Gerhard Jäger*

Model evaluation in computational historical linguistics

https://doi.org/10.1515/tl-2019-0020

Abstract: This is a reply to the comments by Hammarström et al. (This volume) and List (This volume) on the target article *Computational Historical Linguistics* (This volume). There I proposed several methodological principles for research in Computational Historical Linguistics pertaining to suitable techniques for model fitting and model evaluation. Hammarström et al. debate the usefulness of these principles, and List proposes a novel evaluation measure specifically aimed at the task of proto-form reconstruction. This reply will focus on the role of model evaluation in our field.

1 Introduction

A major motivation for writing the target article was to initiate a debate within the research community about standards of model fitting and model evaluation in Computational Historical Linguistics (CHL). I am grateful for and honored by the thoughtful comments it received, which deepened my understanding of issues involved. In the following, I will address some of specific points brought up by Hammarström et al. (Section 2) and List (Section 3) and conclude with some general considerations.

2 Reply to Hammarström et al.

The authors give a very careful and thought-provoking discussion of my methodological suggestions, as well as of the individual steps I took in the pilot study. Let me first focus on the general questions they raise, before I address some of the more specific points.

In the target article, I placed emphasis on the principle "**Separation of training and test data.** Different data sets are used for training and evaluating a model" (p. 156). Hammarström et al. (p. 235) comment on this:

***Corresponding author: Gerhard Jäger,** Seminar für Sprachwissenschaft, Universität Tübingen, Tübingen, Germany, E-mail: gerhard.jaeger@uni-tuebingen.de

> As we have outlined it here, separation of training and test data is a tool for assessing the quality of a predictive model. This fits well for Machine Learning since Machine Learning is mostly concerned with prediction. However, in computational historical linguistics another kind of statistical modelling is also important, namely explanatory modelling. Explanatory modelling tries to causally explain a data-generating process with the help of a theoretical framework [...].

The authors stress the distinction between **predictive modeling** on the one hand, and **explanatory modeling** on the other hand. Holding out test data for evaluating a model that has been fitted against a different set of test data is a well-established litmus test for the quality of predictive models but, as the authors argue, of lesser relevance for explanatory models. They furthermore point out (p. 235):

> Phylogenetic inference is an exemplary case of explanatory modelling: Based on a theoretical framework of cognate evolution we try to infer a language family tree which explains our data as well as possible. However, we are not primarily concerned with predicting new data points, e. g. the cognate classes present in languages that will evolve in the future. Separation of training and test data is in general not applicable to explanatory models [...].

In their insistence on keeping explanatory and predictive modeling apart, the authors cite Shmueli (2010) as an important source, where a forceful case for the importance of this distinction is made. According to Shmueli, explanatory modeling is "the use of statistical models for testing causal explanations" (p. 290) and predictive modeling "the process of applying a statistical model or data mining algorithm to data for the purpose of predicting new or future observations" (p. 291). The prototypical scenario for explanatory modeling would be one where the data-generating process is well-understood and captured by a generally accepted statistical model. An example might be estimating the chemical composition of a star from the spectral analysis of the electromagnetic waves it emits. The task of predicting the outcome of an election would be an extreme case of purely predictive modeling, where any type of data from macro-economic performance measures to psychological questionnaires might be useful, regardless whether the causal connections between independent and dependent variables are understood or not. Here, testing the model with data from past elections is indispensable to validate it.

Shmueli (2010) also stresses, however, that "it is important to establish prediction as a necessary scientific endeavor beyond utility, for the purpose of developing and testing theories" (p. 292). Furthermore, "the omission of predictive modeling for theory development results not only in academic work becoming irrelevant to practice, but also in creating a barrier to achieving significant scientific progress [...]. In areas such as bioinformatics, where there is little theory

and an abundance of data, predictive models are pivotal in generating avenues for causal theory." (p. 305)

In my view, prediction is important for explanatory modeling especially when there is no well-supported theory of the causal structure of the data-generating process. Testing the predictive performance of a model sheds light on the issue whether or not the model captures the relevant aspects of the data generating process. Prediction does not suffice to confirm a model, but it can falsify it. In a domain like CHL where (a) the causal factors involved are only partially understood and (b) only part of the established insights are captured by existing statistical models, prediction, I'd like to argue, is an important sanity check for our models, even though the ultimate goal is explanation.

