form a subclade with posterior probability 1 within this peripheral subgroup. These same subgroups (‘core’ Central, ‘peripheral’ Central, and Kyushu) were strongly supported in the strict clock analyses as well, but the structure of the clade as a whole was less resolved. The lects spoken east of the Japanese Alps appeared as an outgroup to the large Central Japanese clade but with less coherence (posterior probabilities 0.5–0.7). Finally, the position of some individual lects (e.g. Tokyo, Hachijo, Toyama, Shimane) also varied considerably between models.

Within the Ryukyuan clade, Amami and Okinawa emerged as clear subgroups, together forming a stable Northern Ryukyuan subclade—all posterior probabilities > 0.99 across all tree priors and clock

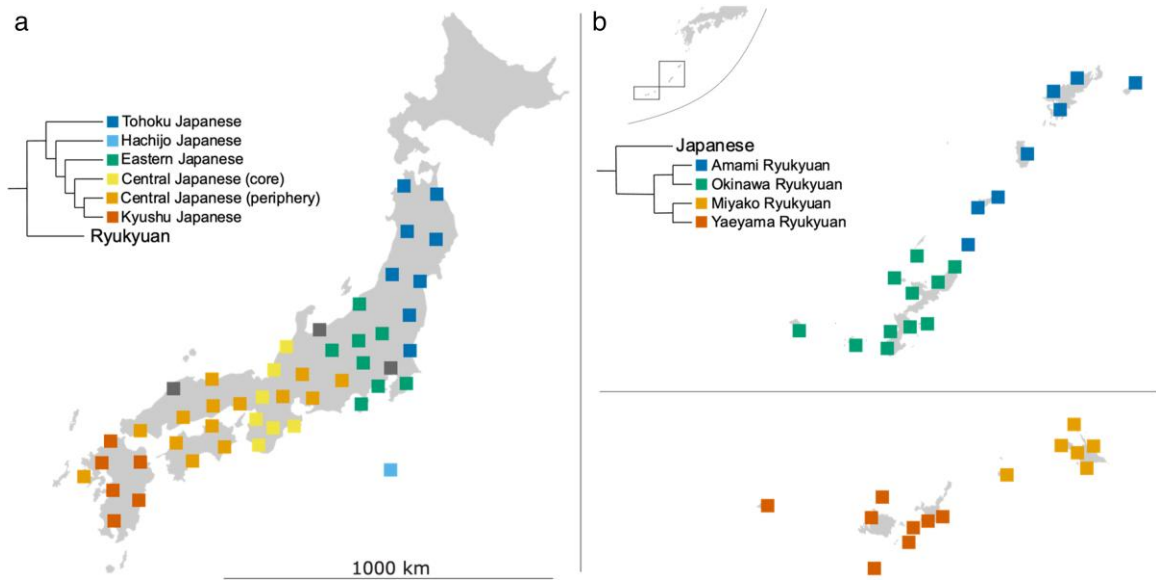


Figure 1. Map showing the locations of the lects included in this study, with summarized topologies for (a) the Japanese clade and (b) the Ryukyuan clade, with main subgroups discussed in the running texts colour-coded—dark grey indicating that the position of that lect remains unresolved.

models. Within Southern Ryukyuan (which was monophyletically constrained), a clear Miyako subgroup emerged with a posterior probability of 1 across all analyses. A monophyletic Macro-Yaeyama subgroup emerged consistently in the strict clock analyses (0.99 across both tree priors), but its posterior probability was considerably lower in the relaxed clock analyses (0.49 across both tree priors).

3.2 Dating

The upper bounds of the 95 per cent Bayesian credible intervals (BCIs) for the estimated age of the entire family (12,000–880 BCE) did not entirely rule out an initial diversification during the Jomon period. However, the full BCIs and median values (470–380 BCE) strongly suggest that the initial diversification of Japonic happened during the Yayoi period. The lower bounds of the BCIs (210–70 BCE) predate the Kofun period.

