Variation in the spatial diffusion of structural aspects of language: A case study

Miri Mertner
University of Tübingen

Linguists have long speculated that some aspects of language structure are more prone to borrowing than others. Morphology is thought to be more resistant to contact than syntax and phonology (Nichols, 1992), though this may only hold for specific morphological features (Greenhill et al., 2010; Dunn et al., 2011). The mechanisms which cause this kind of variation, and how universal it is across families and areas, remains to be understood. This is complicated by the fact that understanding stability requires an understanding of language contact and replacement due to language-internal and external factors (Nichols, 2017).

Computational methods provide a promising way to tackle this question. Bayesian phylogenetic methods have already provided valuable insights into this topic, elucidating how stable structural features are within language lineages (Dediu and Cysouw, 2013; Dediu and Levinson, 2012; Dunn et al., 2011). Some methods incorporate the geographical locations of languages in order to control for and quantify the effect of language contact on aspects of language structure, which is a promising direction for understanding the related mechanisms that lead to structural stability in a language (Dediu and Levinson, 2012; Kauhanen et al., 2018; Murawaki and Yamauchi, 2018, for example).

The present study takes a new approach to the question of feature diffusibility using a Bayesian spatial model. The goal of this method is to infer the range within which different types of linguistic features show areal convergence, as well as the strength of that convergence. This will show whether some features are more prone to diffusion than others, providing an insight into the stability of features within different linguistic domains which could be useful for quantitative work on language phylogenetics. Additionally, it could shed light on past migrations and contact networks.

A limitation of the method is that it is computationally intensive, hence the present study will be a case study including all the languages of Africa for which sufficient data is available. As well as inferring the geographical range of different aspects of language structure within Africa, the model can also be used to find areas where languages have had particularly intensive or long-term contact with each other. Some of the areas that might be found include the Macro-Sudan belt, as described by Güldemann (2008, 2018), and the Rift Valley in Tanzania (Kießling, Mous, and Nurse, 2007).

This paper draws on previous work by Guzmán Naranjo and Mertner (2022) and Guzmán Naranjo, Mertner, and Urban (2023). This model builds on *multivAreate*, the model presented by Guzmán Naranjo, Mertner, and Urban (2023), which can control for phylogenetic relationships between languages and inter-feature correlations while inferring the areal signal of linguistic features. It can also impute missing data and incorporate information about feature universality using priors, a set of innovations also found in Ranacher et al. (2021). Unlike *multivAreate*, this model does not assume linguistic features to be spatially independent. They are assigned to one of the following groups: phonology, word order, nominal categories, and verbal categories. Each group is modelled using a single latent Gaussian process (GP). This means that information about the range and strength of the spatial signal is shared within groups, but not across them. Thus, systematic variation in the diffusibility of phonological features, word order, nominal categories, and verbal categories can be explored.

The phonological data is acquired from PHOIBLE (Moran and McCloy, 2019). For the morphosyntactic data, features from Grambank (Skirgård et al., 2023b,a) and a database based on WALS (Dryer and Haspelmath, 2013), curated by Mark Donohue (pers. comm), are used (Kalyan and Donohue, 2023). The model is coded in Stan (Carpenter et al., 2017) and and run in R. Results will be presented with reference to current literature, and some limitations and future prospects of the method will be discussed.

References

Carpenter, Bob, Andrew Gelman, Matthew Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell (2017). “Stan: A Probabilistic Programming Language”. In: *Journal of Statistical Software, Articles* 76.1, pp. 1–32.

Dediu, Dan and Michael Cysouw (2013). “Some Structural Aspects of Language Are More Stable than Others: A Comparison of Seven Methods”. en. In: *PLoS ONE* 8.1. Ed. by John P. Hart, e55009.

Dediu, Dan and Stephen C. Levinson (2012). “Abstract Profiles of Structural Stability Point to Universal Tendencies, Family-Specific Factors, and Ancient Connections between Languages”. en. In: *PLoS ONE* 7.9. Ed. by Alex Mesoudi, e45198.

Dryer, Matthew S. and Martin Haspelmath, eds. (2013). *WALS Online (v2020.3).* Zenodo.

Dunn, Michael, Simon J. Greenhill, Stephen C. Levinson, and Russell D. Gray (2011). “Evolved structure of language shows lineage-specific trends in word-order universals”. en. In: *Nature* 473.7345, pp. 79–82.

Greenhill, S J, Q D Atkinson, A Meade, and R D Gray (2010). “The shape and tempo of language evolution”. en. In: *Proceedings of the Royal Society B: Biological Sciences* 277.1693, pp. 2443–2450.

Güldemann, Tom (2008). “The ”Macro-Sudan belt”: Towards identifying a linguistic area in northern Sub-Saharan Africa”. In: *A linguistic geography of Africa, 151-185.* Ed. by Bernd Heine and Derek Nurse, pp. 151–185.

— (2018). “Language contact and areal linguistics in Africa”. In: *The Languages and Linguistics of Africa.* Ed. by Tom Güldemann, pp. 445–545.

Guzmán Naranjo, Matías and Miri Mertner (2022). “Estimating areal effects in typology: A case study of African phoneme inventories”. In: *Linguistic Typology* 27.2.

Guzmán Naranjo, Matías, Miri Mertner, and Matthias Urban (2023). *Contact and diffusion with multinomial probit models.* Presentation at the 56th Annual Meeting of the Societas Linguistica Europaea.

Kalyan, Siva and Mark Donohue (2023). “The Dimensions of Morphosyntactic Variation: Whorf, Greenberg and Nichols were right”. In: *Linguistic Typology at the Crossroads* 3.2, pp. 132–190.

Kauhanen, Henri, Deepthi Gopal, Tobias Galla, and Ricardo Bermúdez-Otero (2018). “Geospatial distributions reflect rates of evolution of features of language”. en. In: *arXiv:1801.09637 [cond-mat, physics:nlin, physics:physics].* arXiv: 1801.09637.

Kießling, Roland, Maarten Mous, and Derek Nurse (2007). “The Tanzanian Rift Valley area”. In: *A Linguistic Geography of Africa.* Ed. by Bernd Heine and Derek Nurse. Cambridge Approaches to Language Contact. Cambridge: Cambridge University Press, pp. 186–227.

Moran, Steven and Daniel McCloy (2019). *PHOIBLE 2.0.*

Murawaki, Yugo and Kenji Yamauchi (2018). “A statistical model for the joint inference of vertical stability and horizontal diffusibility of typological features”. en. In: *Journal of Language Evolution* 3.1, pp. 13–25.

Nichols, Johanna (1992). *Linguistic diversity in space and time.* en. Chicago: University of Chicago Press.

— (2017). “Diversity and Stability in Language”. In: *The Handbook of Historical Linguistics.* Ed. by Brian D. Joseph and Richard D. Janda. John Wiley Sons, Ltd. Chap. 5, pp. 283–310.

Ranacher, Peter, Nico Neureiter, Rik van Gijn, Barbara Sonnenhauser, Anastasia Escher, Robert Weibel, Pieter Muysken, and Balthasar Bickel (2021). “Contact-Tracing in Cultural Evolution: A Bayesian Mixture Model to Detect Geographic Areas of Language Contact”. In: *Journal of the Royal Society Interface* 18.181, pp. 1–15.

Skirgård, Hedvig et al. (2023a). “Grambank Reveals Global Patterns in the Structural Diversity of the World’s Languages”. In: *Science Advances* 9 (16).

Skirgård, Hedvig et al. (2023b). *Grambank v1.0.* Version v1.0. Dataset.