There are several methods for predictive checks besides the partitioning of the available data into a training set and a test set. So the insistence on this particular method in the target article was arguably too strict. The best-known example for estimating out-of-sample performance of a fitted model based on training data alone is the **Akaike Information Criterion** (AIC; Akaike 1974). In the context of Bayesian inference, the **Widely Applicable Information Criterion** (WAIC; Watanabe 2010) and **Pareto-smoothed leave-one-out cross-validation** (Vehtari et al. 2017) are frequently being used for that purpose. However, these methods are not directly applicable to phylogenetic inference since the individual data points are not independent. In the target article, I performed an indirect predictive test of a phylogenetic model by using it as a component of a subsequent step – proto-form reconstruction – which is amenable to predictive testing. It is possible, however, to perform cross-validation of phylogenetic inference on the basis of cognate-class data directly. For instance, one could replace a randomly chosen subset of the cognate classifications by *undefined*, infer a posterior distribution of phylogenies, and predict the values of the missing cells. To my knowledge, none of the existing software packages for phylogenetic inference affords this kind of predictive check, so this would be a valuable programming project.

Another useful predictive check of model performance is (a) to use a fitted model to simulate many artificial data sets and (b) to test whether the observed data fall into the distribution of possible data as characterized by the simulations. If the observed data are unlike the simulated ones, the model is faulty. In the context of Bayesian statistics, this approach goes by the name of **Posterior Predictive Sampling** (PPS). This method is not only useful to asses the general suitability of a model, but one can also utilize it to probe which aspects of a model are reliable and which are not. For instance, Hammarström et al. point out (p. 242) that the model for phylogenetic inference I used in the pilot study does not capture the fact that there is typically one word per meaning in each language. While this constraint is enforced as part of the data-collection process–the method of

collecting Swadesh lists imposes this to a certain degree, and I picked out one word per meaning when converting the raw word lists to character matrices–there is nothing in the probabilistic model which would exclude a smaller or larger number of cognate classes per meaning.

To test this expectation statistically, I (a) computed the proportion of 1-cells among the non-missing cells in the cognate-class character vector for each language in the data used in the pilot study, and (b) determined the standard deviation of this measure across the languages in the sample. Let us call this quantity s_{emp}. Without missing data, s_{emp} should be 0, since each language would contain one 1-entry per concept. Due to a varying amount of missing entries in the ASJP (Wichmann et al. 2016) data, $s_{\text{emp}} \approx 0.002$.

In the next step I picked 1,000 samples from the posterior distribution of phylogenies and model parameters obtained via Bayesian phylogenetic inference and used them to simulated 1,000 character matrices of the same size as the training data. For each of these 1,000 simulated data sets, I computed s. This yielded a sample with mean $\approx$ 0.014 and 95% highest posterior density interval [0.009, 0.019]. The empirically observed value clearly falls outside the distribution of possible values predicted by the model being used for phylogenetic inference (see Figure 1). This result emphasizes Hammarström et al.'s point that the pattern *one word per concept* that holds for the training data is not predicted to hold by the statistical model used. This is a direct consequence of the fact that cognate-class characters are assumed to be stochastically independent by the model, while in reality they are not.

As Hammarström et al. correctly point out, this deficiency of the model can be sidestepped by assuming only one character per concept, with n cognate classes as possible values. As remarked in their comment, this approach comes with its own problems pertaining to model fitting. Another option worth exploring is to model discrete characters like cognate classes via latent continuous variables that determine the probability of observing a given trait value (cf. Felsenstein 2005).

This is not the place to explore any of these issues in detail. Suffice it to say that we are still quite far from a full understanding of the causal factors involved in the data generating process in phylogenetic inference (and other aspects of CHL), and even further from adequately capturing these factors in an explanatory probabilistic model. As long as this is the case, predictive checks are an important tool to test in what respects our preliminary models perform well, and in what respects they do not.

Let me conclude this section with addressing some of the other issues brought up in Hammarström et al.'s comment.

In the target article, I also proposed the methodological principle "**Only raw data as input**" (p. 156). The commentators remarked (p. 237) that this "may not

