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.

2 Universal Dependency for Historical Languages (UD4HL): Towards Standardized Syntactic Data for Historical Languages

Luca Brigada Villa[1,2], Erica Biagetti[2], Chiara Zanchi[2], Silvia Luraghi[2]

[1] University of Bergamo
[2] University of Pavia

Over the past few decades, historical linguistic research has been enriched with the creation of treebanks for several ancient languages. Most developers have adopted the same annotation schemes employed for treebanks of modern languages, often choosing between the two de facto standards of the Penn Treebank phrase-structure format and the Prague Dependency Treebank (PDT) format. The PROIEL scheme (https://dev.syntacticus.org/proiel.html), which integrates Dependency Grammar with elements of Lexical Functional Grammar and was originally designed for a parallel treebank of translations of the Gospels in old Indo-European (IE) languages, has been applied to several other texts and is nowadays regarded as a further standard for the annotation of historical IE languages (Eckhoff et al. 2018). The multiplication of projects has led to an ever-growing number of historical treebanks that are incompatible with one another. As a result, new treebanks are created for languages that have already been annotated, but according to a different formalism from the one adopted by the authors. Recently, the annotation scheme designed within the Universal Dependency initiative (UD; Nivre et al. 2016, https://universaldependencies.org) has established itself as the standard for dependency annotation. As it favors comparative research, several constituency and dependency treebanks of ancient languages have been converted to UD (notably, we have no knowledge of dependency treebanks being converted to the Penn scheme), and others are now being developed according to this scheme. Yet the achievement of a comparable dataset for historical languages is still hampered by problems related to: a) coverage and balance of each sub-corpus, b) errors caused by the conversion process, and c) the absence of sufficiently clear and adequate guidelines for the annotation of historical languages.

In this paper, we present the state of the art, some issues and possible solutions to obtain corpora as representative as possible of historical languages. In order not to contribute to the flourishing of individual initiatives, we will open a UD working group dedicated to the annotation of such languages in UD: Universal Dependency for Historical Languages (UD4HL). In this group, we plan to address the following issues with the community. First, tools designed to convert the treebanks to the UD format, such as UDConverter (https://github.com/thorunna/UDConverter) and proiel-cli (https://github.com/proiel/proiel-cli), need to be further improved to produce cleaner outputs. Second, we aim to stimulate a revision process of both converted and native UD treebanks that tackles one construction type at a time (cf. Brigada Villa et al. 2022 , Biagetti et al. 2022): this will make it possible to fix errors caused by the conversion and to provide accurate and consistent guidelines for the annotation of new texts. Finally, the conllu format employed by UD features a MISC (miscellaneous) field that can be enriched with information that is not strictly syntactic but useful for studies on the syntax of historical languages, and is currently underexploited. We propose to add various types of information, such as e.g., metrical information for poetic texts or semantic information regarding the animacy of verbal arguments (PROIEL that had such information in its native format, but this has not been included in the UD converted treebanks). Findings and conclusions reached within the working group will be presented at the conference.

Biagetti, Erica, Chiara Zanchi and Francesco Mambrini. Universal Homeric Dependencies? Towards a complete and updated UD treebank of the Homeric poems. Delbrück Symposium on Indo-European Syntax. Università di Verona, November 9-12 2022.

Brigada Villa, Luca, Erica Biagetti and Chiara Zanchi (2022). Annotating “Absolute” Preverbs in the Homeric and Vedic Treebanks. Proceeding of the Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2022; Marseille, June 25, 2022).

Eckhoff, Hanne, Bech, Kristin, Eide, Kristine, Bouma, Gerlof, Haug, Dag T. T., Haugen, Odd E. & Jøhndal, Marius. 2018b. The PROIEL treebank family: A standard for early attestations of Indo-European languages. Language Resources and Evaluation 52(1): 29–65.

Nivre, Joakim, De Marneffe, Marie-Catherine, Ginter, Filip, Goldberg, Yoav, Hajic, Jan, Manning, Christopher D., McDonald, Ryan, et al. 2016. Universal Dependencies V1: A Multilingual Treebank Collection. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16), 1659–66.

