Computer-assisted approaches in historical and typological language comparison

Johann-Mattis List
Max Planck Institute for the Science of Human History, Jena

Abstract: The workshop invites papers that deal with *computer-assisted* (as opposed to pure computational or pure qualitative) approaches to historical and typological language comparison. Computer-assisted approaches are hereby understood as procedures involving different stages of qualitative *and* quantitative data analysis, ranging from the initial preparation of lexical or structural data, via automatic or manual annotation, up to qualitative or quantitative analysis, that yield a specific result, be it a linguistic reconstruction system linking proto-forms to aligned reflexes, a phylogeny that lists inferred word histories, or tools for exploratory data analysis. By focusing on *computer-assisted approaches*, we hope to foster a more intensive collaboration between classical and computational linguists. In addition to detailed descriptions of concrete tasks in historical and typological language comparison, we also encourage submissions dealing with data standards enhancing data sharing and reuse, as well as the presentation of purely qualitative approaches for which no computational solutions exist so far.

Keywords: computational historical linguistics, classical historical linguistics, comparative method, phylogenetic reconstruction, qualitative data analysis

Workshop description

By comparing the languages of the world historically, we can gain invaluable insights into human prehistory. By comparing them typologically, we can gain invaluable insights into the fundamentals of perception and cognition. The classical methods for historical and typological language comparison date back to the early 19th century and have been constantly refined and improved since then. Thanks to the comparative method for historical language comparison, linguists have made ground-breaking insights into language change in general and into the history of many specific language families in specific (Campbell and Poser 2008), and external evidence has often confirmed the validity of the findings (McMahon 2005). Thanks to large-scale approaches to typological comparison (Greenberg 1963; Dryer and Haspelmath 2013), we have gained many new insights into "universal" patterns recurring independently across the world's languages.

With increasing amounts of data, however, the methods to prepare, compare, and analyze data, which are largely based on manual labor, reach their practical limits. As a result, scholars are now increasingly trying to automatize different aspects of the classical comparative method in

historical linguistics (Kondrak 2000; Prokić, Wieling, and Nerbonne 2009; List 2014), or to automatize the retrieval of typological information (Bender 2017; Malaviya et al. 2017). On the other hand, the last decade has seen a large number of attempts to analyze cross-linguistic data statistically, be it to uncover universal factors that shape linguistic diversity independently of language history (Everett et al. 2015; Blasi et al. 2016), to gain insights into the past of specific language families (Bouckaert et al. 2012; Chang et al. 2015), to understand the dynamics underlying lexical and grammatical evolution (Greenhill et al. 2017), or to arrive at a better understanding of areal factors in language history (Cathcard et al. 2018).

Purely computational applications, however, are not capable of replacing experts' experience and intuition, and given that most of the computational methods for data preparation still largely lag behind human judgments, it is not surprising that most of the computational analyses still rely on manually annotated data. In a situation where computers cannot replace experts and experts do not have enough time to analyze the increasing amounts of data, a new framework, neither completely computer-driven, nor ignorant of the help computers provide, becomes urgent. Such frameworks are well-established in biology and translation, where computational tools cannot provide the accuracy needed to arrive at convincing results, but do assist humans to digest large data sets.

This has led to a situation in which computational methods can only be carried out by a small number of experts who have a strong background in programming. Since computational experts do not necessarily always have a strong background or interest in linguistic topics, this has lead to a certain split in the field, with classical linguists being often dismissive and sceptical with respect to computer-based applications, and computational linguists being unsatisfied with the lack of interest in the multiple opportunities which quantitative and digital approaches have to offer.

That both classical and computational analyses could profit from each other has been increasingly demonstrated in *computer-assisted frameworks* in which classical linguists collaborate closely with computational linguists, with the data being analyzed both qualitatively and quantitatively (List 2016). In these frameworks, computational methods can be used in various ways to assist experts in qualitative analysis, be it (1) by pre-processing large datasets automatically before having experts manually correct the results (Hill and List 2017), (2) by visualizing large datasets in a convenient way that allows experts for a quick inspection (List et al. 2018; List 2017), (3) or by using automatic methods to check expert annotations for internal consistency (Kolipakam et al. 2018).

What is important for a successful application of computer-assisted methods are the detailed *workflows* that experts use to retrieve and analyze information both quantitatively and

qualitatively. Since they usually require a complicated mixture of programming using different software packages, data annotation using different formats, and statistical analysis using different models, computer-assisted approaches are not (yet) easy to apply, especially for scholars with little experience in programming or data handling. What further exacerbates the more widespread sharing and reuse of computational and computer-assisted approaches that have been proposed in the past is that the information provided in the articles that discuss them is usually very sparse.