The 95 per cent BCIs of the Japanese clade tMRCA range from 980 to 1,420 CE, which clearly suggests diversification in the Middle Japanese period (9th–16th century), but it is less certain whether it occurred after the Early Middle Japanese period (posterior probabilities: 0.592–0.661).

The lower bounds of the 95 per cent BCIs (970–980 CE) of the Ryukyuan clade tMRCA predate the *Gusuku* period (11th century onwards), suggesting

diversification prior to this (posterior probabilities: 0.991–0.994), coinciding with the first population movements from the mainland into the Ryukyu islands. [Figure 2](#) summarizes the findings across the combinations of clock models and tree priors, with reference to the findings reported by [Lee and Hasegawa \(2011\)](#).

4. Discussion and conclusion

The analyses presented here strongly support a primary split between mainland Japanese and Ryukyuan, in line with both traditional views in dialectology and work in (computational) historical linguistics ([Pellard 2009, 2015; Lee and Hasegawa 2011](#)). We found no support for the recently revived proposal grouping together the Kyushu and Ryukyu lects into one clade (cf. [De Boer 2020; Igarashi 2023a](#)). Lexical correspondences for this proposal are generally only found in non-basic vocabulary ([Jarosz 2019; Jarosz and Orlandi 2023](#)), which, together with other commonly cited evidence—e.g. shared sound changes occurring in one specific verb class ([Hattori 1978; Igarashi 2023b](#))—are more likely to result from contact from when Proto-Ryukyuan was still spoken in Kyushu (see also [Pellard 2015, 2021](#)). Regarding the age of the family, [Pellard \(2015\)](#) previously noted the temporal discrepancy between the start of the Yayoi period around 1,000 BCE and the first split between Japanese and

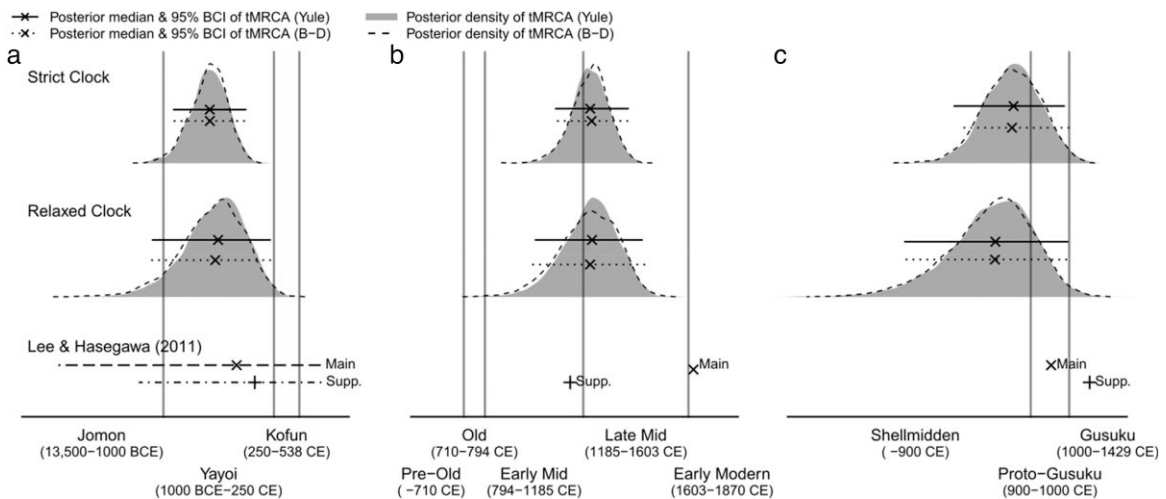


Figure 2. Posterior estimates of tMRCA for (a) all taxa, (b) the contemporary Japanese lects, and (c) the Ryukyuan lects. For each of the (a)–(c), the top and middle parts present the posterior estimates from our study for the strict clock analyses and relaxed clock analyses, respectively. The shaded areas represent the posterior densities estimated from the analyses with the calibrated Yule prior, while the dashed contour lines represent those estimated from the analyses with the calibrated birth-death prior. For a given clock model, the solid bars represent the 95 per cent Bayesian credible intervals from the analyses with the calibrated Yule prior, while dotted bars represent the intervals from using the calibrated birth-death prior. Corresponding posterior medians are indicated by the crosses on the bars. The bottom part of each panel presents the posterior estimates presented in the main text ('Main') and supplementary materials ('Supp.') by Lee and Hasegawa (2011), where the posterior medians are available for all tMRCA estimates, but Bayesian credible intervals for the root only.

