Imputing typological values via phylogenetic inference

Gerhard Jäger
Tübingen University / SfS, Wilhelmstr. 19, 72074 Tübingen, Germany
gerhard.jaeger@uni-tuebingen.de

Abstract

This paper describes a workflow to impute missing values in a typological database, a subset of the World Atlas of Language Structures (WALS). Using a world-wide phylogeny derived from lexical data, the model assumes a phylogenetic continuous time Markov chain governing the evolution of typological values. Data imputation is performed via a Maximum Likelihood estimation on the basis of this model. As back-off model for languages whose phylogenetic position is unknown, a k-nearest neighbor classification based on geographic distance is performed.

1 Introduction

Precise knowledge of the typological diversity of natural languages is essential for several scientific disciplines. It provides clues about cognitive biases in language learning and processing, intrinsic tendencies in language change, a window into deep time regarding prehistoric population spread and population contact, as well as scaffolds for setting up NLP infrastructure for understudied languages.

While a growing body of digital data collections has become available in this regard (e.g., Dryer and Haspelmath, 2013; Bickel et al., 2018), these resources are sparse and skewed both with respect to the languages and the features covered. Given that obtaining high-quality typological qualification for underdocumented languages is a highly demanding task, this situation is likely to persist. It is therefore worthwhile to leverage statistical and machine learning methods to impute missing typological feature values from existing resources. Precisely this is the topic of the *2020 shared task*[1] of the ACL *Special Interest Group on Typology* (SIGTYP). The organizers made three subset of the WALS data (Dryer and Haspelmath, 2013) available, see Table 1. The task is to impute the $2,410$ missing values from the totality of known values.

[1] https://sigtyp.github.io/st2020.html, (Bjerva et al., 2020)

2 General approach

The present approach is based on two simplifying assumptions:

- The value of a typological feature in a given language is stochastically independent from the values of other features, and
- typological feature values are transmitted only vertically, i.e. from ancestor language to descendant language.

Both assumptions are known to be wrong. The first one is falsified by the manifold findings regarding *implicative universals* (established in Greenberg, 1963 and amply confirmed in subsequent typological research). The second assumption ignores the impact of *language contact* on typological properties. Research in areal linguistics has established beyond doubt that typological properties can be transmitted horizontally though (see, e.g., Campbell, 2006 for an overview). It seems worthwhile, however, to test how well a model based on these idealizations fares with regard to typological predictions. This establishes a baseline to be improved upon in future research.

The approach pursued here assumes that the phylogeny, i.e. the family tree representing their genealogical interrelatedness, including branch lengths reflecting the time between divergence events, of the languages under consideration is known. I used the techniques described in (Jäger, 2018) to infer such a phylogeny from lexical data (see next section for details).

Following much work in computational phylogenetics (see, e.g., Felsenstein, 2004, and Dunn

Proceedings of the Second Workshop on Computational Research in Linguistic Typology, pages 36–42

dataset	# languages	# features	# revealed values	# blinded values	release date
training	1,125	185	42,698	0	26 March 2020
development	83	182	3,246	0	26 March 2020
test	149	183	3,056	2,410	1 July 2020

Table 1: Subsets of WALS used in the Shared Task

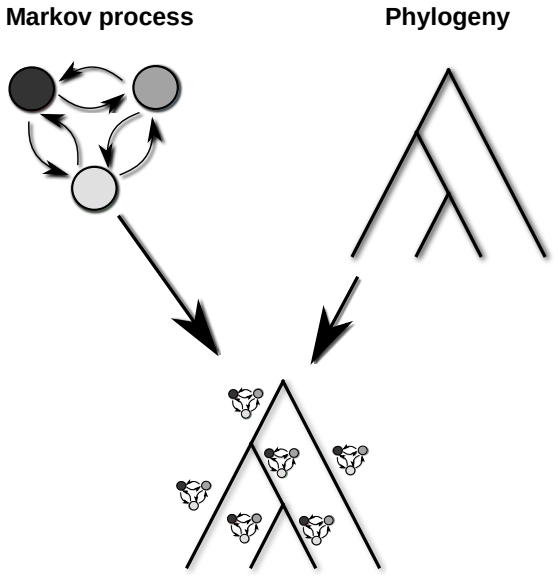

Figure 1: Phylogenetic CTMC

et al., 2011 for an application to linguistic typology), the model assumes that typological values evolve according to a *Continuous Time Markov Chain* process (CTMC). This means that a mutation, i.e., a change of the value of a typological feature, can occur at any time according to a probability density governed by a rate parameter. If a mutation occurs, another possible value of that feature is chosen according to a given probability distribution.

