Proto-word reconstruction with NNs

1 Work to build on

- Reversible-jump MCMC methods Bouchard-Côté et al. (2013)
- Conditional random fields & RNN Ciobanu and Dinu (2018)
- RNNs Meloni et al. (2019)

2 Data sources

- Wiktionary
  - Many datapoints of dubious quality
  - Would have to do most of the data extraction ourselves
  - Data dumps available here
  - Technically it should be possible possible to get RDF data
- Indo-European lexical cognacy database
  - Used by famous Bouckaert et al. (2012)
  - No longer maintained (since 2016)
  - Only Indo-European data, which may be a bit over-investigated
  - But: plain TSV
- Evolution of human language project (used in Hruschka et al. (2015)). provides cognate data for several Eurasian language families (Altaic, Tungusic, Mongolic, Japonic...)
  - I didn't know the format (dBase/.dbf), don't know exactly how to use
  - Somewhat outdated (2013)
  - Pro: Many languages from many families
  - Could try to reconstruct proto-Altaic (which is a deprecated clade) or proto-Transeurasian

3 Model architectures

- Code letters for phonological features
  - Word = $n_{letters} \times n_{features}$ array

$$\begin{Bmatrix} l/f & \pm cons & \pm vow & \pm cont & \pm front & .. \\ t & 1 & 0 & 0 & 1 & \end{Bmatrix}$$

References

Bouchard-Côté, A., Hall, D., Griffiths, T. L., and Klein, D. (2013). Automated reconstruction of ancient languages using probabilistic models of sound change. *Proceedings of the National Academy of Sciences*, 110(11):4224–4229.

Bouckaert, R., Lemey, P., Dunn, M., Greenhill, S. J., Alekseyenko, A. V., Drummond, A. J., Gray, R. D., Suchard, M. A., and Atkinson, Q. D. (2012). Mapping the origins and expansion of the indo-european language family. *Science*, 337(6097):957–960.

Ciobanu, A. M. and Dinu, L. P. (2018). Ab Initio: Automatic Latin Proto-word Reconstruction. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1604–1614, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Hruschka, D., Branford, S., Smith, E., Wilkins, J., Meade, A., Pagel, M., and Bhattacharya, T. (2015). Detecting Regular Sound Changes in Linguistics as Events of Concerted Evolution. *Current Biology*, 25(1):1–9.

Meloni, C., Ravfogel, S., and Goldberg, Y. (2019). Ab Antiquo: Proto-language Reconstruction with RNNs. *arXiv:1908.02477 [cs]*. arXiv: 1908.02477.