Harald Hammarström*, Philipp Rönchen, Erik Elgh and Tilo Wiklund

On computational historical linguistics in the 21st century

https://doi.org/10.1515/tl-2019-0015

1 Introduction

We welcome Gerhard Jäger's framing of Computational Historical Linguistics: its history and background, its goals and ambitions as well as the concrete implementation by Jäger himself. As Jäger explains (pp. 151–153), the comparative method can be broken down into seven steps and there have been attempts to formalise/automatise (some of) the steps since the 1950s. However, Jäger contrasts the work in the 1960–2000s on various steps as "mostly constituting isolated efforts" and, in contrast, characterises the biologically inspired work of the 2000s as a "major impetus". It is difficult to find the motivation for this division as the latter group, like the former, also concern themselves with only a subpart of the comparative method and, furthermore, rely fundamentally on subjective cognate judgments done by humans (as also acknowledged by Jäger later, pp. 156–157).

2 The goals of a program for computational historical linguistics

Jäger (p. 156) sets up four key principles for a computational historical linguistics: replicability, rigorous evaluation, separation of training and test data, and only

***Corresponding author: Harald Hammarström,** Department of Linguistics and Philology, Uppsala University, Box 635, 751 26 Uppsala, Sweden, E-mail: harald.hammarstrom@lingfil.uu.se
Philipp Rönchen, Department of Linguistics and Philology, Uppsala University, Box 635, 751 26 Uppsala, Sweden, E-mail: philipp.ronchen@lingfil.uu.se
Erik Elgh, Department of Linguistics and Philology, Uppsala University, Box 635, 751 26 Uppsala, Sweden, E-mail: erik.elgh@gmail.com
Tilo Wiklund, Department of Mathematics, Uppsala University, Uppsala, Sweden; Department of Linguistics and Philology, Uppsala University, Box 635, 751 26 Uppsala, Sweden, E-mail: tilo.wiklund@math.uu.se

raw data as input. Jäger states that these principles are "following the standards in statistical NLP" (p. 156). We strongly support Jäger's idea that computational historical linguistics should be governed by strict methodological principles but disagree on what the most important principles are and in which contexts they are applicable.

Replicability. This is a general scientific principle not restricted to computational historical linguistics.

Separation of training and test data. This is indeed a very important principle in NLP. However, we would like to argue that it has only limited applicability outside of the field of Machine Learning. Separation of training and test data is based on the idea that predictions from a data set about which we have information should generalise to new data sets (see Figure 1a for an illustration). Therefore, a method is not evaluated on the whole original data set but only on a