Ryukyuan occurring as much as 800 years later, dated to around 200 BCE (Lee and Hasegawa 2011).⁴ The models presented here estimated the Japanese-Ryukyuan split to have occurred around 400–500 BCE, which means this discrepancy is also present in our results, albeit to a lesser degree. While there is no clear explanation, Vovin (2017) has suggested that Japonic was once a more diverse language family than it is today, and perhaps these estimated ages reflect the partial loss of this diversity.

The topology of the mainland Japanese clade uncovered in our analyses neatly unifies the two divisions recognized in traditional Japanese dialectology: the East-West division based on various structural features, and the center-periphery division based on patterns of lexical diffusion (see Kawaguchi and Inoue 2002). At the same time, further work is required to determine the position of divergent lects such as Shimane and Hachijo. In addition, the age estimates provided by the models raise more questions than they provide

answers, as they suggest that the contemporary mainland Japanese lects share a common ancestor that is much more recent than Old Japanese. This is puzzling given the attested differences between the central and eastern varieties of Old Japanese, some of which persist to this day. Further research is needed to explain this apparent paradox. Attestations of Eastern Old Japanese are not abundant enough to precisely determine its lexical differentiation (Vovin and Ishisaki-Vovin 2021), but perhaps whatever diversity there was has been slowly eroded over time under the influence of historically high-prestige and (later) standard(-like) varieties.

The major subgroups uncovered in the topology of the Ryukyuan clade largely followed geographical segmentation—albeit with some uncertainty around some Macro-Yaeyama lects (Hateruma and Yonaguni). Further divisions of the subgroups also largely followed previously suggested classifications—see e.g. Lawrence (2000, 2006) on Okinawa and Yaeyama, respectively, and Pellard (2009) on Miyako. The ancestor of the Ryukyuan branch was dated to around the turn of the 9th century, which coincides with the first traces of cereal farming appearing in Amami and Okinawa between the 8th and 10th centuries (see also Jarosz *et al.* 2022). The relatively short branch between this ancestor and the Southern Ryukyuan subgroup suggests that the entire island chain was settled fairly soon after the initial movement into the Ryukyu islands, representing

⁴At the same time, Pellard (2015, 2021) has suggested that the split between Japanese and Ryukyuan happened during the Kofun period (250–500 CE), which would make the discrepancy even larger. Recent ancient DNA studies have identified a component in Kofun period individuals not present in Yayoi period individuals (Cooke *et al.* 2021) and found that both contemporary mainland Japanese and Ryukyuan populations are genetically more similar to these Kofun individuals than they are to the Yayoi individuals (Cooke *et al.* 2023). This raises questions about who exactly the people associated with this DNA component were, and how—if at all—they influenced the linguistic situation.

a burst-like expansion also found in Austronesian (Gray et al. 2009).

In sum, the analyses confirm previous findings on the age of the family as a whole, provide greater certainty about the linguistic history and internal structure of mainland Japanese, as well as a more detailed perspective on the history of the Ryukyuan languages. We present our updated phylogenetic tree of the Japonic language family to be used as a resource in broader studies on language evolution.

Finally, our results show that other types of data can successfully complement lexical data—the standard in phylolinguistic studies. In this regard, our study contrasts with Macklin-Cordes et al. (2021), who found that the addition of phonotactic characters did not produce a better model. In the Japonic case, they were particularly useful for inferring the internal branch structure, likely because of its relatively shallow time-depth. Moreover, the phonotactic data added information as binary characters, which makes integration with the binary cognacy data more straightforward—cf. Macklin-Cordes et al. (2021), who needed to incorporate phonotactics as frequency-based features. Future work on the finer structure of other language families might thus benefit from binary phonotactic data, especially as it can be easily extracted from the lexical data.

