# Imputing typological values via phylogenetic inference

Gerhard Jäger

Tübingen University

The 2020 Conference on Empirical Methods in Natural Language Processing

*SIGTYP 2020: The Second Workshop on Computational Research in Linguistic Typology*

`Shared Task: Prediction of Typological Features`

November 19, 2020

# SIGTYP 2020 Shared Task

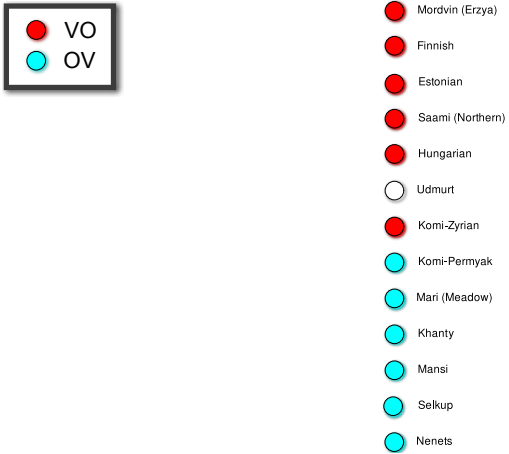
- **given:** subset of World Atlas of Language Structures (WALS; Dryer and Haspelmath, 2013):
  - 45,944 typological language-feature-value triplets
  - from 1,357 languages
  - covering 185 features
- **task:** infer the unkown values for a given set of
  - 2,410 language-feature-value triplets
  - from 149 of these languages and
  - 178 of these features
- `https://sigtyp.github.io/st2020.html`

## simplifying assumptions

- only vertical transmission of feature values
- features are mutually independent

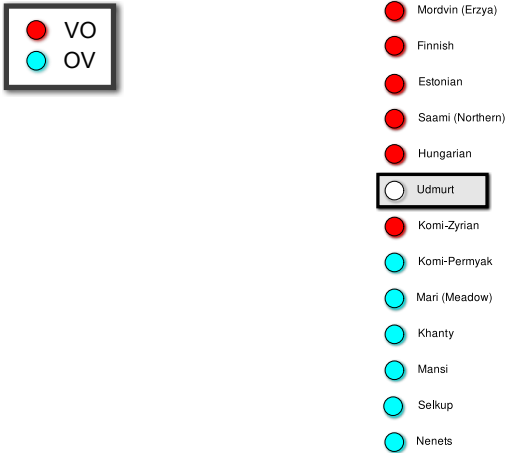
**How far can we get with this?**

# Basic idea



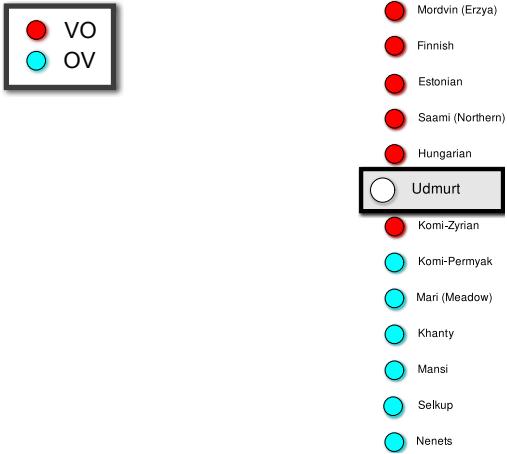

- missing value for a certain feature in a certain language
- feature value in genetically related languages is known

# Basic idea



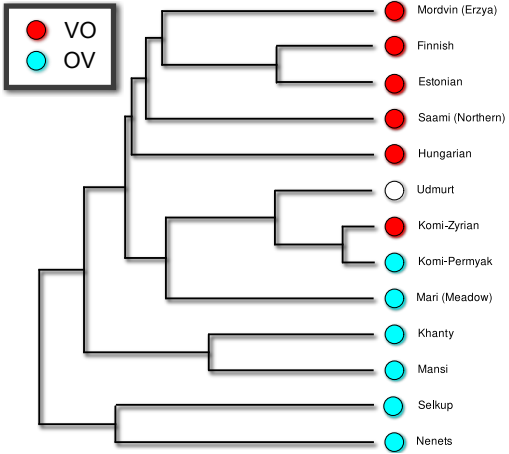

- missing value for a certain feature in a certain language
- feature value in genetically related languages is known