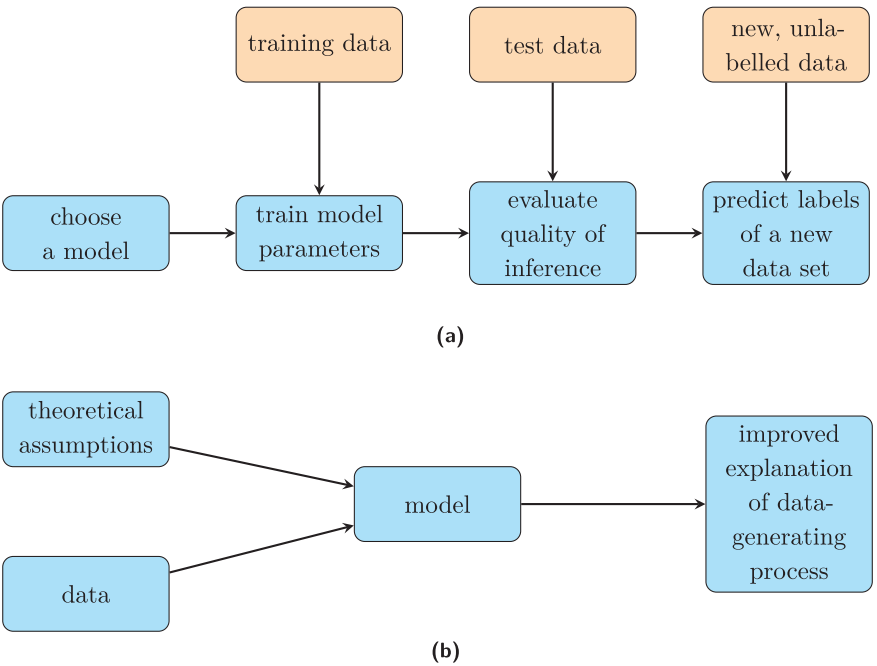

Figure 1: While machine learning is mainly concerned with prediction, in computational historical linguistics we are often interested in improving our explanation of some process of language change. (a) Schematic description of the typical workflow in prediction in Machine Learning, where data is separated into training and test data. (b) Schematic description of the typical workflow in explanatory modelling.

reserved part of it which is not used for training and which therefore works as a proxy for a "new unseen" data set. Thereby "overfitting", i.e. the problem that a method might explain existing data well but fail to make valid predictions about new data data points is minimised.

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 (see Figure 1b). Explanatory modelling is sometimes just called inference, but the term inference can also include prediction. Here, we use the distinction between explanatory and predictive modelling as it is made by Shmueli (2010).

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 (see Shmueli 2010: 296–297 for this point as well as some examples in which some partitioning of the data set can still be useful for explanatory models).

As an example, imagine that we have cognate data for one hundred languages of a given family, and that we split our dataset into two sets (Dataset A and Dataset B) that contain the data from 50 of these languages each. Now imagine that we construct a language tree for Dataset A using phylogenetic methods. Then there is no straight-forward way to use Dataset B to test whether the tree for Dataset A is right or wrong – a language tree does not make any predictions for languages which are not contained in the tree.

In his case study, Jäger himself is using Machine Learning algorithms (Sections 3.2 and 3.6), but also performing statistical hypothesis testing (Section 3.1) and phylogenetic inference (Sections 3.4 and 3.5). The latter two are concerned with explanatory modelling, and here training and test data are not useful concepts.

Rigorous evaluation. Jäger is of course right that models should be rigorously evaluated. However, "model evaluation" can mean several quite different things. For predictive modelling, the only quality of the model one is interested in is its predictive power, and the quality of prediction can often efficiently be approximated by a separation of training and test data. In explanatory modelling, a good model should do two different things: It should reflect our theoretical

assumptions about the process that we study, and it should explain the data as well as possible (Shmueli 2010). It happens that explanatory models also have *some* predictive component. In this case the model should also be able to predict some qualities of new data points well, but only as a supplement to the other two goals.

Let us again take the phylogenetic construction of a language family tree as an example: A good phylogenetic model should reflect our theoretical assumptions about language change well, i.e. that languages exchange a certain fraction of the vocabulary over time and that it is unlikely that the same word will have appeared twice in two different parts of the family tree (without contact events), and it should explain the cognate data that we have for known languages of the family.

Usually, there is a trade-off between these two aims of an explanatory model: By adjusting our theoretical assumptions, we can fit the model better to our data. The choice of model parameters can be viewed as instance of "adjusting our theoretical assumptions based on the data". However, if we change our theoretical assumptions too much to fit the data, we will be left with mere *descriptions* of the data set, i.e. we lose our ability to obtain novel insights about the process that we study. This phenomenon is related to the kind of overfitting that occurs in predictive modelling (which, in predictive modelling, can often be avoided by separation of training and test data).

In the following, we give some methodological principles which we think can help finding a model (explanatory or partly predictive) that fits the data well but avoids overfitting:

- *Make theoretical assumptions as transparent as possible*. An explanatory model only makes inferences based on a theoretical framework and to determine the applicability of the model depends of course on that the framework is known. This point is of additional importance in computer-based modelling: When software is developed for the purpose of modelling, certain assumptions need to be made to make computations feasible. However, quite often these assumptions are very opaque and can even go *against* the theoretical properties one assumes for the process (see Symons and Alvarado 2016: 6–8 and Winsberg 2010: 120f.). A good example in phylogenetic inference of an assumption which is made to make computations feasible but which goes against theory is the coding of cognate classes as binary characters, see Section 3.4.2.
- *As far as possible, modelling assumptions should be tested for internal consistency and consistency with the data*. This can be done by separating training and test data (for predictive model components), by measuring model fit or by testing single model assumptions with statistical tests. However, while such techniques can falsify a given model, they can never "prove" its

appropriateness. This always depends on a theoretical framework which is falsifiable.