If a language splits into two daughter lineages, both daughter branches continue to evolve according to two independent copies of the mother language's CTMC. This is illustrated in Figure 1.

Let n be the number of values a given feature can assume. By way of a further simplification, I assume that the choice of a mutation target follows a uniform distribution. The probability of a language being in state b at the end of a time interval of length t if it was in state a at the beginning of the interval is given by (1) (known as the Jukes-Cantor model of molecular evolution, see for instance Felsenstein, 2004).

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

The parameter r ($r \in \mathbb{R}^+$) is the *rate* of evolution, i.e., the expected number of mutations per unit of time.

Suppose the phylogeny, the parameters of the CTMC and the states at the tips of the phylogeny are observed. It is then possible to compute the posterior probability distribution of the state at the root of the tree via a postorder (bottom-up) recursion through the phylogeny.

The likelihood of an observed state at a tip is 1, and the likelihood of all other states is 0. The likelihood of state a at an internal node α, conditional on the observed states at all tips descending from α, is given by the following equation, where $\mathrm{dr}(\alpha)$ are the immediate daughter nodes of α.

$$\mathcal{L}_\alpha(a) = \prod_{\beta \in \mathrm{dr}(\alpha)} \sum_{b \in \text{states}} P(a|b;t_{\alpha,\beta},r)\mathcal{L}_\beta(b),$$

Here $t_{\alpha,\beta}$ is the length of the branch from α to β.

Applying this equation recursively via postorder traversal through the phylogeny, we obtain the likelihood of the individual states at the root.

By assuming a uniform prior over states and applying Bayes' rule, the posterior probability of state a at the root of the phylogeny is

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

This algorithm is called *ancestral state reconstruction* (ASR). A detailed study of applications of ASR in linguistics for lexical evolution is given in (Jäger and List, 2018). An individual step of the recursion is graphically illustrated in Figure 2.

It is a convenient feature of the Jukes-Cantor model that it is *time reversible*. This means that the likelihood of a state at a node, given partial knowledge of the states at other nodes, remains

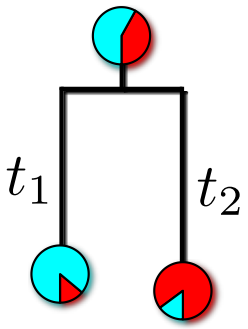

Figure 2: Elementary step of ASR

constant if the phylogeny is re-rooted and thereby the time arrow at some branches is reversed. Due to this property, it is possible to use ASR to impute unknown values at a tip of a tree. Suppose a phylogeny and the states of a feature at some, but not all, tips is known. To estimate the probability distribution over states at a particular tip α with missing value, the following steps are performed:

1. Prune the phylogeny by removing all tips with unknown value except α.
2. Reroot the tree so that α becomes the root.
3. Apply ASR.

This is graphically illustrated in Figure 3. The training set of the Shared Task contains the value of twelve Uralic languages for the feature *Order of Verb and Object*. The value for Udmurt (which is OV) is contained in the development set. Applying the described method (and using a known phylogeny and an estimated value for r; see next section), the posterior probability for OV in Udmurt, given the information in the training set, is ≈ 0.86. The maximum likelihood estimation of this feature value is therefore OV, which happens to be correct.

3 Data and methods

3.1 Phylogeny

In (Jäger, 2018) a method is described how to infer a world tree of languages from the 40-item Swadesh lists collected by the *Automated Similarity Judgment Project* (ASJP; Wichmann et al., 2020). The original paper used version 17 of ASJP. The results of applying precisely the same workflow to version 18 are made available at `https://osf.io/sdca4/` and were used here. More precisely, the phylogeny