3 From Old Data to Fresh Phylogenies — A Linguistic Data Journey in the Times of CLDF

Christoph Rzymski[1]

[1] Dep. of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology Leipzig

Historical linguistics involves the study of language change over time, and is often aided by the use of cross-linguistic data. Cross-Linguistic Data Formats (CLDF, Forkel et al. 2018) provides a standardized way to represent and exchange such data, while cldfbench (Forkel & List 2020) is a workflow model that facilitates the management and analysis of CLDF data. In this study, we demonstrate how CLDF and cldfbench can be used to tackle commonplace tasks in historical linguistics, such as analyzing word lists to identify cognates and building phylogenies. By using CLDF as both input and output, we aim to show how these tools can help streamline the process of working with cross-linguistic data in historical linguistics, from the initial stage of collecting data from “old sources” (i.e., physical sources such as dictionaries and language documentation materials) to the final stage of constructing phylogenies that represent the relationships between languages.

We will demonstrate how to automatically compute cognates (List 2018, List 2021) in word lists using resources such as Concepticon (List 2022) and Glottolog (Hammarström 2022), and how to use these lists as input for BEAST (Bouckaert et al. 2014) to compute phylogenies. Since cldfbench supports a workflow that involves using “raw” source data and converting it to one or more CLDF datasets with the help of custom configurations and/or additional Python code, we aim to showcase how this can be utilized to prepare datasets for individual research questions. CLDF, cldfbench, and the aforementioned workflows can help researchers to efficiently process and analyze large amounts of data, and facilitate the integration of data from multiple sources.

Overall, our goal is to demonstrate the utility of CLDF and CldfBench for researchers in the field of historical linguistics, and to encourage their adoption as standard tools for handling cross-linguistic data. By showcasing innovative approaches to working with standardized cross-linguistic data, we hope to inspire new ideas and perspectives on how to build fresh phyologenies from “old data”.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.H., Xie, D., Suchard, M., Rambaut, A., & Drummond, A. (2014). BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. PLoS Computational Biology, 10(4), e1003537.

Forkel, R., List, J.M., Greenhill, S., Rzymski, C., Bank, S., Cysouw, M., Hammarstrom, H., Haspelmath, M., Kaiping, G., & Gray, R. (2018). Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. Scientific data, 5(1), 1–10.

Forkel, R., & List, J.M. (2020). CLDFBench. Give your Cross-Linguistic data a lift. In Proceedings of the Twelfth International Conference on Language Resources and Evaluation (pp. 6997-7004). European Language Resources Association (ELRA).

Hammarström, H., Forkel, R., Haspelmath, M., & Bank, S. (2022). CLLD Glottolog 4.7. Max Planck Institute for Evolutionary Anthropology. https://glottolog.org.

List, J. M., Walworth, M., Greenhill, S. J., Tresoldi, T., & Forkel, R. (2018). Sequence comparison in computational historical linguistics. Journal of Language Evolution, 3(2), 130-144.

List, J.M, & Forkel, R. (2021). LingPy. A Python library for historical linguistics. Version 2.6.9. URL: https://lingpy.org, DOI: https://zenodo.org/badge/latestdoi/5137/lingpy/lingpy. With contributions by Greenhill, Simon, Tresoldi, Tiago, Christoph Rzymski, Gereon Kaiping, Steven Moran, Peter Bouda, Johannes Dellert, Taraka Rama, Frank Nagel. Leizpig: Max Planck Institute for Evolutionary Anthropology.

List, J.M., Tjuka, A., Rzymski, C., Greenhill, S., & Forkel, R. (2022). CLLD Concepticon 3.0.0. Max Planck Institute for Evolutionary Anthropology. https://concepticon.clld.org.

