Using Demographic Data with Bayesian Phylogenies: A Japanese Case Study

Richard Littauer // University of Malta & Saarland University
richard.littauer@gmail.com // @richlitt

Take-away

Lee and Hasegawa (2011) used a Bayesian phylogenetic analysis on 59 Japanese dialects to show that it is highly probable that the Japanese language developed in the last 2000 years. But their shallow branches are problematic (Whitman 2011).

By comparing the shallow branches with demographic data, it should be possible to check the validity of Bayesian metrics for dialects on a shorter timeframe. Here, I present plans for future work, in the hopes of fostering discussion about shallow branches in larger Bayesian phylogenetic trees.

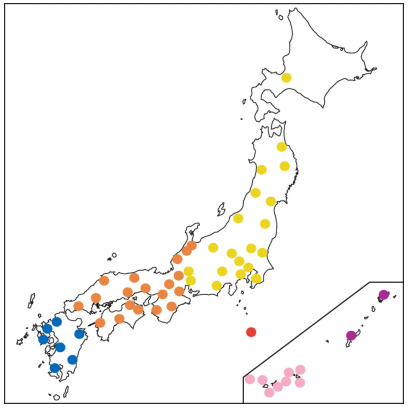
Lee & Hasegawa 2011

The use of Bayesian phylogenetic methods to trace population expansion and language change has been frequent in recent years. Such statistical studies, working on cognate lists, have shown to a reasonable degree possible lineages of languages from many language families.

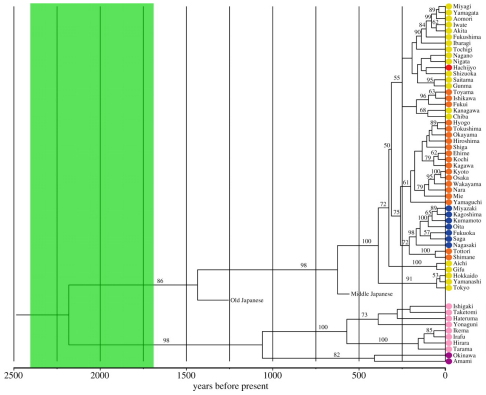
In this vein, Lee and Hasegawa (2011) used a Bayesian phylogenetic analysis on 59 Japanese dialects [see map at right] to show that it is highly probable that they have a common ancestor within the the last 2000 years, with the introduction of rice farming and the Yayoi expansion into the Japanese archipelago, instead of in the past 12,000-30,000 years, which is the projected time span for inhabitation of the archipelago by a hunter-gatherer society. To do this, they ran BEAST* (which uses Bayesian Markov chains Monte Carlo sampling methods) cognate lists from the various dialects they compiled, as well as a separate, control dataset. This has been used before on Semitic languages (Kitchen et al. 2009). They use a probabilistic divergence time calibration prior for only one historical known event, the switch of populations mirroring the imperial court moving from Kyoto to Edo (Tokyo).

In their proposed tree, the majority of the dialect splits occurred in the past three hundred years; however, "the phylogeny selected by Hasegawa and Lee is problematic in its shallower branches, which represent all non-Ryūkyūan branches as descended from Early Middle Japanese, but it is not clear that this affects their overall results." And later: "We know that the actual date of dispersal [of the Japonic languages in the archipelago] must be earlier, but we do not know how much. Even Lee and Hasegawa's date, first century BCE, leaves a 900-year lag between the archaeological event and the linguistic evidence for dispersal." (Whitman 2011: 155)

While it is admittedly not clear how much minor differences in shallow branches influence the final decision, it should be possibly to analyze the shallow branches for their historical accuracy, or for consistent errors. This would potentially cast light on the use of Bayesian phylogenetic methods for shallow analysis, helping to understand recent chronological change where arduous, manual analysis would have had to be performed before. It could also show the problems with using Bayesian analysis to diagnose recent language change. One way of analyzing the shallow branches is to compare them to relative research, as has been overviewed in Whitman (2011), and in Lee & Hasegawa (2011) in passing. Another way would be to use other datasets to compare the splits statistically.



Map of Japonic Languages
Lee S , Hasegawa T Proc. R. Soc. B doi:10.1098/rspb.2011.0518



Maximum clade credibility tree of Japonic languages.

All node heights in the tree are scaled to match the posterior median node heights. The value on each branch of the tree is the posterior probability, showing the percentage support for a node following a particular branch. Posterior probabilities below 50% are not shown. The green bar represents the age range predicted by the farming/language theory (1700–2400 YBP).
Lee S , Hasegawa T Proc. R. Soc. B doi:10.1098/rspb.2011.0518

Methodology

Accurate census data for cities and regions has been gathered in Japan for roughly the past three hundred years. This data can be extracted from translated tables on Wikipedia**, and then fact checking these tables using the original sources and Japanese open census data. This data can then be compared to the proposed splits to see if growth correlates with splitting in a predictable fashion – a deeper historical analysis is planned for particular cases.

As well as checking the shallow branches of the Bayesian tree, the differences and possible causes of dialect shift and creation can be more closely examined. In particular, possible correlations between urban growth and dialect splitting, following Trudgill's (1974) analysis of city size influence, can be explored by comparing rate of change in the lexicon against rate of change in population of the cities. In certain cases, I hope to rerun the Bayesian analyses using BEAST on a smaller subset of the dialects in order to ascertain their probable divergence, ordering the possible changes using geographical distance combined with population size both as a proxy for contact in the new model, and as a robust signal of deviation from the standard (Wieling et al., 2011).

Future Directions

As this poster is meant to foster discussion and gather ideas as to how to approach this research, the work presented here is preliminary at best. As such, there is much that could fit under 'future work', such as:

- Gather demographic data from direct, Japanese sources (instead of, as here, Wikipedia).
- Mine the demographic and dialect data using traditional statistical methods to see if there are predictable patterns of movement and language change.
- Develop an add-on to BEAST or another program to deal with demographic data in shallow branches.
- Compare shallow branches on larger, family trees, where the shallow branches are not so well worked out by historical linguists.

References

• Kitchen, A., Ehret, C., Assefa, S. & Mulligan, C. J. 2009 Bayesian phylogenetic analysis of Semitic languages identifies an Early Bronze Age origin of Semitic in the Near East. Proc. R. Soc. B 276, 2703 – 2710. (doi:10. 1098/rspb.2009.0408)
• Lee, S. and Hasegawa, T. (2011). Bayesian phylogenetic analysis supports an agricultural origin of Japonic languages. Proc. R. Soc. B. 10.1098/rspb.2011.0518.
• Trudgill, P. (1974). Linguistic change and diffusion: Description and explanation in sociolinguistic dialect geography. Language in Society, 2:215–246.
• Whitman, J. (2011). Northeast Asian linguistic ecology and the advent of rice agriculture in Korea and Japan. Rice, 4:149–158. 10.1007/s12284-011-9080-0.
• Wieling, M., Nerbonne, J., and Baayen, R. H. (2011). Quantitative social dialectology: Explaining linguistic variation geographically and socially. PLoS ONE, 6(9):e23613. 1

*BEAST. http://beast.bio.ed.ac.uk/Main_Page
**http://en.wikipedia.org/wiki/Demographics_of_Japan_before_Meiji_Restoration