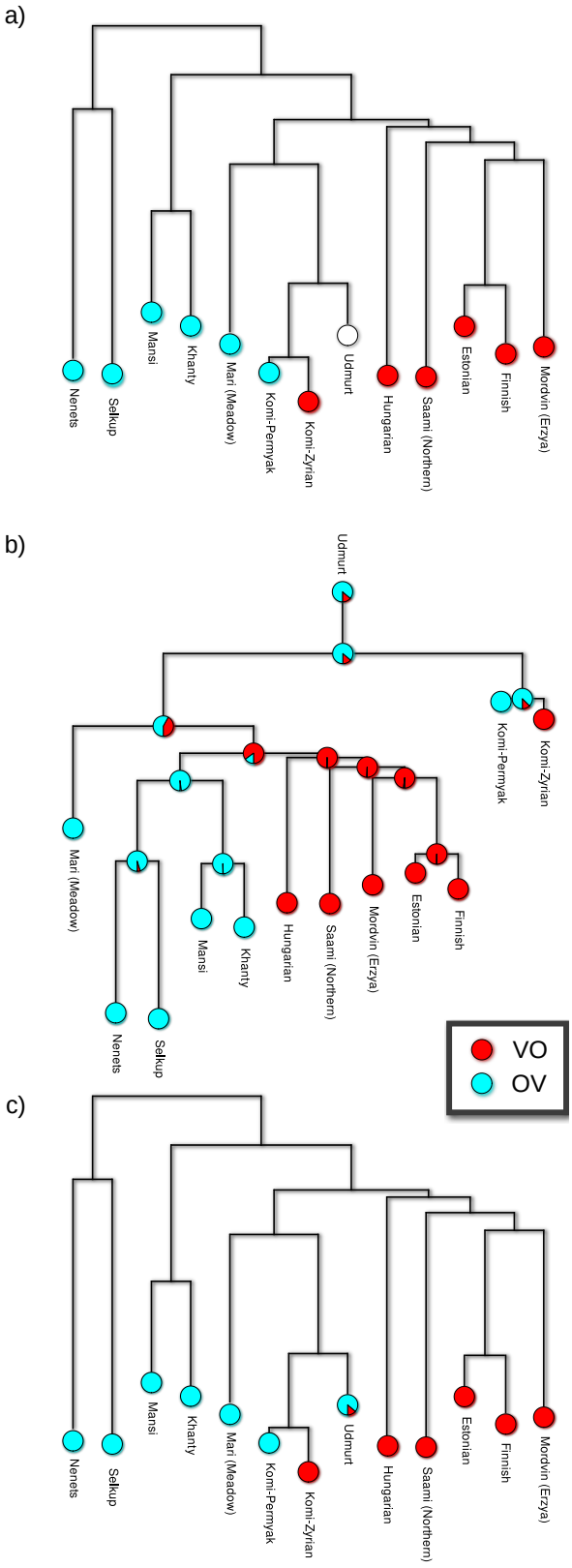

Figure 3: Value imputation via rerooting and ASR. a) Original phylogeny with missing value for Udmurt. b) Rerooting and recursive application of ASR. c) Resulting imputation.

`RAxML_bestTree.world_sc_ccGlot` was used, which is the result of using Maximum Likelihood phylogenetic inference over all characters and using the Glottolog classification (Hammarström et al., 2020) as constraint tree.

3.2 Matching WALS and ASJP

Among the 7,346 doculects for which the method described in the previous subsection provides cognate classification, only one doculect per glottocode was retained — generally the one with the least number of missing entries in ASJP. The metadata from WALS v.2020 were used to match glottocodes to WALS codes.

Among the 1,357 languages in the union of the three datasets from the Shared Task, 1,212 could be uniquely matched to an ASJP doculect with the same glottocode in this way.

For the 124 languages in the training set and the six languages in the development set for which no corresponding ASJP doculect could be identified in this way, I used the following procedure to define an ASJP proxy:

1. Use the closest geographic neighbor (according to great-circle distance, using the geographic coordinates supplied with the Shared Task data) within the same WALS genus among the 1,212 languages identified above.

2. If the language in question is a singleton within its genus, choose the closest geographic neighbor within its Glottolog family.

3. If the language is an isolate, choose the closest geographical neighbor.

The ASJP world tree was pruned to the 1,212 doculects which correspond to languages within the Shared Task data.

3.3 Phylogenetic value imputation

ASR, and therefore phylogenetic value imputation, depends on the rate parameter r. This value was estimated by maximizing the total marginal log-likelihood of all defined values within the training set under the CTMC model and the ASJP phylogeny. The log-likelihood as a function of r is shown in Figure 4. The Maximum-Likelihood estimation is $r \approx 5.13$.

Using this parameter estimate, the missing values from the test set for the 134 languages which directly correspond to an ASJP doculect were imputed. All known values from the training, development and test sets were used as input for the inference.