4 Phlorest: A Database of Consistent and Reusable Language Phylogenies

Robert Forkel[1], Simon Greenhill[2]

[1] Dep. of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology Leipzig
[2] School of Psychology, University of Auckland, Auckland

The last few decades have seen the publication of many language phylogenies. These phylogenies have proven to be incredibly powerful tools for making inferences about language relationships (e.g. Gray, Drummond, and Greenhill 2009; Kolipakam et al. 2018; Remco R. Bouckaert, Bowern, and Atkinson 2018; Chang et al. 2015; Greenhill et al. 2022), or as a backbone for testing hypotheses about language change (e.g. Dunn et al. 2011), linguistic reconstructions (e.g. Carling and Cathcart 2021), and evolutionary processes (e.g. Greenhill et al. 2017). Often the results of these phylogenetic studies are repurposed by other researchers to test other hypotheses Watts et al. (2016). Or the results themselves are controversial e.g. witness the arguments about the age of Indo-European Chang et al. (2015) or the debates about language universals Dryer (2011).

We therefore need good ways for researchers to obtain, inspect, compare them, and reuse these phylogenies. However, to date this re-use is hard, often requiring detailed phylogenetic knowledge to identify the relevant files, understand their formats, and extract the critical information. Phlorest is a database of published language phylogenies that aims to standardise the outputs of these analyses to make them Findable, Accessible, Interoperable, and Reusable (Wilkinson et al. 2016). Phlorest collects published language phylogenies into a single database in a consistent and easily usable format (CLDF, Forkel et al. 2018). Currently, Phlorest contains 42 phylogenies, covering a total of 4266 varieties from 2172 languages. Each analysis is preprocessed to a consistent format, providing a summary tree and a posterior tree sample, linked where possible to the raw data. Each taxon in the analysis is mapped to catalogues like Glottolog (https://glottolog.org) and D-PLACE (https://d-place.org/) so that users can readily identify which languages were included in each analysis.
In this talk we will present Phlorest and discuss the benefits it provides. First, phlorest enables replicability and reuse of these trees. Second, having these phylogenies aligned in time and space enables us to compare patterns and processes across the globe. Third, phlorest allows us to scale up to bigger questions by combining trees into super trees. Finally, phlorest allows us to highlight interesting big picture findings from historical linguistics to the wider public, providing a highly visible resource that brings this research to a wider audience.

Bouckaert, Remco R., Claire Bowern, and Quentin D. Atkinson. 2018. “The Origin and Expansion of Pama-Nyungan Languages Across Australia.” Nature Ecology & Evolution. https://doi.org/10.1038/s41559-018-0489-3.

Bouckaert, Remco R, et al.. 2012. “Mapping the Origins and Expansion of the Indo-European Language Family.” Science 337 (6097): 957–60. https://doi.org/10.1126/science.1219669.

Carling, Gerd, and Chundra Cathcart. 2021. “Reconstructing the Evolution of Indo-European Grammar.” Language 97 (3): 561–98. https://doi.org/10.1353/lan.2021.0047.

Chang, Will, Chundra Cathcart, David Hall, and Andrew Garrett. 2015. “Ancestry-Constrained Phylogenetic Analysis Supports the Indo-European Steppe Hypothesis.” Language 91 (1): 194–244. https://doi.org/10.1353/lan.2015.0005.

Dryer, Matthew S. 2011. “The Evidence for Word Order Correlations.” Linguistic Typology 15: 335–80. https://doi.org/10.1515/LITY.2011.024.

Dunn, Michael, Simon J. Greenhill, S. C. Levinson, and Russell D. Gray. 2011. “Evolved Structure of Language Shows Lineage-Specific Trends in Word-Order Universals.” Nature 473 (7345): 79–82. https://doi.org/10.1038/nature09923.

Forkel, Robert, et al. 2018. “Cross-Linguistic Data Formats, Advancing Data Sharing and Re-Use in Comparative Linguistics.” Scientific Data 5 (1): 180205. https://doi.org/10.1038/sdata.2018.205.

Gray, Russell D., Alexei J Drummond, and Simon J. Greenhill. 2009. “Language Phylogenies Reveal Expansion Pulses and Pauses in Pacific Settlement.” Science 323 (5913): 479–83. https://doi.org/10.1126/science.1166858.

