Using Demographic Data with Bayesian Phylogenies: A Japanese Case Study

Richard Littauer

January 30, 2012

Abstract

In this paper I compare urban demographics with proposed Bayesian phylogenies. The use of Bayesian phylogenic 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 Bayesian phylogenic analysis on 59 Japanese dialects to show that it is highly probable that the current Japanese language developed in the last 2000 years. In their proposed tree, the majority of the dialect splits occurred in the past three hundred years; however, the recent, or 'shallow' trees, are notably problematic in some respects, which may affect the overall results (Whitman, 2011).

Here I test this analysis by combining similar Bayesian analyses based on the word list used by Lee and Hasegawa with diachronic data regarding the demographics of the main cities and regions. Accurate census data has been gathered in Japan for roughly the past three hundred years. This coincides with a modern surge in population and increasing urbanisation, and by looking at urban growth and population expansion, combined with geographic data of the dialects involved, the differences and possible causes of dialect shift and creation can be more closely graphed, visualised, and examined. In particular, I explore possible correlations between urban growth and dialect splitting, following Trudgill's (1974) analysis of city size influence, by comparing rate of change in the lexicon against rate of change in population of the cities. In certain cases, I rerun the Bayesian algorithm on a smaller subset of the dialects in order to ascertain their probable divergence, using geographical distance both as a proxy for contact in the new model, and as a robust signal of deviation from the standard when combined with population size (Wieling et al., 2011).

There have been few or no studies done testing Bayesian phylogenic analysis against population data combined with geographical coordinates, as the shallow results are often not robust compared to deeper branches of the proposed trees; by comparing the shallow branches with data gathered from city populations, it may be possible to check the validity of Bayesian metrics for dialects in a shorter timeframe. I will present more fully this new methodology and the theoretical implications of short-term Bayesian analysis, as well as my preliminary results.

References

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.