Phylogenetic signals of language external factors

Christian Bentz, University of Tübingen

Recent large-scale quantitative studies have elicited potential links between "internal" structural features of languages – e.g. phoneme inventories, inflectional marking, lexical diversity – and so-called language "external" factors, such as population size, climate and geography (Atkinson 2011; Lupyan & Dale 2010; Everett, Blasi & Roberts 2015; Bentz 2016; Bentz, Verkerk, Kiela, Hill & Buttery 2015; Bentz & Winter 2013; Lewis & Frank submitted). These studies argue that links between "internal" and "external" factors can give us a window into past processes of language change, and thus a handle on explaining current linguistic diversity. A potential confound in such statistical models is the genealogical relatedness of languages, and their resulting non-independence.

Phylogenetic signal analyses are a tool from evolutionary biology used to estimate the degree of phylogenetic clustering of species and their phenotypic properties (see e.g. Münkemüller, Laveregne, Bzeznik, Dray, Jombart, Schiffers & Thuiller 2012; Revell, Harmon & Collar 2008). These methods have also been applied to estimate phylogenetic signals of linguistic features such as motion event encoding in Indo-European languages (Verkerk 2014), and lexical diversity in Indo-European, Bantu and Austronesian languages (Bentz et al. 2015). However, phylogenetic signals for external factors are generally unknown. This study uses language family trees from Glottolog, WALS, Autotyp and Ethnologue, as published in Dediu (2015), as well as information on population sizes, latitude, longitude and altitude to clarify how much phylogenetic signal there is in language external factors. Averaging across all the available phylogenetic trees for 21 language families (with more than 20 languages) we find that there is a consistent cline of phylogenetic signals: latitude > longitude > altitude > population size.

This has two main implications: 1) geographic properties of languages have – across the board – higher phylogenetic signals than population properties, and are hence more likely to be preserved over time. This suggests that deep processes of language change might be most strongly reflected in geographic patterns. As a consequence, 2) controlling for non-independence of languages is more important when predicting linguistic structure from geographic data than from population size data.

Atkinson, Q. D. (2011). Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science*, 332, 346–349.

Bentz, C. (2016). The Low-Complexity-Belt: Evidence For Large-scale Language Contact In Human Prehistory? In S.G. Roberts, C. Cuskley, L. McCrohon, L. Barceló-Coblijn, O. Fehér & T. Verhoef (eds.) *The Evolution of Language: Proceedings of the 11th International Conference (EVOLANG11)*. Available online: http://evolang.org/neworleans/papers/93.html

Bentz, C., Verkerk, A., Kiela, D., Hill, F., & Buttery, P. (2015). Adaptive communication: Languages with more non-native speakers tend to have fewer word forms. *PLoS ONE*, 10 (6), e0128254. doi: 10.1371/journal.pone.0128254

Bentz, C. and Winter, B. (2013). Languages with more second language learners tend to lose nominal case. *Language Dynamics and Change* 3, 1-27.

Dediu, D. (2015). Language family classifications as Newick trees with branch length [database and software tool]. *GitHub,* online at https://github.com/ddediu/lgfam-newick

Everett, C., Blasi, D., & Roberts, S. (2015). Climate, vocal folds, and tonal languages: Connecting the physiological and geographic dots. *Proceedings of the National Academy of Sciences* 112.1322-1327.

Lewis, M. & Frank, M. (submitted). Linguistic niches emerge from pressures at multiple timescales.

Münkemüller, T., Laveregne, S., Bzeznik, B., Dray, S., Jombart, T., Schiffers, K., & Thuiller, W. (2012). How to measure and test phylogenetic signal. *Methods in Ecology and Evolution*, 3, 743-756.

Revell, L. J., Harmon, L. J., & Collar, D. C. (2008). Phylogenetic signal, evolutionary process, and rate. *Systematic Biology*, 57 (4), 591-601.

Verkerk, A. (2014). Diachronic change in Indo-European motion event encoding. Journal of Historical Linguistics, 4 (1), 40-83.