Greenhill, Simon J., Hannah J. Haynie, Robert M Ross, Angela M. Chira, Johann-Mattis List, Lyle Campbell, Carlos A. Botero, and Russell Gray. 2022. “A Recent Northern Origin for the Uto-Aztecan Family,” August. https://doi.org/10.31235/osf.io/k598j.

Greenhill, Simon J., Chieh-Hsi Wu, Xia Hua, Michael Dunn, Stephen C. Levinson, and Russell D. Gray. 2017. “Evolutionary Dynamics of Language Systems.” Proceedings of the National Academy of Sciences, 201700388. https://doi.org/10.1073/pnas.1700388114.

Kolipakam, Vishnupriya, Fiona M Jordan, Michael Dunn, Simon J Greenhill, Remco Bouckaert, Russell D Gray, and Annemarie Verkerk. 2018. “A Bayesian Phylogenetic Study of the Dravidian Language Family.” Royal Society Open Science 5: 171504. https://doi.org/10.1098/rsos.171504.

Levy, Roger, and H Daumé III. 2011. “Computational Methods Are Invaluable for Typology, but the Models Must Match the Questions.” Linguistic Typology 15: 393–99. https://doi.org/10.1515/LITY.2011.026.

Watts, Joseph, Oliver Sheehan, Quentin D. Atkinson, Joseph Bulbulia, and Russell D. Gray. 2016. “Ritual Human Sacrifice Promoted and Sustained the Evolution of Stratified Societies.” Nature 532 (7598): 228–31. https://doi.org/10.1038/nature17159.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. “The FAIR Guiding Principles for Scientific Data Management and Stewardship.” Scientific Data 3 (1): 160018. https://doi.org/10.1038/sdata.2016.18.

5 Configurable Language-Specific Tokenization for CLDF Databases

Johannes Dellert[1], Verena Blaschke[2]

[1] Institute of Linguistics, University of Tübingen
[2] Center for Information and Language Processing, LMU Munich

In any workflow for computational historical linguistics, tokenization of IPA sequences is a crucial preprocessing step, as it shapes the alignments which provide the input of algorithms for cognate detection and proto-form reconstruction. This is also true for EtInEn (Dellert 2019), our forthcoming integrated development environment for etymological theories. An EtInEn project can be created from any CLDF database such as the ones that have been aggregated and unified by the Lexibank initiative (List ea. 2022). Whereas the tools for preparing CLDF databases (Forkel & List 2020) encourage the application of a uniform tokenization across all languages in a dataset, our view is that in many contexts, it is more natural to tokenize phonetic sequences in ways that differ between languages. To provide a simple example, many geminates in Italian need to be aligned to consonant clusters in other Romance languages (e.g. notte vs. Romanian noapte "night"), which is much easier if they are tokenized into two instances of the same consonant, whereas geminates in Swedish are best treated as cognate to their shortened counterparts in other Germanic languages.

To provide comprehensive support for such cases, EtInEn includes configurable language-specific tokenizers as an additional abstraction layer that allows to reshape forms after the import, and also serves as a generic way to bridge phonetic surface forms and the underlying forms that historical linguists are primarily interested in. Each tokenizer is defined by a token alphabet which is used for greedy tokenization, a list of allophone sets which can be used to abstract over irrelevant subphonemic distinctions, and a list of non-IPA symbols that are defined in terms of phonetic features. The initial state of each tokenizer is based on an analysis of the tokens used by the imported CLDF database. Tokenizer definitions are stored in a human-editable plain-text format which we would like to propose as a new standard.

In EtInEn, tokenizer definitions are manipulated through a graphical editor in which the potential tokens for each language are arranged in the familiar layout of consonant and vowel charts, enhanced by additional panels for diphthongs and tones. Currently defined tokens are highlighted, and allophone sets are summarized under their canonical symbols. Basic edit operations serve to group several sounds into an allophone set, and to join or split a multi-symbol sequence, such as a diphthong or a sound with a coarticulation. More complex operations support workflows for parallel configuration of multiple tokenizers.