# Basic idea



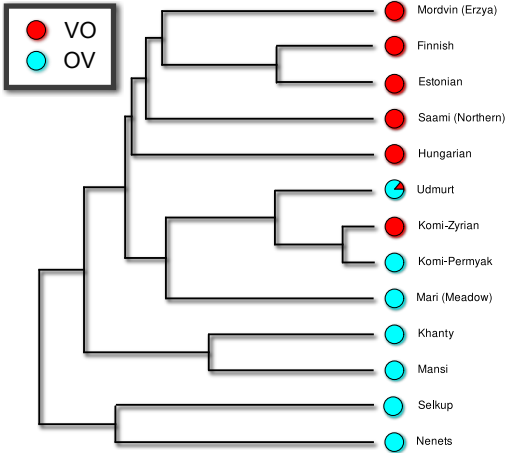

- missing value for a certain feature in a certain language
- feature value in genetically related languages is known

# Basic idea




- missing value for a certain feature in a certain language
- feature value in genetically related languages is known
- use family tree

# Basic idea




- missing value for a certain feature in a certain language
- feature value in genetically related languages is known
- use family tree
- interpolate feature value from related languages

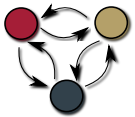
# Continuous time Markov chains on a tree

- continuous time Markov process (CTMC) with discrete state space
- characterized by $Q$-matrix, e.g.

$$Q = r \begin{pmatrix} -2 & 1 & 1 \\ 1 & -2 & 1 \\ 1 & 1 & -2 \end{pmatrix}$$

- Here: **Jukes-Cantor model** (originally developed for DNA evolution)
  - all rates are equal
  - global rate $r$ (expected number of mutations per unit of time) as parameter to be estimated

**Markov process**



# Continuous time Markov chains on a tree

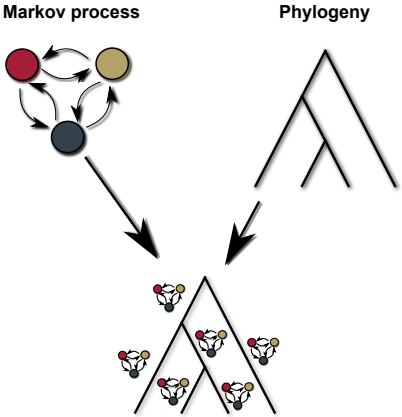

- (unordered) tree $\mathcal{T}$ with branch lengths
- Here:
  - inferred from lexical data
  - branch lengths represent amount of lexical change

# Continuous time Markov chains on a tree

- phylogenetic CTMC
- independent copies of CTMC on each branch of the tree
- likelihood of a branch of length $t$ with rate $r$, states $a$ and $b$ and top and bottom and $n$ possible states:

$$P(b|a;t,r) = \frac{1}{n}\begin{cases} 1+(n-1)e^{-tr} \text{ if } a=b \\ 1-e^{-tr} \text{ else} \end{cases}$$

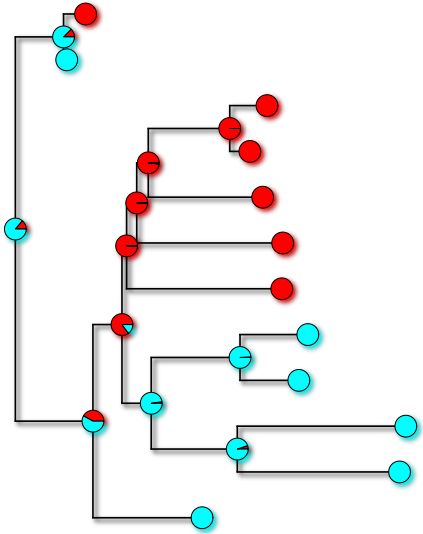
- likelihood of entire tree is product of branch likelihoods
- unkown states are marginalized out
- marginal likelihood can be efficiently computed via dynamic programming (bottom-up recursion through the tree, cf. Felsenstein, 2004)