By bringing together scholars from the classical and the computational camps, we hope to foster a closer future collaboration that integrates both quantitative and qualitative approaches. Topics for papers include (but are not limited to):

- Computer-assisted approaches to study language contact in specific linguistic areas.
- Computer-assisted approaches to study language history in form of networks of phylogenies.
- Papers discussing the compilation of large annotated datasets in historical linguistics and language typology.
- Workflows for linguistic reconstruction (phonology, lexicon, syntax).
- Tools for exploratory data analysis in historical linguistics and language typology.
- Standards and best practices for data curation and reuse.
- Qualitative workflows and computer-assisted cases studies.
- Presentation of linguistic problems for which only qualitative workflows exist so far.

References

Bender, Emily. 2017. “Linguistic Typology in Natural Language Processing.” *Linguistic Typology* 20 (3): 645–60.

Blasi, Damián E., Søren Wichmann, Harald Hammarström, Peter Stadler, and Morten H. Christiansen. 2016. “Sound–meaning Association Biases Evidenced Across Thousands of Languages.” *Proceedings of the National Academy of Science of the United States of America* 113 (39): 10818–23.

Bouckaert, Remco, Philippe Lemey, Michael Dunn, Simon J. Greenhill, Aalexander 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): 957–60.

Campbell, Lyle, and William John Poser. 2008. *Language Classification: History and Method*. Cambridge: Cambridge University Press.

Cathcart, Chundra, Gerd Carling, Niklas Johansson, and Erich Round. 2018. “Areal Pressure in Grammatical Evolution.” *Diachronica* 35 (1): 1–34.

Chang, Will, Chundra Cathcart, David Hall, and Andrew Garret. 2015. “Ancestry-Constrained Phylogenetic Analysis Ssupport the Indo-European Steppe Hypothesis.” *Language* 91 (1): 194–244.

Dryer, Matthew S., and Martin Haspelmath, eds. 2013. *WALS Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info/.

Everett, C., D. E. Blasi, and S. G. Roberts. 2015. "Climate, Vocal Folds, and Tonal Languages: Connecting the Physiological and Geographic Dots." *Proceedings of the National Academy of Sciences of the United States of America* 112 (5): 1322–7.

Greenberg, Joseph Harold. 1963. *The Languages of Africa*. Bloomington: Indiana University Press.

Greenhill, S. J., C. H. Wu, X. Hua, M. Dunn, S. C. Levinson, and R. D. Gray. 2017. "Evolutionary Dynamics of Language Systems." *PNAS* 114 (42): E8822–E8829.

Hill, Nathan W., and Johann-Mattis List. 2017. "Challenges of Annotation and Analysis in Computer-Assisted Language Comparison: A Case Study on Burmish Languages." *Yearbook of the Poznań Linguistic Meeting* 3 (1): 47–76.

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): 1–17.

Kondrak, Grzegorz. 2000. "A New Algorithm for the Alignment of Phonetic Sequences." In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, 288–95.

List, Johann-Mattis. 2014. *Sequence Comparison in Historical Linguistics*. Düsseldorf: Düsseldorf University Press.

———. 2016. "Computer-Assisted Language Comparison: Reconciling Computational and Classical Approaches in Historical Linguistics." Jena: Max Planck Institute for the Science of Human History.

———. 2017. "Using Network Models to Analyze Old Chinese Rhyme Data." *Bull. Chin. Ling.* 9 (2): 218–41.

List, Johann-Mattis, Simon J. Greenhill, Cormac Anderson, Thomas Mayer, Tiago Tresoldi, and Robert Forkel. 2018. "CLICS². An Improved Database of Cross-Linguistic Colexifications Assembling Lexical Data with Help of Cross-Linguistic Data Formats." *Linguistic Typology* 22 (2): 277–306.

Malaviya, Chaitanya, Graham Neubig, and Patrick Littell. 2017. "Learning Language Representations for Typology Prediction." *ArXiv E-Prints*, no. 1707.09569: 1–7. https://arxiv.org/abs/.

McMahon, April. 2005. "Special Issue: Quantitative Methods in Language Comparison - Introduction." *Transactions of the Philological Society* 103 (2): 113–19.

Prokić, Jelena, Martijn Wieling, and John Nerbonne. 2009. "Multiple Sequence Alignments in Linguistics." In *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education*, 18–25.