Additional non-IPA symbols can be given semantics in terms of a combination of phonetic features, and declared to be part of the token set for any language. On the representational level, this provides the option to use non-IPA symbols for form display, whereas underlyingly, the system will interpret the symbols in terms of their features. On the conceptual level, underspecified definitions provide support for metasymbols. In addition to some predefined metasymbols (such as V for vowels and C for consonants), the user can assign additional symbols to arbitrary classes of sounds. These are then available throughout EtInEn for various purposes, such as concisely representing the conditioning environments for a soundlaw, or summarizing the probabilistic output of an automated reconstruction module.

In addition to configurable tokenizers, EtInEn provides the option to define form-specific tokenization overrides, allowing to substitute the result of automated tokenization with any sequence over the current token alphabet for the relevant language. This is currently our strategy for handling otherwise challenging phenomena such as metathesis or root-pattern morphology, which we normalize into alignable and concatenative representations. This forms a bridge to existing standards for representing morphology in the CLDF framework (e.g. Schweikhard & List 2020), which currently only support the annotation of morpheme boundaries in terms of simple splits in phonetic IPA sequences.

Dellert, Johannes (2019): "Interactive Etymological Inference via Statistical Relational Learning." Workshop on Computer-Assisted Language Comparison at SLE-2019.

Forkel, Robert and Johann-Mattis List (2020): "CLDFBench. Give your Cross-Linguistic data a lift." Proceedings of LREC 2020, 6997-7004.

List, Johann-Mattis, Robert Forkel, S. J. Greenhill, Christoph Rzymski, Johannes Englisch & Russell Gray (2022): "Lexibank, A public repository of standardized wordlists with computed phonological and lexical features." Scientific Data 9.316, 1-31.

Schweikhard, Nathanael E. and Johann-Mattis List (2020): “Developing an annotation framework for word formation processes in comparative linguistics.“ SKASE Journal of Theoretical Linguistics 17(1), 2-26.

6 A computational evaluation of regularly recurring sound correspondences

Frederic Blum[1], Johann-Mattis List[1,2]

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

Regularly recurring sound correspondences are the main tools of the comparative method (Anttila 1972; Lass 1997). The cognate judgements which are based on these correspondences are also used in the phylogenetic approaches to historical linguistics that have received widespread attention in recent years (Greenhill et al. 2020). However, regularity is often more an intuitive notion than a quantified evaluation, and irregularity is argued to be more common than expected from the Neogrammarian hypothesis (Durie & Ross 1996; Labov 1981). Given the recent devel- opment of computational methods in historical linguistics and the availability of cross-linguistic comparative formats (Forkel et al. 2018; List 2019), we are now able to improve our workflows in this regard.

We provide a computational machinery that can be used as a means to improve the anno- tation of cognates in a standardized data set. For this, we focus on a quantitative measure for assessing the regularity of sound correspondences across cognates. This can, for example, be used to compare the results of different automated methods of cognate judgements and align- ments, or to identify possible errors in expert cognate annotations. Our workflow proceeds in four stages. In the first stage, we carry out a phonetic alignment analysis (List et al. 2018) of all cognate sets in a standardized wordlist. In the second stage, we preprocess the phonetic alignments by excluding spurious alignment sites (columns in a multiple phonetic alignment). In the third stage, we search for recurring correspondences across our aligned cognate sets and determine potentially regular correspondence patterns. In a fourth stage, we score the overall regularity of the individual cognate sets in our data by counting how many sites in the align- ments can be represented by recurring (regular) correspondence patterns, and how many are unique.

In the talk, we showcase the functionality of this workflow using data from the Pano-Tacanan language family. We will focus on two key issues: the automated detection of potential false positive cognate judgements, as well as the detection of potential false negatives. Potential false positives are identified as words in a cognate set with very low regularity in the correspondence patterns across the data set. For the detection of potential false negatives, we compare two different sets of cognate annotations of the same data. If no second expert annotation is available, the first annotation can be compared to an automated judgement of cognacy (List 2019). We identify all cognate words above a custom regularity threshold that are assigned different cognacy in the first set of annotations, but are part of the same cognate set in the second annotation. We show how different thresholds influence the results and discuss possible further applications and developments of this workflow.