Acknowledgements

We thank two anonymous reviewers for the helpful comments on an earlier version of this manuscript.

Conflict of interest statement. None declared.

Funding

This research was supported by the Scandinavia-Japan Sasakawa Foundation (grant number GA22-SWE-0073; ‘A linguistic database of the Ryukyu Islands to study the history of Japan’).

Data availability

The data underlying this article, as well as the file used in the analyses are available in an OSF repository, at <https://osf.io/5yqf9/>.

References

- Arakaki, K. (2000) *A Descriptive Comparative Study of the Ryukyuan Dialects—Phonology, Verbal Morphology, and Particles*. Chiba, Japan: Chiba University.
- Asato, S. et al. (2004) *Okinawa-ken no rekishi*. Tokyo, Japan: Yamagawa shuppan.
- Asato, S. and Doi, N. (1999) *Okinawa-jin wa doko kara kita ka? Ryūkyū=Okinawa-jin no kigen to seiritsu*. Naha, Japan: Bōdā inku.
- Beyer, R. et al. (2019) ‘Environmental Conditions do not Predict Diversification Rates in the Bantu Languages’, *Heliyon*, 5: e02630. <https://doi.org/10.1016/j.heliyon.2019.e02630>
- Bouckaert, R. R. et al. (2019) ‘BEAST 2.5: An Advanced Software Platform for Bayesian Evolutionary Analysis’, *PLoS computational biology*, 15: e1006650. <https://doi.org/10.1371/journal.pcbi.1006650>
- Bouckaert, R. R., Bown, C. and Atkinson, Q. D. (2018) ‘The Origin and Expansion of Pama–Nyungan Languages Across Australia’, *Nature Ecology & Evolution*, 2: 741–9. <https://doi.org/10.1038/s41559-018-0489-3>
- Chang, W. et al. (2015) ‘Ancestry-constrained Phylogenetic Analysis Supports the Indo-European Steppe Hypothesis’, *Language*, 91: 194–244. <https://doi.org/10.1353/lan.2015.0005>
- Cooke, N. P. et al. (2021) ‘Ancient Genomics Reveals Tripartite Origins of Japanese Populations’, *Science Advances*, 7: eabh2419. <https://doi.org/10.1126/sciadv.abh2419>
- Cooke, N. P. et al. (2023) ‘Genomic Insights into a Tripartite Ancestry in the Southern Ryukyu Islands’, *Evolutionary Human Sciences*, 5: e23. <https://doi.org/10.1017/ehs.2023.18>
- de Boer, E. (2020) ‘The Classification of the Japonic Languages, in *The Oxford Guide to the Transeurasian Languages*, pp. 40–58. Oxford, UK: Oxford University Press.
- de Boer, E. et al. (2020) ‘Japan Considered from the Hypothesis of Farmer/Language Spread’, *Evolutionary Human Sciences*, 2: e13. <https://doi.org/10.1017/ehs.2020.7>
- de Boer, E. (2024) *Using Tonal Data to Recover Japanese Language History*. Amsterdam, the Netherlands: John Benjamins.
- Dockum, R. ‘Phylogeny in phonology: how Tai sound systems encode their past’, in *Proceedings of the Annual Meetings on Phonology*. New York, USA: Linguistic Society of America. 2017.
- Dunn, M. et al. (2011) ‘Evolved Structure of Language Shows Lineage-Specific Trends in Word-Order Universals’, *Nature*, 473: 79–82. <https://doi.org/10.1038/nature09923>
- Ferraz Gerardi, F. et al. (2023) ‘Lexical Phylogenetics of the Tupi-Guaraní Family: Language, Archaeology, and the Problem of Chronology’, *PLoS One*, 18: e0272226. <https://doi.org/10.1371/journal.pone.0272226>
- Fritz, S. A. and Purvis, A. (2010) ‘Selectivity in Mammalian Extinction Risk and Threat Types: A New Measure of Phylogenetic Signal Strength in Binary Traits’, *Conservation Biology*, 24: 1042–51. <https://doi.org/10.1111/j.1523-1739.2010.01455.x>
- Fujiwara, Y. (1962) *Hogengaku [Dialectology]*. Tokyo, Japan: Sansendo.
- Gray, R. D., Drummond, A. J. and Greenhill, S. J. (2009) ‘Language Phylogenies Reveal Expansion Pulses and Pauses in Pacific Settlement’, *Science*, 323: 479–83. <https://doi.org/10.1126/science.1166858>
- Greenhill, S. J. et al. (2017) ‘Evolutionary Dynamics of Language Systems’, *Proceedings of the National Academy of Sciences*, 114: E8822–9. <https://doi.org/10.1073/pnas.1700388114>

