A global search for linguistic areas using Bayesian mixture modelling

Anna Graff[1,2,3], David Inman[1,3], Natalia Chousou-Polydouri[1,3], Nico Neureiter[3,4], Peter Ranacher[3,4], Russell Barlow[5], Alena Witzlack-Makarevich[6], Russell Gray[5], Chiara Barbieri[1,2,3], Balthasar Bickel[1,3]

1 Department of Comparative Language Science, University of Zurich
2 Department of Evolutionary Biology and Environmental Studies, University of Zurich
3 Centre for the Interdisciplinary Study of Language Evolution (ISLE), University of Zurich
4 Department of Geography, University of Zurich
5 Department of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology
6 Department of Linguistics, The Hebrew University of Jerusalem

Keywords: Bayesian mixture model, linguistic areas, language contact, database comparison, dataset curation

Languages whose speakers stand in contact with one another tend to become more similar through borrowing of lexical, phonological and morphosyntactic features. *Linguistic areas* denote regions where such contact effects take place over prolonged periods such that multiple unrelated (or distantly related) languages share a variety of structural properties (e.g. Campbell, 2006). Defining such areas in a mathematically grounded, non-impressionistic way has long been a challenge, such that many proposed areas remain disputed in terms of their boundaries and which languages belong to them (Campbell, 2017). However, new tools provide objective approaches to determine areas: While an early Bayesian mixture model (Daumé III, 2009) identified areas sharing features beyond levels expected due to relatedness, a new model, sBayes (Ranacher et al., 2021), goes further by disentangling contact from both inheritance and universal tendencies.

Despite the availability of these tools, the question of linguistic diffusion has never been approached with a global, bottom-up procedure, in which the same criteria are applied universally to reveal where grammatical features bundle in statistically significant ways. Such an approach would address a common matter of dispute in areal linguistics: the use of global data consisting of a large and uniform set of linguistically independent features would reduce possible suspicions that detected areal signals represent an artifact of subjectively having cherry-picked a particular set of features to maximise specific suspected areal signals. This talk will present both methodological updates to sBayes and results from new global-scale implementations of the algorithm, thus representing a timely contribution to computational historical linguistics.

Methodologically, we improve sBayes by updating the model comparison method it employs. While the release version of sBayes employs the deviance information criterion (DIC) to identify the optimal number of areas given the input data, we now replace the DIC with Pareto-smoothed importance sampling leave-one-out cross-validation (PSIS-LOO) (Vehtari et al., 2017), a state-of-the-art alternative approach to Bayesian model comparison that makes fewer assumptions about the data than the DIC.

Additionally, we apply the updated algorithm on a global level, comparing results obtained using data from different linguistic databases. Our study will comprise both analyses using the global database Grambank (Skirgård et al., submitted), covering 195 logically independent morphosyntactic features for 2,430 languages, and analyses using over 300 logically independent features relating to morphosyntax, phonology and the lexicon from the databases AUTOTYP (Bickel et al., 2021), PHOIBLE (Moran and McCloy, 2019), WALS (Dryer and Haspelmath, 2013) and Lexibank (List et al., 2021) for over 4,000 languages. At this workshop, we will present our curated datasets as well as methods we have developed to handle sparse language-data matrices. Additionally, we will present first results of global areal analyses using sBayes.

References

Bickel, B., Nichols, J., Zakharko, T., Witzlack-Makarevich, A., Hildebrandt, K., Rießler, M., et al. (2021). *The AUTOTYP database*. Zenodo doi:10.5281/zenodo.4574513.

Campbell, L. (2006). "Areal Linguistics: A Closer Scrutiny," in *Linguistic Areas: Convergence in Historical and Typological Perspective*, eds. Y. Matras, A. McMahon, and N. Vincent (London: Palgrave Macmillan UK), 1–31. doi:10.1057/9780230287617_1.

Campbell, Lyle. 2017. Why is it so Hard to Define a Linguistic Area? In Raymond Hickey (ed.), The Cambridge handbook of areal linguistics, 19-39. Cambridge: Cambridge University Press. doi:10.1017/9781107279872.003

Daumé III, H. (2009). Non-Parametric Bayesian Areal Linguistics. *arXiv:0906.5114 [cs]*. Available at: http://arxiv.org/abs/0906.5114 [Accessed June 4, 2021].

Dryer, M. S., and Haspelmath, M. eds. (2013). *WALS Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology Available at: https://wals.info/ [Accessed June 4, 2021].

Skirgård, H., Haynie, H. J., Hammarström, H., Blasi, D. E., Collins, J., Latarche, J., et al. (submitted) Grambank data reveal global patterns in the structural diversity of the world's languages.

List, J.-M., Forkel, R., Greenhill, S. J., Rzymski, C., Englisch, J., and Gray, R. D. (2021). Lexibank: A public repository of standardized wordlists with computed phonological and lexical features. In Review doi:10.21203/rs.3.rs-870835/v1.

Moran, S., and McCloy, D. eds. (2019). *PHOIBLE 2.0*. Jena: Max Planck Institute for the Science of Human History Available at: https://phoible.org/ [Accessed December 2, 2021].

Ranacher, P., Neureiter, N., Gijn, R. van, Sonnenhauser, B., Escher, A., Weibel, R., et al. (2021). Contact-tracing in cultural evolution: a Bayesian mixture model to detect geographic areas of language contact. *bioRxiv*, 2021.03.31.437731. doi:10.1101/2021.03.31.437731.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat Comput* 27, 1413–1432. doi:10.1007/s11222-016-9696-4.