PHYLOGENETIC MULTILEVEL MODELS REVEAL A SIMPLICITY BIAS IN URALIC

Yingqi Jing*[1], Miina Norvik[23], Outi Vesakoski[4], and Michael Dunn[1]

*Corresponding Author: yingqi.jing@lingfil.uu.se
[1]Department for Linguistics and Philology, Uppsala University, Uppsala, Sweden
[2]Department of Modern Languages, Uppsala University, Uppsala, Sweden
[3]Institute of Estonian and General Linguistics, University of Tartu, Tartu, Estonia
[4]Department of Finnish language and Finno-Ugric linguistics, University of Turku, Turku, Finland

One central topic in comparative linguistics is to investigate the evolutionary rates of typological traits between different levels of language systems. Varying rate of change could take place due to functional constraints such as general principles of efficiency in communication and learning (Zipf, 1949; Culbertson & Kirby, 2016; Hahn, Jurafsky, & Futrell, 2020), or language contact and second language acquisition that may affect the linguistic complexity (McWhorter, 2011; Lupyan & Dale, 2010; Bentz & Berdicevskis, 2016; Housen, De Clercq, Kuiken, & Vedder, 2019). Previous research on rate variation across domains is however unclear. Greenhill et al. (2017) and Carling and Cathcart (2021) show that grammatical features tend to evolve faster than lexical items (i.e., basic vocabulary) and morphological features at least in Austronesian and Indo-European, whereas Dediu and Cysouw (2013) show that syntactic features including word orders are more stable than morphological features, or there is no much rate difference across domains (Greenhill, Atkinson, Meade, & Gray, 2010). Newly available Uralic typological data allows us to estimate the rates of change across typological domains (Norvik et al., 2022) and understand the evolutionary dynamics of the Uralic language system.

With the publications of the Uralic Typological Database (Norvik et al., 2022), we use Bayesian phylogenetic inference to estimate the rates of change for different typological features in the history of Uralic. We go beyond the earlier approaches to fit phylogenetic models for each individual feature or estimate the average rates across all features. Instead, we introduce a novel multilevel phylogenetic Continuous-time Markov Chain model to investigate the evolutionary trajectories of 110 features across different domains (phonology, morphology and syntax) in 33 Uralic languages. The hierarchical model allows us to jointly infer the evolutionary rates at both population and group levels, guarding against overfitting and underfitting (Nalborczyk, Batailler, Lœvenbruck, Vilain, & Bürkner, 2019; Stan Development Team, 2021). We first validate the model with simulated

data and then apply our approach to Uralic typological data.

Our results reveal a slight directional change towards simple character states at the population level (mean rate of $q_{01} = 6.8$ and 95% CI = [0.92, 23]; mean rate of $q_{10} = 8.26$ and 95% CI = [0.76, 24.1]), suggesting that losing a complex feature on average takes around 250 years less time than gaining a complex one. In each domain, we also observe consistent simplicity biases, though the estimated rates of change are quite similar across different domains (mean rate in phonology: 8.1 and 95% CI = [0.79, 30.3]; mean rate in morphology: 7.29 and 95% CI = [0.52, 27.9]; mean rate in syntax: 8.6 and 95% CI = [0.47, 35]).

The evolutionary biases towards simplified states can be driven by the general economic principle, which reduces the linguistic complexity to facilitate communication (Zipf, 1949). The observed trends are quite consistent across domains, suggesting that forces of simplification are persistent in the whole Uralic language system. The simplification of language systems would also be expected in certain contact situations (e.g., imperfect learning or language shift), which may lead to the loss of complexity or redundancy (Heine & Kuteva, 2005; Grünthal et al., 2022). It is also worth noting that even though the differences in rate biases are slight, these evolutionary preferences can be amplified in a long period of language evolution or learning. Our results are not consistent with previous work that suggests unequal rates of change across domains in other languages families. Instead, we show that language evolves at very similar rates across typological domains in Uralic. Further research is needed to expand our approach to other families to see whether there is a constant rate constraint in the multilayer system of languages (Kroch, 1989; Kauhanen & Walkden, 2018).

References

Bentz, C., & Berdicevskis, A. (2016). Learning pressures reduce morphological complexity:: Linking corpus, computational and experimental evidence. In *26th international conference on computational linguistics* (pp. 222–232). Osaka, Japan.

Carling, G., & Cathcart, C. (2021). Reconstructing the Evolution of Indo-European Grammar. *Language*, *97*(3), 561–598.

Culbertson, J., & Kirby, S. (2016). Simplicity and specificity in language: Domain-general biases have domain-specific effects. *Frontiers in Psychology*, *6*(1964).

Dediu, D., & Cysouw, M. (2013). Some Structural Aspects of Language Are More Stable than Others: A Comparison of Seven Methods. *PLoS ONE*, *8*(1), e55009.

Greenhill, S. J., Atkinson, Q. D., Meade, A., & Gray, R. D. (2010). The shape and tempo of language evolution. *Proceedings of the Royal Society B: Biological Sciences*, *277*(1693), 2443–2450.

Greenhill, S. J., Wu, C.-H., Hua, X., Dunn, M., Levinson, S. C., & Gray, R. D.

(2017). Evolutionary dynamics of language systems. *Proceedings of the National Academy of Sciences*, 201700388.

Grünthal, R., Heyd, V., Holopainen, S., Janhunen, J., Khanina, O., Miestamo, M., Nichols, J., Saarikivi, J., & Sinnemäki, K. (2022). Drastic demographic events triggered the uralic spread. *Diachronica (to appear)*.

Hahn, M., Jurafsky, D., & Futrell, R. (2020). Universals of word order reflect optimization of grammars for efficient communication. *Proceedings of the National Academy of Sciences*, 201910923.

Heine, B., & Kuteva, T. (2005). *Language contact and grammatical change.* Cambridge University Press.

Housen, A., De Clercq, B., Kuiken, F., & Vedder, I. (2019). Multiple approaches to complexity in second language research. *Second language research*, *35*(1), 3–21.

Kauhanen, H., & Walkden, G. (2018). Deriving the Constant Rate Effect. *Natural Language & Linguistic Theory*, *36*(2), 483–521.

Kroch, A. S. (1989). Reflexes of grammar in patterns of language change. *Language Variation and Change*, *1*(3), 199–244.

Lupyan, G., & Dale, R. (2010). Language structure is partly determined by social structure. *PloS one*, *5*(1), e8559.

McWhorter, J. H. (2011). *Linguistic simplicity and complexity: Why do languages undress?* Walter de Gruyter.

Nalborczyk, L., Batailler, C., Lœvenbruck, H., Vilain, A., & Bürkner, P.-C. (2019). An introduction to bayesian multilevel models using brms: A case study of gender effects on vowel variability in standard indonesian. *Journal of Speech, Language, and Hearing Research*, *62*(5), 1225–1242.

Norvik, M., Jing, Y., Dunn, M., Forkel, R., Honkola, T., Klumpp, G., Kowalik, R., Metslang, H., Pajusalu, K., Piha, M., Saar, E., Saarinen, S., & Vesakoski, O. (2022). Uralic typology in the light of new comprehensive data sets. *Journal of Uralic Linguistics*, *1*(1), 4-42.

Stan Development Team. (2021). *RStan: the R interface to Stan.* (R package version 2.21.2)

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort.* Cambridge, MA: Addison-Wesley Press.