- Greenhill, S. J., Blust, R. and Gray, R. D. (2008) 'The Austronesian Basic Vocabulary Database: From Bioinformatics to Lexomics', *Evolutionary Bioinformatics*, 4: EBO.S893. <https://doi.org/10.4137/EBO.S893>
- Greenhill, S. J., Heggarty, P. and Gray, R. D. (2020) 'Bayesian Phylolinguistics', in Janda, R. D., Joseph, B. D., Vance, B. S. (eds.) *The Handbook of Historical Linguistics*, 1st edn, pp. 226–53. New Jersey, USA: Wiley.
- Guillon, M. and Mace, R. (2016) 'A Phylogenetic Comparative Study of Bantu Kinship Terminology Finds Limited Support for its co-Evolution with Social Organisation', *PLoS One*, 11: e0147920. <https://doi.org/10.1371/journal.pone.0147920>
- Hanihara, K. (1991) 'Dual Structure Model for the Population History of the Japanese', *Nichibunken Japan Review*, 2: 1–33. <https://www.jstor.org/stable/25790895>
- Hattori, S. (1954) 'On the Method of Glottochronology and the Time-Depth of Proto-Japanese', *Gengo Kenkyu (Journal of the Linguistic Society of Japan)*, 1954: 29–77. https://doi.org/10.11435/gengo1939.1954.26-27_29
- Hattori, S. (1973) 'Japanese Dialects', in Hoenigswald, H. M. (ed.) *Diachronic, Areal, and Typological Linguistics*, pp. 368–400. Berlin, Boston: De Gruyter.
- Hattori, S. (1976) 'Ryūkyū hōgen to hondo hōgen [The Ryukyuan Dialects and the Mainland Dialects]', in Iha Fuyū 100th Anniversary Commemoration Association (ed.) *Okinawagaku no reimei [The Dawn of Okinawan Studies]*, pp. 7–55. Naha, Japan: Okinawa Bunka Kyōkai.
- Hattori, S. (1978–1979) 日本祖語について(1-22)[About Proto-Japanese 1-22], 月刊言語 [Preprint].
- Haynie, H. J. and Bowern, C. (2016) 'Phylogenetic Approach to the Evolution of Color Term Systems', *Proceedings of the National Academy of Sciences*, 113: 13666–71. <https://doi.org/10.1073/pnas.1613666113>
- Heggarty, P. et al. (2023) 'Language Trees with Sampled Ancestors Support a Hybrid Model for the Origin of Indo-European Languages', *Science*, 381: eabg0818. <https://doi.org/10.1126/science.abg0818>
- Heled, J. and Drummond, A. J. (2012) 'Calibrated Tree Priors for Relaxed Phylogenetics and Divergence Time Estimation', *Systematic Biology*, 61: 138–49. <https://doi.org/10.1093/sysbio/syr087>
- Heled, J. and Drummond, A. J. (2015) 'Calibrated Birth–Death Phylogenetic Time-Tree Priors for Bayesian Inference', *Systematic Biology*, 64: 369–83. <https://doi.org/10.1093/sysbio/syu089>
- Hirayama, T., ed. (1966) *Ryūkyū hōgen no sōgōteki kenkyū [A Comprehensive Study of the Ryukyuan Dialects]*. Tokyo, Japan: Meiji-shoin.
- Hirayama, T., ed. (1967) *Ryūkyū Sakishima hōgen no sōgōteki kenkyū [A Comprehensive Study of the Ryukyuan Sakishima Dialects]*. Tokyo, Japan: Meiji-shoin.
- Hirayama, T. (1992–1994) *Gendai Nihongo Hōgen Daijiten [Dictionary of Contemporary Japanese Dialects]*. Tokyo, Japan: Meiji-shoin.
- Honkola, T. et al. (2013) 'Cultural and climatic changes shape the evolutionary history of the Uralic languages', *Journal of Evolutionary Biology*, 26: 1244–53.
