1 Introduction by the Organizers

Gerhard Jäger[1], Robert Forkel[2], Johann-Mattis List[2,3]

[1] Institute of Linguistics, University of Tübingen
[2] Dep. of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology Leipzig
[3] Chair of Multilingual Computational Linguistics, University of Passau

Computational approaches play an increasingly important role in mainstream historical linguistics. Along with these contributions, we note an increased need for standards which drive the curation and sharing of data in historical linguistics (annotated texts, wordlists, collections of structural data, information on phylogenies, etc.). While there have been attempts towards standardization in the past, most prominently reflected in the Cross-Linguistic Data Formats initiative (Forkel et al. 2018), which has been adopted by several teams working on computational and quantitative approaches in the field of historical linguistics, there are still many types of data for which no standards and examples of best practice exist, although they serve frequently as input or output of studies in historical linguistics (e.g. language phylogenies as collected in Greenhill's (2022) "Phlorest collection"). Considering in addition that many new data collections have been published lately (Dellert et al. 2020, List et al. 2022, Kaiping and Klamer 2018), it seems about time to consolidate and discuss which methods we have at our disposal in order to explore highly standardized collections of cross-linguistic data.

The workshop intends to bring together scholars from three different backgrounds: those who work actively on the development of new standards for cross-linguistic data in historical linguistics in particular and comparative linguistics in general, those who design new methods and workflows to explore and exploit standardized data, and those who conduct full-scale analyses of standardized data in order to address concrete scientific problems. The contributions to the workshop can be assigned to one of three key topics: (1) Standards for Cross-Linguistic Data in Historical Linguistics, (2) Methods and Analyses for the Exploitation of Standardized Cross-Linguistic Data, and (3) Research Questions Requiring New/Better Data. Contributions related to key topic (1) present existing standards for linguistic data that have not yet been introduced in historical linguistics, propose new standards for those cases in which standards are lacking, or discuss the role that standards could or should play in historical linguistics (their use, their limits). Contributions to key topic (2) present new methods by which standardized cross-linguistic data can be explored as well as new full-fledged analyses in which specific research questions are addressed by means of workflows that involve standardized cross-linguistic datasets. Contributions to key topic (3) initiate broader discussions on particular research questions that cannot yet be solved but might be solved in the future if sufficiently standardized cross-linguistic data would be available.

Greenhill, Simon J. (2022): Phlorest. https://github.com/phlorest

Dellert, Johannes, Thora Daneyko, Alla Münch, Alina Ladygina, Armin Buch, Natalie Clarius, Ilja Grigorjew, et al. 2020. "NorthEuraLex: A Wide-Coverage Lexical Database of Northern Eurasia." Language Resources & Evaluation54: 273–301. https://doi.org/10.1007/s10579-019-09480-6.

Forkel, Robert, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping, and Russell D. Gray. 2018. "Cross-Linguistic Data Formats, Advancing Data Sharing and Re-Use in Comparative Linguistics." Scientific Data5 (180205): 1–10.

Kaiping, Gereon A., and Marian Klamer. 2018. "LexiRumah: An Online Lexical Database of the Lesser Sunda Islands." PLOS ONE13 (10): 1–29. https://doi.org/10.1371/journal.pone.0205250.

List, Johann-Mattis, Robert Forkel, Simon J. Greenhill, Christoph Rzymski, Johannes Englisch, and Russell D. Gray. 2022. "Lexibank, a Public Repository of Standardized Wordlists with Computed Phonological and Lexical Features." Scientific Data9 (316): 1–31.