Anttila, Raimo. 1972. An Introduction to Historical and Comparative Linguistics. New York: The Macmillan Company.
Durie, Mark & Malcolm Ross. 1996. The Comparative Method Reviewed: Regularity and Irregularity in Language Change. New York, Oxford: Oxford University Press.
Forkel, Robert, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping & Russell D. Gray. 2018. Cross-linguistic data formats, advancing data sharing and re-use in comparative linguistics. Scientific Data 5(1). https://doi.org/10.1038/sdata.2018.205.
Greenhill, Simon J., Paul Heggarty & Russell D. Gray. 2020. Bayesian phylolinguistics. In Barbara S. Vance Richard D. Janda Brian D. Joseph (ed.), The Handbook of Historical Linguistics, chap. 11, 226–253. Wiley. https://doi.org/10.1002/9781118732168.ch11.
Labov, William. 1981. Resolving the Neogrammarian Controversy. Language 57(2). 267–308. https://doi.org/10.2307/413692.
Lass, Roger. 1997. Historical linguistics and language change. Cambridge: Cambridge University Press.
List, Johann-Mattis. 2019. Automatic inference of sound correspondence patterns across multiple languages. Computational Linguistics 45(1). 137–161. https://doi.org/10.1162/coli_a_ 00344.
List, Johann-Mattis, Mary Walworth, Simon J. Greenhill, Tiago Tresoldi & Robert Forkel. 2018. Sequence comparison in computational historical linguistics. Journal of Language Evolution 3(2). 130–144. https://doi.org/10.1093/jole/lzy006.

7 Exploring the Geographical Distribution of Missing Data Using Approximate Gaussian Processes

Miri Mertner[1], Matías Guzmán Naranjo[1]

[1] Institute of Linguistics, University of Tübingen

Gaussian processes (GPs) have several qualities that make them well-suited to spatial statistics, as they allow us to add non-linear effects to a model in a flexible way (see e.g. McElreath, 2020, Chapter 14, for an explanatory example). A GP essentially estimates the effect that every observation has on every other observation in the form of a covariance matrix, which can then be used, for example, as a predictor in a model. In linguistic typology, they have been used as a way to control for spatial autocorrelation between languages, as well as for inferring probable ranges of contact between languages (Guzmán Naranjo & Mertner, 2022). However, they can be prohibitively slow to use with large datasets, such as the global sample of languages included in WALS or Glottolog. Therefore, in order to use them on such large datasets, an approximation of the GP is required.

One of the cases in which a large dataset is necessary to make meaningful inferences is in the exploration of the distribution of missing data in linguistic databases such as WALS (Dryer & Haspelmath, 2013) and ASJP (Wichmann et al., 2022). Using approximate GPs implemented in the programming language Stan (Stan Development Team), the present study will focus on uncovering areal biases in the distribution of missing linguistic data. Geographical and social correlates which could help explain the causal factors behind a higher or lower density of missing data in a particular area will also be tested, such as landscape roughness, climate, and population size. A better understanding of the factors which lead to geographical imbalances in the distribution of missing data could, among other things, improve our ability to impute missing data as part of statistical modelling work.

Dryer, Matthew S. & Haspelmath, Martin (eds.) 2013. The World Atlas of Language Structures Online. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info, accessed on 2023-01-01.)

Guzmán Naranjo, Matías and Mertner, Miri. 2022. Estimating areal effects in typology: a case study of African phoneme inventories. Linguistic Typology. https://doi.org/10.1515/lingty-2022-0037

McElreath, R. (2020). Statistical Rethinking: A Bayesian Course with Examples in R and Stan (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781315372495

Stan Development Team. 2022. Stan Modeling Language Users Guide and Reference Manual, 2.31. https://mc-stan.org

Wichmann, Søren, Eric W. Holman, and Cecil H. Brown (eds.). 2022. The ASJP Database (version 20).