# Ancestral state reconstruction

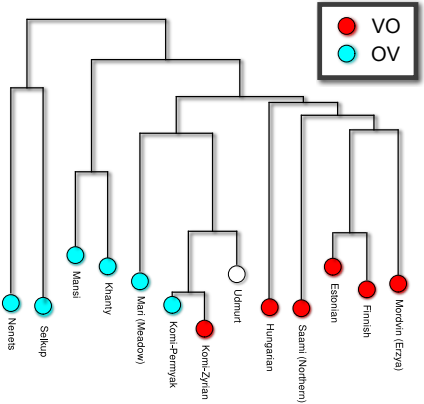
- usually states at the tips are observed
- ancestral state can be inferred via Bayes' rule:

$$P(a|\mathrm{root}) = \frac{\mathcal{L}_{\mathrm{root}}(a)P(a)}{\sum_{b\in\mathrm{states}} \mathcal{L}_{\mathrm{root}}(b)P(b)}$$



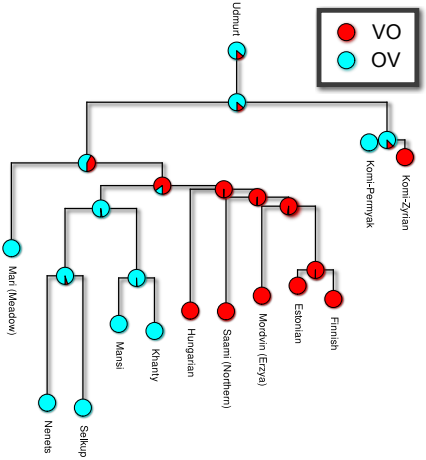
# Value imputation

- Jukes-Cantor model is **time reversible**



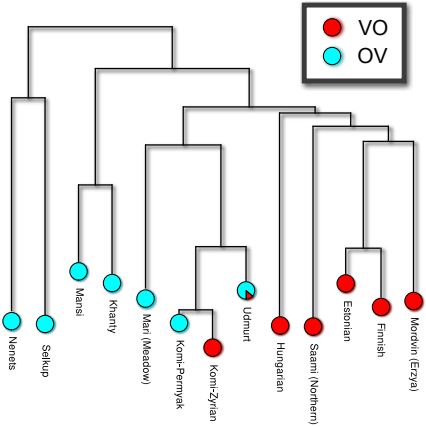
# Value imputation

- Jukes-Cantor model is **time reversible**
- rerooting a tree does not alter likelihood



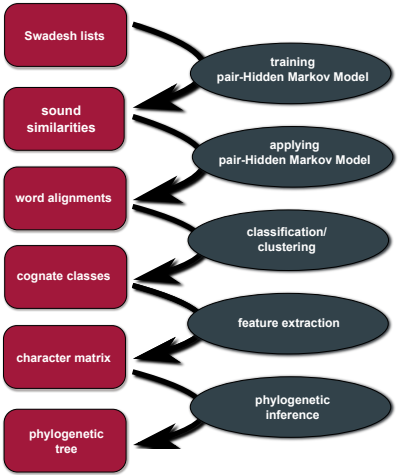
# Value imputation

- Jukes-Cantor model is **time reversible**
- rerooting a tree does not alter likelihood
- value imputation at tip can be done by placing the root of the tree at the tip in question