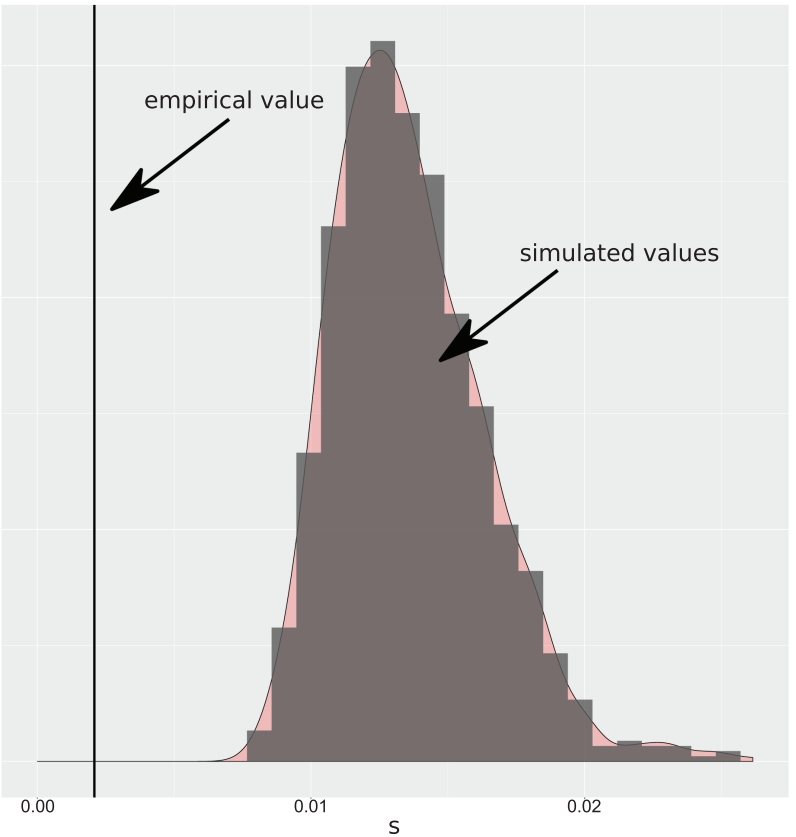

Figure 1: Posterior predictive check: empirical value and predicted distribution of *s*.

be the optimal name for the desired somewhat more extensive principle". Upon reflection, I agree with this assessment. The important point here is that the process of data collection and preparation should be as close as possible to the ideals of transparency and replicability. Expert judgments such as manual cognate classification are not *per se* unsuitable for model training purposes, provided that standards like consulting multiple annotators and documenting inter-annotator agreement are observed.

I also agree with the commentators that the task of "Demonstration of genetic relationship" (p. 158 of the target article/p. 238 of the comment) is not really amenable to the kind of statistical and computational modeling I envision. By definition, a group of languages is demonstrably genetically related if it is possible to trace a sufficiently substantial part of their core vocabularies and

grammatical properties to a common proto-language, from which the observed languages inherit them via an uninterrupted chain for vertical transmission (thus excluding transmission via language contact). At present time, such a demonstration is only possible via traditional, non-computational methods.

Further, the commentators remark (p. 241) with regard to the evaluation of the automatic cognate clustering step that "one would have expected accuracy numbers on cognate detection using the method actually presented for the data on Romance (or at least a subfamily of similar depth)." The ASJP data used for the pilot study are not annotated with cognate judgments. As a proxy, I used a mapping between the word lists from ASJP and from IELex (Dunn 2012) that was produced within the Evolaemp ERC project at Tübingen University. (This mapping is included in the GitHub repository for inspection.) The intersection of the data used in the pilot study with IELex contains 420 entries, i. e. roughly one quarter of the 2,058 ASJP entries used in the pilot. For this subset, the automatic cognate detection achieved a B-cubed precision of 0.99 and a recall of 0.62, which amounts to an F-score of 0.76. These numbers indicate that the automatic clustering method produces barely any false positives but a fair amount of false negatives.

The pilot study used a novel type of characters for phylogenetic inference dubbed "soundclass-concept characters", which are discussed by Hammarström et al. as well. In this connection I refer the interested reader to Jäger (2018), where I motivate and discuss this issue in greater depth.

Regarding the final step of proto-form reconstruction, the commentators find it "opaque [...] why the progressive alignment is not used directly to search for a minimally different proto-form" (p. 243). Searching directly for a proto-form minimally different from Latin would defy the general approach of unsupervised learning, paired with testing on unseen data. The method used in the pilot study identifies the likeliest proto-sound in each column of the progressive alignment instead.

3 Reply to List

Mattis List's contribution to this volume nicely illuminates the value of Open Science practice. While he tried to replicate my evaluation of the proto-form reconstructions of the pilot study, using the code that I made available via GitHub, he spotted a bug in my Python code. It had a minor effect on the quantitative results, while the qualitative results were not affected. I repeat the relevant sentences from Subsection 3.8 of the target article (p. 175), with corrections in bold: *The extant Romance doculects have an average score of* ***0.627****. The most*

conservative doculect, Sardinian, has a score of ***0.489****, and the least conservative,* ***French****,* ***0.703****.*

The corrected version of Figure 6 of the target article (p. 177) is given here as Figure 2.

However, List's main point in his comment is to propose a novel method for evaluating proto-form reconstructions. Rather than comparing sound-by-sound whether the reconstruction correctly identifies the phonetic value of the gold standard, List's method focuses on to what degree the reconstruction captures the phonemic contrasts within the gold standard, regardless of the specific phonetic values of the individual phonemes.