- Hudson, M. J., Nakagome, S. and Whitman, J. B. (2020) 'The Evolving Japanese: The Dual Structure Hypothesis at 30', *Evolutionary Human Sciences*, 2: e6. <https://doi.org/10.1017/ehs.2020.6>
- Huisman, J. L., Majid, A. and Van Hout, R. (2019) 'The Geographical Configuration of a Language Area Influences Linguistic Diversity', *PLoS One*, 14: e0217363. <https://doi.org/10.1371/journal.pone.0217363>
- Igarashi, Y. (2023a) 'Controversy about the phylogenetic position of Kyushu and Ryukyuan languages: current situation and future prospects'. The Origin and Spread of the Japonic Languages: Putting Together Linguistics, Genetics, and Archaeology.
- Igarashi, Y. (2023b) '現代九州諸方言における旧上二段動詞の「下二段化」は九州・琉球祖語仮説を支持するか?', 言語研究, 163: 1–31. https://doi.org/10.11435/gengo.163.0_1
- Itoyō, K., Hino, S. and Sato, R. (1982–1984) *Kouza Hougengaku (Volumes 4–10)*. Tokyo, Japan: Kokusho Kankoukai.
- Janhunen, J. (2023) 'Tungusic in Time and Space', in *The Tungusic Languages*, pp. 517–37. Milton Park, UK: Routledge.
- Jarosz, A. (2019) 'Non-Core Vocabulary Cognates in Ryukyuan and Kyushu', in *Proceedings of International Symposium Approaches to Endangered Languages in Japan and Northeast Asia-Poster Session*, pp. 8–29. Tokyo, Japan: Tokyo University of Foreign Studies.
- Jarosz, A. et al. (2022) 'Demography, Trade and State Power: A Tripartite Model of Medieval Farming/Language Dispersals in the Ryukyuan Islands', *Evolutionary Human Sciences*, 4: e4. <https://doi.org/10.1017/ehs.2022.1>
- Jarosz, A. and Orlandi, G. (2023) 'Common Kyushu-Ryukyuan Substratum in Maritime Vocabulary: A Preliminary Analysis', *Lingua Posnaniensis*, 65: 7–46. <https://doi.org/10.14746/linpo.2023.65.2.1>
- Jordan, F. M. et al. (2009) 'Matrilocal Residence is Ancestral in Austronesian Societies', *Proceedings of the Royal Society B: Biological Sciences*, 276: 1957–64. <https://doi.org/10.1098/rspb.2009.0088>
- Kawaguchi, Y. and Inoue, F. (2002) 'Japanese Dialectology in Historical Perspectives', *Revue belge de Philologie et d'Histoire*, 80: 801–29. <https://doi.org/10.3406/rbph.2002.4642>
- Kindaichi, H. (1955) 'Nihongo (hōgen) [Japanese (Dialects)]', in Ichikawa, S., Kōzu, H. (eds.) *Sekai Gengo Gaisetsu*, pp. 212–38. Tokyo, Japan: Kenkyusha.
- Koganebuchi, K. et al. (2023) 'Demographic History of Ryukyuan Islanders at the Southern Part of the Japanese Archipelago Inferred from Whole-Genome Resequencing Data', *Journal of Human Genetics*, 68: 759–67. <https://doi.org/10.1038/s10038-023-01180-y>
- Koile, E. et al. (2022) 'Phylogeographic Analysis of the Bantu Language Expansion Supports a Rainforest Route', *Proceedings of the National Academy of Sciences*, 119: e2112853119. <https://doi.org/10.1073/pnas.2112853119>
- Kupchik, J. (2023) *Azuma Old Japanese: A Comparative Grammar and Reconstruction*. Berlin, Germany: Walter de Gruyter GmbH & Co KG.