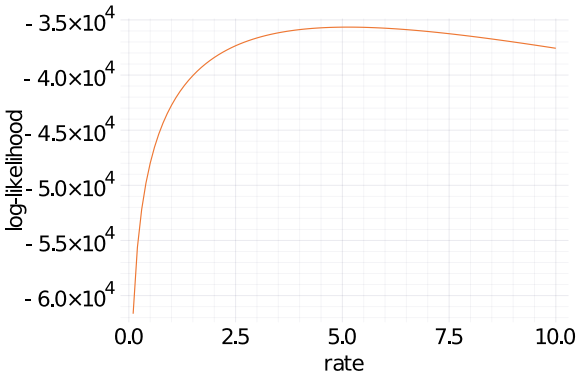

Figure 4: Log-likelihood of training set as a function of the mutation rate

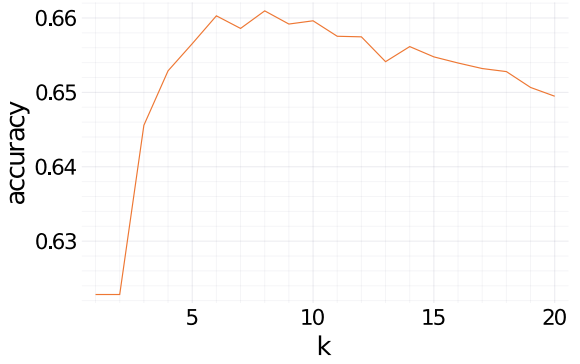

Figure 5: Average accuracy of geographic knn classification on the training set (20-fold cross-validation) as a function of k

3.4 Geographical k-nearest neighbor back-off

15 languages in the test set do not correspond to an ASJP doculect with the same glottocode. For the missing values from these languages, phylogenetic value imputation was therefore not possible. As back-off for these languages model I chose k-nearest neighbor based on geographical distances.

The value of k was estimated using 20-fold cross-validation over all feature-value pairs of the training set (see Figure 5). The optimal cross-validation accuracy was achieved at $k = 8$. Using this value, the missing values for the 15 test languages missing an ASJP counterpart were predicted via geographical knn-classification from the union of the training and the validation set.

4 Error analysis

To assess which factors influence the performance of the phylogenetic imputation method, I conducted a 20-fold cross-validation on all language-feature pairs for which the language corresponds to a tip in the ASJP phylogeny. This provides a dataset of 44,598 language-feature pairs for which both a goldstandard value and an inferred value are available. Not surprisingly, the accuracy of the predictions depend on the following three factors:

- Entropy of the feature value. The more values a feature can take and the more evenly the possible values are distributed, the harder it is to predict the correct value.
- Size of the language family. The inferred phylogeny used here is fairly reliable within language families but less so across families. Therefore predictions based on phylogenetically close languages is the more reliable the more close relatives a language has.
- Coverage of the feature. The more languages have a known value within the training data, the easier it is to impute a missing value in the test data.

The effect of these predictors is visualized in Figure 6.

To test whether each of those three factors are relevant given the other two, I conducted a Bayesian mixed-effects logistic regression with accuracy of prediction as dependent variable, feature entropy, log-transformed size of language family and log-transformed feature coverage as fixed effects, and language family and feature as random effects. The analysis was carried out using the R-package `brms` (Bürkner, 2017), which is based on the the programmning lanuage *Stan* (Carpenter et al., 2017).

The results are shown in Table 2.

It demonstrates that feature entropy has a credible negative effect, and both family size and feature coverage have a credible positive effect on accuracy.

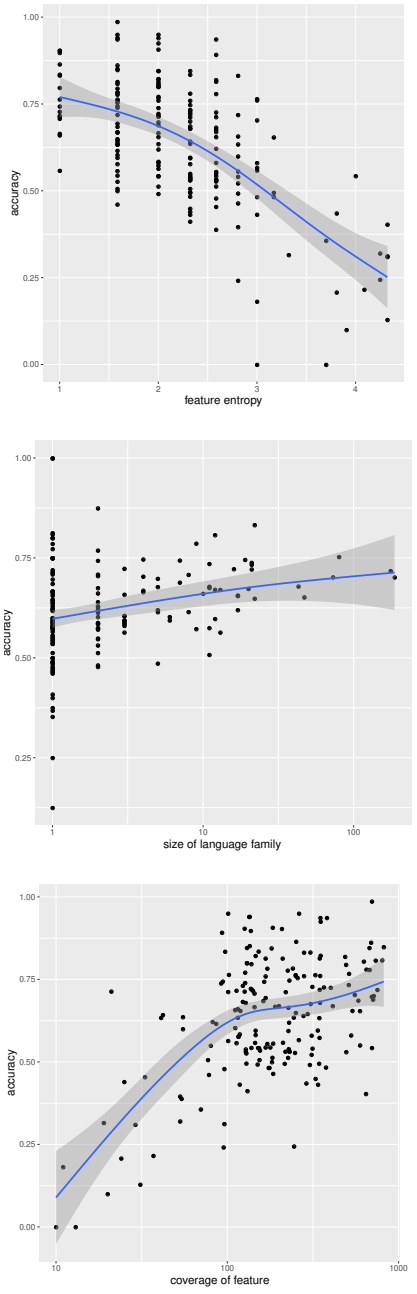

Figure 6: Impact of feature entropy, language family size and feature coverage on accuracy of phylogenetic imputation.

