Swadesh spaces. Deep learning in phylogenetic linguistics.

Gerhard Jäger, Seminar für Sprachwissenschaft, University of Tübingen, Germany

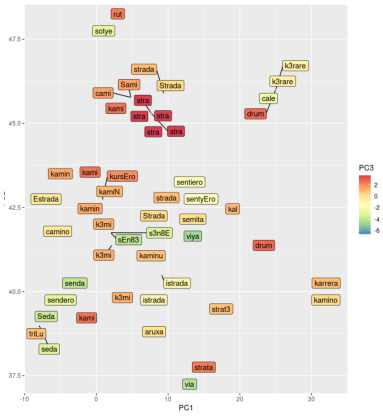
keywords: phylogenetics, deep learning, vector space embedding

Computational historical linguistics has benefitted immensely from the recent advances in applied statistics and machine learning. Bayesian phylogenetic inference (as exemplified by Bouckaert et al. 2012 or Bowern and Atkinson 2012) has become an indespensible scaffold for reconstructing prehistoric language change processes (e.g., Haynie and Bowern 2016, Carling and Cathcart 2021) and statistical explorations in typology while controlling for non-independences of languages due to common descent (Verkerk et al. 2021, Jäger and Wahle 2021). At the same time, various clustering and alignment techniques from machine learning have been deployed for the task of *cognate detection* (List et al. 2018, among many others).

The data types and methodologies which are common in computational historical linguistics differ markedly from those used in neighboring fields though. Linguists predominantly deal with *discrete* data, be it character sequences, cognate classes, phylogenetic trees, or typological feature matrices. The currently most fertile lines of research in machine learning, and in data science in general, focus on *continuous* data though. This especially stark in the context of Deep Learning, where information is uniformly represented as high-dimensional vectors. It also applies to Bayesian statistics though, where one of the most advanced software package, *Stan* (Carpenter et al. 2017), which discourages work with discrete data types.

The focus on discrete data is not just unusual in the wider field of data science, it is often sub-optimal also because not all relevant information is represented adequately. For instance, classifying words into cognate classes brushes many phonological and morphological phenomena that could inform, e.g., phylogenetic inference.

In the talk I will present a programm for representing, processing and analysing comparative and historical language data in high-dimensional vector spaces. This change of perspective opens up a plethora of powerful techniques for the field.



In a pilot study to be presented at the conference, this general program is exemplified. The crucial algorithmic tools are adopted from neural machine translation. I used a sequence-to-sequence architecture consisting of a *encoder* and a *decoder*, which are both based on a Recurrent Neural Network (a GRU, to be precise). The encoder takes a string of sound symbols as input and encodes it in a 256-dimensional numerical vector. The decoder takes such a vector as input and generates a string. The architecture was trained as an *auto-encoder*. This means that the decoder recovers the input string from a vector representation that was produced by the encoder. This architecture was trained with the Swadesh lists from the *Automated Similarity Judgment Program* (ASJP; Wichmann et al. 2020). This resulted in a vector representation for each ASJP entry. Since the string representation can be (almost perfectly) recovered from the vector representation, it can be said that the vectors contain the same information as the strings.

To illustrate the structure of the resulting vector space, a Principial Component Analysis for Romance words for *path* was conducted. The results are shown the the Figure above. It can be seen that different cognate classes correspond to different major regions, while sound change is represented by smaller spatial displacement.

Obvious routes of enquiry in the future include (1) cognate detection via Gaussian Mixtures, (2) ancestral state reconstruction, cognate prediction and missing value imputation via a combination of multivarate inference and the deep decoder, and (3) modeling of language contact via Gaussian Processes.

References

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

Claire Bowern and Quentin Atkinson. Computational phylogenetics and the internal structure of Pama-Nyungan. *Language*, 88(4):817–845, 2012.

Gerd Carling and Chundra Cathcart. Reconstructing the evolution of indo-european grammar: Supplementary material. *Language*, 97(3), 2021.

Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1):1–32, 2017.

Hanna J. Haynie and Claire Bowern. Phylogenetic approach to the evolution of color term systems. *Proceedings of the National Academy of Sciences of the United States of America*, 113(48):13666–13671, 2016.

Gerhard Jäger and Johannes Wahle. Phylogenetic typology. *Frontiers in Psychology*, 12, 2021. doi: doi.org/10.3389/fpsyg.2021.682132.

Johann-Mattis List, Mary Walworth, Simon J. Greenhill, Tiago Tresoldi, and Robert Forkel. Sequence comparison in computational historical linguistics. *Journal of Language Evolution*, 3(2):130–144, 2018.

Annemarie Verkerk, Hannah Haynie, Russell Gray, Simon Greenhill, Olena Shcherbakova, and Hedvig Skirgård. Revisiting typological universals with Grambank. paper presented at the DGfS annual meetig, February 2021.

Søren Wichmann, Eric W. Holman, and Cecil H. Brown. The ASJP database (version 20). `http://asjp.clld.org/`, 2020.