- Lawrence, W. (2000) ‘八重山方言の区画について [On the Classification of the Yaeyama Dialects]’, in Ishigaki, S. (ed.) 宮良當壯記念論集, pp. 547–59. Okinawa, Japan: Hirugi.
- Lawrence, W. (2006) ‘沖繩方言群の下位区分について [On the Subclassification of the Okinawan Dialects]’, *Okinawa Bunka*, 100: 101–18.
- Lee, S. and Hasegawa, T. (2011) ‘Bayesian Phylogenetic Analysis Supports an Agricultural Origin of Japonic Languages’, *Proceedings of the Royal Society B: Biological Sciences*, 278: 3662–9. <https://doi.org/10.1098/rspb.2011.0518>
- Macklin-Cordes, J. L., Bowern, C. and Round, E. R. (2021) ‘Phylogenetic Signal in Phonotactics’, *Diachronica*, 38: 210–58. <https://doi.org/10.1075/dia.20004.mac>
- Matsunami, M. et al. (2021) ‘Fine-Scale Genetic Structure and Demographic History in the Miyako Islands of the Ryukyu Archipelago’, *Molecular Biology and Evolution*, 38: 2045–56. <https://doi.org/10.1093/molbev/msab005>
- Miyake, M. H. (2013) *Old Japanese: A Phonetic Reconstruction*. Milton Park, UK: Routledge.
- Moseley, C., ed. (2010) *Atlas of the World’s Languages in Danger*. Paris, France: UNESCO Publishing.
- Nakamoto, M. (1976) *Ryukyubougen On’in no Kenkyuu [A Study of the Phonology of the Ryukyu Dialects]*. Tokyo, Japan: Hosei Daigaku Shuppankyoku.
- Nakamoto, M. (1981) *Zusetsu Ryukyugo Jiten [Illustrated Dictionary of the Ryukyuan Languages]*. Tokyo, Japan: Rikitomo Shobo.
- Orme, D. et al. (2012) ‘Caper: comparative analyses of phylogenetics and evolution in R’, R package version 0.5, 2, p. 458. <https://doi.org/10.32614/cran.package.caper>
- Pearson, R. (2013) *Ancient Ryukyu: An Archaeological Study of Island Communities*. Honolulu, USA: University of Hawaii Press.
- Pellard, T. (2009) *Ōgami: Éléments de description d’un parler du Sud des Ryūkyū*. Paris, France: Ecole des Hautes Etudes en Sciences Sociales.
- Pellard, T. (2015) ‘1. the Linguistic Archeology of the Ryukyu Islands’, in Heinrich, P., Miyara, S. and Shimoji, M. (eds.) *Handbook of the Ryukyuan Languages*, pp. 13–38. Berlin, Germany: De Gruyter.
- Pellard, T. (2021) ‘Nichiryū shogo no keitō bunrui to bunki ni tsuite (日琉諸語の系統分類と分岐について)’, in *Firudo to bunken kara miru Nichiryū shogo no keitō to rekishi (フィールドと文献から見る日琉諸語の系統と歴史)[The Phylogeny and History of the Japonic Languages Seen from the Field and from Philological Records]*, pp. 2–16. Tokyo, Japan: Kaitakusha.
- Peng, V. M. and Peng, F. C. C., (1990) ‘The Application of the Comparative Method to Japanese Dialects’, in Peng, F. C. C. et al. (eds) *Varieties of Language Behavior*. Bunka Hyoron Publishing.
- Robbeets, M. et al. (2021) ‘Triangulation Supports Agricultural Spread of the Transeurasian Languages’, *Nature*, 599: 616–21. <https://doi.org/10.1038/s41586-021-04108-8>
- Shcherbakova, O. and Allasonnière-Tang, M. (2024) ‘Evolutionary Pathways of Complexity in Gender Systems’, *Journal of Language Evolution*, 8: 120–133. <https://doi.org/10.1093/jole/lzae001>