This proposal is potentially highly valuable for CHL, beyond the envisaged use as an evaluation for reconstructions. It can equally be applied to comparing documented languages, and it provides a measure of similarity. Unlike similarity measures commonly in use, List's measure will treat two languages as identical if they only differ by a regular (unconditioned) sound change. It is to be expected that it will therefore discount differences between dialects of the same languages while amplifying the distance between different languages. Also, it is applicable

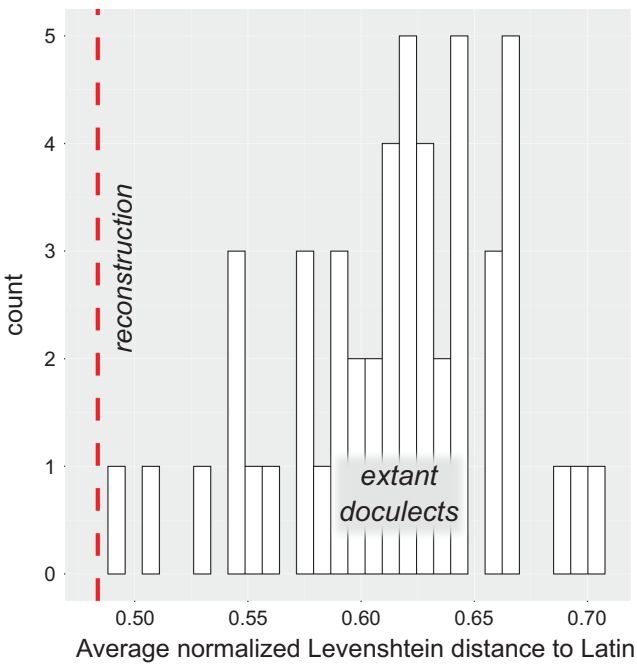

Figure 2: Average normalized Levenshtein distance to Latin words: reconstruction (dashed line) and extant Romance doculects (white bars) [correction].

to orthographic descriptions with different alphabets or sound-letter correspondence conventions, provided the orthographic conventions are phonetic.

To flesh out this idea in a fully algorithmic way, some further decisions have to be made. The method relies on a pairwise string alignment method. List writes: "This step can be carried out automatically, but it may be useful to manually adjust it, specifically in those cases where we expect a high degree of substantial differences" (p. 254). Rather than taking recourse to manual corrections, a fully automated version should arguably use an alignment method that is already sensitive to language-specific correspondence patterns, such as List's own LexStat method (List 2014) or Dellert's (2019) *Information-Weighted Sequence Alignment*. A more ambitious approach would be to search for the alignment that maximizes overall similarity. It remains to be seen whether this is possible in an efficient way.

Another issue to be addressed is how multiple synonymous entries for the same concept are to be treated. Here some method of assessing the probability of cognacy might be required to select for each concept the pair of words that are most likely to participate in regular sound correspondences.

4 Conclusion

One of the central points I wanted to make in the target article was that **evaluating** computational and statistical models in historical linguists should be a central concern in our research practice. Both comments reinforce that point. Also, both comments emphasize that evaluation should be performed from the perspective of our understanding of the causal fabric of language change, e. g. with regard to the organization of the lexicon or to the importance of regular sound changes. I fully concur with this, and I'd like to thank again the commentators for their inspiring contributions.

Funding: 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).

References

Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19. 716–723.

Dellert, J. 2019. *Information-theoretic causal inference of lexical flow*. Berlin: Language Science Press.

Dunn, M. 2012. Indo-European lexical cognacy database (IELex). http://ielex.mpi.nl/.
Felsenstein, J. 2005. Using the quantitative genetic threshold model for inferences between and within species. *Philosophical Transactions of the Royal Society B: Biological Sciences* 360(1459). 1427–1434.
Jäger, G. 2018. Global-scale phylogenetic linguistic inference from lexical resources. *Scientific Reports* 5, 2018. https://www.nature.com/articles/sdata2018189.
List, J.-M. 2014. *Sequence comparison in historical linguistics*. Düsseldorf: Düsseldorf University Press.
Shmueli, G. 2010. To explain or to predict? *Statistical Science* 25(3). 289–310.
Vehtari, A., A. Gelman, and J. Gabry. 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and computing* 27(5). 1413–1432.
Watanabe, S. 2010. Asymptotic equivalence of Bayes cross validation and Widely Applicable Information Criterion in singular learning theory. *Journal of Machine Learning Research* 11. 3571–3594.
Wichmann, S., E. W. Holman, and C. H. Brown. 2016. The ASJP database (version 17). http://asjp.clld.org/.

Article note: The code used when conducting the evaluation of the cognate clustering from the pilot study, the posterior predictive check (Section 2) and the corrected evaluation (Section 3) are available for download and inspection from https://github.com/gerhardJaeger/protoRomance/reply.