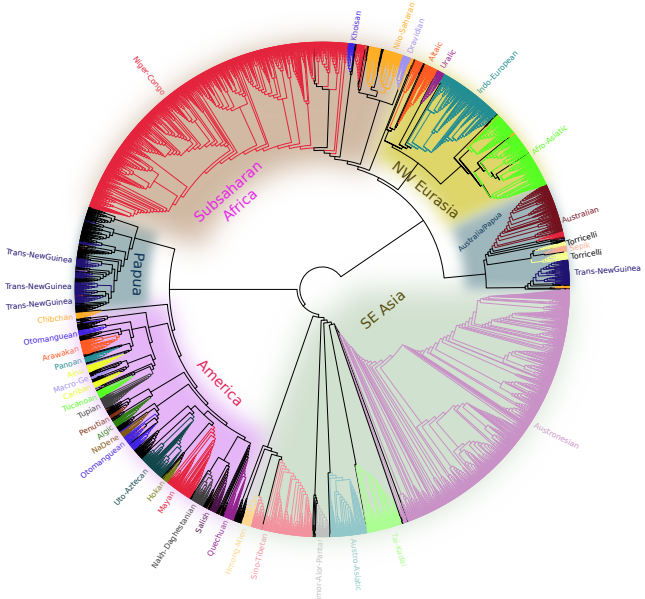


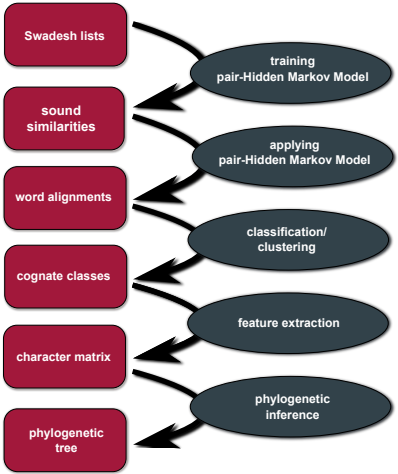
# Inferring a tree

- workflow from (Jäger, 2018)
- data taken from ASJP database (Wichmann et al., 2020)
- captures 7,346 languages and dialects
- overlap with Shared Task languages: 1,212 languages
- tree inference constrained by Glottolog classification (Hammarström et al., 2020)



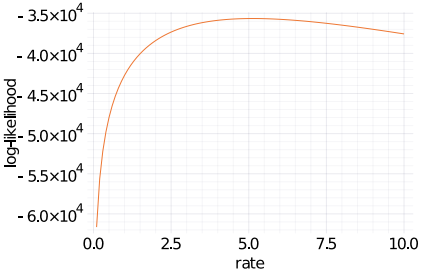
# Inferring a tree



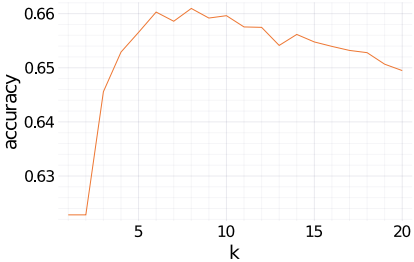
# Estimating the mutation rate

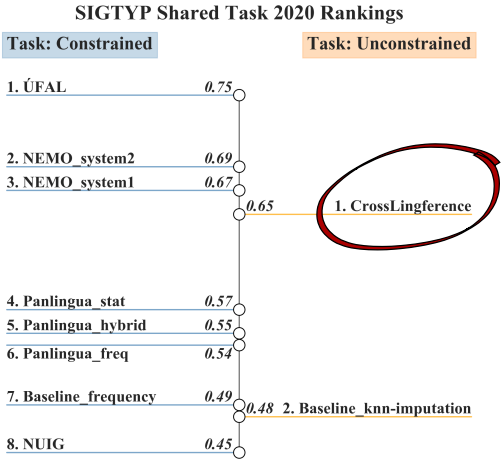
- mutation rate $r$ is *a priori* unknown
- assumption: all features evolve with same rate (unrealistic, but otherwise difficult to estimate due to data sparseness)
- ML estimation using all known feature values
- estimation: $r \approx 5.13$ (has no simple linguistic interpration)



# Geographical back-off

- 15 languages from test set have no counterpart in ASJP $\Rightarrow$ phylogenetic imputation not possible
- fallback option: $k$-nearest neighbor estimation based on geographical distance
- optimal value of $k$ estimated as 8 based on 20-fold cross-validation over known data




SIGTYP Shared Task 2020 Rankings
Task: Constrained
Task: Unconstrained
1. ÚFAL 0.75
2. NEMO_system2 0.69
3. NEMO_system1 0.67
0.65 1. CrossLingference
4. Panlingua_stat 0.57
5. Panlingua_hybrid 0.55
6. Panlingua_freq 0.54
7. Baseline_frequency 0.49
0.48 2. Baseline_knn-imputation
8. NUIG 0.45

# Conclusion

- imputation of typological feature values via phylogenetic inference achieves competitive performance, even though
  - cross-feature correlation and
  - language contact

  are ignored
- future work:
  - embedding of feature values into continuous high-dimensional space to incorporate cross-feature correlations
  - simultaneous control for phylogenetic and spatial autocorrelation

Matthew S. Dryer and Martin Haspelmath, editors. *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, 2013. `http://wals.info/`.

Joseph Felsenstein. *Inferring Phylogenies*. Sinauer Inc. Publishers, Sunderland, 2004.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. *Glottolog 4.2.1*. Max Planck Institute for the Science of Human History, Jena, 2020. doi: doi.org/10.5281/zenodo.3754591.

Gerhard Jäger. Global-scale phylogenetic linguistic inference from lexical resources. *Scientific Data*, 5, 2018. doi: 10.1038/sdata.2018.189.

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