- Smith, A. H. (1960) ‘The Culture of Kabira, Southern Ryūkyū Islands’, *Proceedings of the American Philosophical Society*, 104: 134–71. <https://www.jstor.org/stable/985656>
- 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: 452–63. <https://www.jstor.org/stable/3143802>
- Swadesh, M. (1955) ‘Towards Greater Accuracy in Lexicostatistic Dating’, *International Journal of American Linguistics*, 21: 121–37. <https://doi.org/10.1086/464321>
- Tadmor, U. (2009) ‘III. Loanwords in the World’s Languages: Findings and Results’, in Haspelmath, M. and Tadmor, U. (eds.) *Loanwords in the World’s Languages*, pp. 55–75. Berlin, Germany: Walter de Gruyter.
- Takahashi, T., Onohara, A. and Ihara, Y. (2023) ‘Bayesian Phylogenetic Analysis of Pitch-Accent Systems Based on Accentual Class Merger: A new Method Applied to Japanese Dialects’, *Journal of Language Evolution*, 8: 169–91. <https://doi.org/10.1093/jole/lzae004>
- Takamiya, H. (2005) *Shima no senshigaku: Paradaisu dewanakarta Okinawa shotō no senshi jidai*. Naha, Japan: Bōdā inku.
- Takubo, Y. (2018) ‘Mutual Intelligibility as a Measure of Linguistic Distance and Intergenerational Transmission’, in *NINJAL International Symposium on Approaches to Endangered Languages in Japan and Northeast: Description, Documentation and Revitalization*.
- Thorpe, M. (1983) *Ryukyuan Language History*. Los Angeles, USA: University of Southern California.
- Tian, Z. et al. (2022) ‘Triangulation fails when neither linguistic, genetic, nor archaeological data support the Transeurasian narrative’, bioRxiv, <https://doi.org/10.1101/2022.06.09.495471>, 12 June 2022, preprint: not peer reviewed.
- Tojo, M. (1927) *Kokugo no Hogen Kukaku [Dialectal Divisions of Japanese]*. Tokyo, Japan: Ikuei Shoin.
- Tojo, M. (1951) *Zenkoku Hogen Jiten [A Dictionary of the Dialects of the Whole Country]*. Tokyo, Japan: Tokyodo.
- Tominaga, Y. (1988) *Dialect Comprehensibility in Japan: A Study in Determining the Linguistic Distance Between Dialects of the Japanese Language*. San Francisco, USA: University of San Francisco.
- Vovin, A. (2011) ‘Why Japonic is not Demonstrably Related to “Altaic” or Korean’, *Historical Linguistics in the Asia-Pacific Region and the Position of Japanese*, pp. 17–25. Osaka, Japan: National Museum of Ethnology.
- Vovin, A. (2017) ‘Origins of the Japanese Language’, in *Oxford Research Encyclopedia of Linguistics*. Oxford, UK: Oxford University Press.
- Vovin, A. and Ishisaki-Vovin, S. (2021) ‘The Eastern Old Japanese Corpus and Dictionary’, *The Eastern Old Japanese Corpus and Dictionary*. Leiden, the Netherlands: Brill.
- Yamada, M. et al. (2020) ‘Experimental Study of Inter-Language and Inter-Generational Intelligibility: Methodology and Case Studies of Ryukyuan Languages.’, *Japanese/Korean Linguistics*, 26: 249–60.
- Yamagiwa, J. K. (1967) ‘On Dialect Intelligibility in Japan’, *Anthropological Linguistics*, 9: 1–17. <https://www.jstor.org/stable/30029037>