5 Results and discussion

According the evaluation script provided by the organizers of the Shared Task, this combination of phylogenetic and geographic knn value imputation achieved an overall accuracy of ≈ 0.68 — as compared to ≈ 0.51 both for the frequency baseline and the knn imputation baseline.

```
  Family: bernoulli
  Links: mu = logit
Formula: y ~ entropy + log10(featureFreq) + log10(famSize) + (1 | glotFam)
               + (1 | feature)
   Data: dat (Number of observations: 44598)
Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup samples = 4000

Group-Level Effects:
~feature (Number of levels: 185)
              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)     0.35      0.03     0.31     0.41 1.00     1456     2384

~glotFam (Number of levels: 164)
              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)     0.31      0.03     0.26     0.36 1.00     1376     2310

Population-Level Effects:
                Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept           0.68      0.24     0.21     1.16 1.00     1061     1849
entropy            -1.08      0.05    -1.17    -0.98 1.00     1408     2072
log10featureFreq    0.57      0.09     0.39     0.75 1.00     1029     1912
log10famSize        0.37      0.05     0.26     0.47 1.00     1056     1534

Samples were drawn using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

Table 2: Mixed-effects logistic regression

As pointed out in the Introduction, this model makes several simplifying assumptions. The most serious one is arguably the presumed independence of typological features. The phylogenetics literature contains techniques for dealing with correlated discrete features (Dunn et al., 2011; Pagel and Meade, 2006). However, applying this approach on a large scale would lead to an explosion of parameters that cannot be estimated with the sparse data available in WALS or similar data sources. Therefore it seems more promising in the long run to attempt an embedding of discrete typological features into a continuous high-dimensional space and combine this with phylogenetic ASR for continuous characters (using, e.g., a Brownian motion model of evolution). Such an approach can be combined with multivariate techniques like PCA to detect correlations between features.

Furthermore, a principled study of the interplay between vertical/phylogenetic and horizontal transmission mechanisms is called for to make further progress in the task of typological value imputation.

6 Data and code

Data and code are freely available at github.com/gerhardJaeger/emnlp2020.

Acknowledgments

Many thanks to the organizers of the Shared Task, and to three anonymous reviewers for their feedback. This research was supported by the DFG-Centre for Advanced Studies in the Humanities *Words, Bones, Genes, Tools* (DFG-KFG 2237) and by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement 834050).

References

Balthasar Bickel, Johanna Nichols, Taras Zakharko, Alena Witzlack-Makarevich, Kristine Hildebrandt, Michael Rießler, Lennart Bierkandt, Fernando Zúñiga, and John B. Lowe. 2018. The AUTOTYP database, release 0.1. https://github.com/autotyp/autotyp-data.

Johannes Bjerva, Elizabeth Salesky, Sabrina Mielke, Aditi Chaudhary, Giuseppe G. A. Celano, Edoardo M. Ponti, Ekaterina Vylomova, Ryan Cotterell, and Isabelle Augenstein. 2020. SIGTYP 2020 Shared Task: Prediction of Typological Features. In *Proceedings of the Second Workshop on Computational Research in Linguistic Typology*. Association for Computational Linguistics.

Paul-Christian Bürkner. 2017. Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, 10(1):395–411.

Lyle Campbell. 2006. Areal linguistics: A closer scrutiny. In Yaron Matras and April McMahon and-Nigel Vincent, editors, *Linguistic Areas*, pages 1–31. Springer.

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

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

Michael Dunn, Simon J. Greenhill, Stephen Levinson, and Russell D. Gray. 2011. Evolved structure of language shows lineage-specific trends in word-order universals. *Nature*, 473(7345):79–82.

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

Joseph Greenberg. 1963. Some universals of grammar with special reference to the order of meaningful elements. In *Universals of Language*, pages 73–113. MIT Press, Cambridge, MA.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2020. *Glottolog 4.2.1*. Max Planck Institute for the Science of Human History, Jena.

Gerhard Jäger. 2018. Global-scale phylogenetic linguistic inference from lexical resources. *Scientific Data*, 5.

Gerhard Jäger and Johann-Mattis List. 2018. Using ancestral state reconstruction methods for onomasiological reconstruction in multilingual word lists. *Language Dynamics and Change*, 8(1):22–54.

Mark Pagel and Andrew Meade. 2006. Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. *The American Naturalist*, 167(6):808–825.

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