Evaluating the phylogenetic signal of morphosyntax

Ruby Sleeman, Elena Anagnostopoulou, Manolis Ladoukakis, Maria-Margarita Makri, Dimitris Michelioudakis, Christos Zioutis & Pavlos Pavlidis
(Goethe University Frankfurt and Institute for Mediterranean Studies - Foundation for Research & Technology Hellas (IMS-FORTH); University of Crete, Rethymno and IMS-FORTH; University of Crete, Iraklio; IMS-FORTH; Aristotle University of Thessaloniki; University of Vienna; University of Crete, Iraklio and FORTH-Institute of Computer Science)

Keywords: phylogenetics, historical linguistics, linguistic typology, Indo-European, morphosyntactic variation

Computational linguistic phylogenetics has so far relied heavily on cognate data, which have been extensively analyzed over the past decades and have produced phylogenies largely aligning with existing knowledge of language history (Bouckaert et al. 2012, Chang et al. 2015, Sagart et al. 2019, a.o.). In contrast, the potential of morphosyntactic characters as a valuable source of data for phylogenetic analysis has been largely overlooked. Such characters can provide insights into aspects that cognate data cannot address, especially with respect to genealogical/historical relationships beyond individual language families. Notably, however, recent studies (e.g. Longobardi et al. 2013) employing morphosyntactic characters have not reconstructed the phylogeny of Indo-European (IE) languages with sufficient accuracy and/or significant statistical support. In this study, we explore the usefulness of the World Atlas of Language Structures (WALS) data for reconstructing phylogenies, with a focus on IE languages. We constructed a table with 425 states of WALS features as (binary) taxonomic characters in 60 IE languages, providing our own values for >70% of the table's cells. It turns out that WALS-type data often contain a strong phylogenetic signal, but fail to yield a purely historical tree of IE. We evaluated the initial characters, which involved reformulating many of them on the basis of theoretical, historical and typological reasoning, and constructed a new table of 530 characters for the same IE languages. We then used this table to generate phylogenies. Although the resulting tree largely aligns with a cognate-based tree, consistent discrepancies are observed. We argue that these discrepancies arise from the quantity and quality of the data employed. While cognate data comprise a few thousand entries, morphosyntactic data are counted in hundreds (at best). Moreover, the morphosyntactic data currently employed for phylogenetic analysis lack qualitative filtering and contain elements prone to horizontal transfer or homoplasy, which obscure the underlying phylogenetic signal. To address these issues, we propose three novel methods that leverage both linguistic expertise and computational approaches to evaluate morphosyntactic data. The first one compares the feature bipartition with a "golden standard" language bipartition in order to identify the features with a high correlation between these two types of information. The second method uses a hill-climbing approach to explore subsets of features which produce a tree close to a commonly accepted phylogenetic tree. Finally, the third method evaluates each morphosyntactic feature using its parsimony score on a target tree.

This research was funded by the ERC ADG 2022 Project 101096554.

References

Bouckaert, Remco R., Philippe Lemey, Michael Dunn, Simon J. Greenhill, Alexander V. Alekseyenko, Alexei J. Drummond, Russell D. Gray, Marc A. Suchard, and Quentin D. Atkinson (2012), Mapping the origins and expansion of the Indo-European language family, *Science* 337(6097), 957-960.

Chang, Will, Chundra Aroor Cathcart, David P. Hall, and Andrew J. Garrett (2015), Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis, *Language* 91(1), 194-244.

Longobardi, G., Cristina Guardiano, Giuseppina Silvestri, Alessio Boattini, and Andrea Ceolin (2013), Toward a syntactic phylogeny of modern Indo-European languages, *Journal of Historical Linguistics*, 3(1), 122-152.

Sagart, Laurent, Guillaume Jacques, Yunfan Lai, Robin J. Ryder, Valentin Thouzeau, Simon J. Greenhill, and Johann-Mattis List (2019), Dated language phylogenies shed light on the ancestry of Sino-Tibetan, *Proceedings of the National Academy of Sciences of the United States of America*, 116(21), 10317-10322.