- *Beware of hidden overfitting*. It is very easy to adjust one's model assumptions without really being aware of that, something we will call "hidden overfitting". Let us give one example of how hidden overfitting is possible, again in relation to phylogenetic inference: Imagine that we construct a language tree of Indo-European, by phylogenetic methods. We infer a recent clade containing both English and French, which is not compatible with the historical record. Imagine that we change the parameters of our model until English and French are sorted in different clades, and then publish the results (but without mentioning that our original model gave a different tree topology). Then we have engaged in a case of hidden overfitting. While it is possible that the model with the adjusted parameters *is* a better model of Indo-European, we do not know that: There are many models which can fit data well but which are theoretically wrong, and we did not change our parameters (that is: our model assumptions) in an internally consistent and transparent way. Our new model has therefore reduced explanatory power.
- *Simple models are better than more complex ones*. Hidden overfitting (or in fact overfitting of any kind) is more likely to happen if our models contain a lot of parameters that we can adjust based on the data. Therefore, if a simple and a more complex model explain a process equally well, the simple model should always be preferred. This is a general scientific principle known as Ockham's razor, but we think it is useful to reconsider the principle in relation to the danger of hidden overfitting.

Only raw data as input. We agree with a principle that strives only to make objective and theoretically arguable transformations of original observations. A case in point are, of course, human cognate judgments known to be subjective to varying degrees (pp. 156–157). But all is not won by using "only raw data" if the subjective human transformation (cognate judgment) is replaced by an objective computational method which is questionable on the same grounds as the original transformation which it sought to replace. For example, cognate judgment algorithms are typically trained on subjective human cognate judgments and/or informally engineered towards them (p. 165). Hence, "only raw data as input" may not be the optimal name for the desired somewhat more extensive principle.

3 A case study: reconstructing proto-Romance

We applaud Jäger's expositionary case study to reconstruct proto-Romance from modern Romance language data. It is a near end-to-end system, step-by-step

described, applied to data in the open database ASJP, https://asjp.clld.org, and as such a star in replicability and a clear starting-point for discussion.

3.1 Demonstration of genetic relationship

Jäger (Section 3.1) gives a statistical argument for the Romance lects being genealogically related. As evidence he gives the fact that all pairs of Romance lects are too similar, based on a string similarity measure he calls PMI-distance, compared to pairs from Papunesia deemed unrelated in Glottolog. While the similarity is clear from simple visualisation it is welcomed that Jäger attempts to formalise this. The procedure he proposes is meant to test the hypothesis that a collection of lects $L_1, \ldots, L_n$ are unrelated. From a somewhat abstract point of view the method proceeds as follows:

1. Produce an estimate $\widehat{F}$ of the cumulative distribution function of the distribution of the PMI-distances between pairs of unrelated lects, based on a set of lects known to be unrelated.
2. Define a test procedure for the hypothesis that two lects L and L' are non-related by treating $\widehat{F}(d(L, L'))$ as a p-value, where $d(L, L')$ is the PMI-distance between L and L'.
3. Evaluate the significance level and power of this test on a separate, independent set of known related and unrelated lects.
4. Compute $\widehat{F}(d(L_1, L_2)), \widehat{F}(d(L_1, L_3)), \ldots, \widehat{F}(d(L_{n-1}, L_n))$ where $d(L_1, L_2)$, $d(L_1, L_3), \ldots, d(L_{n-1}, L_n)$ is the collection of all PMI-distances between $L_1, \ldots, L_n$. Treat these values as p-values of separate tests and apply the Holm–Bonferroni method to compensate for multiple tests.
5. If the Holm–Bonferroni method allows us to reject all hypotheses reject the hypothesis that the lects contain some pair of non-related ones.

In the first two steps, Jäger uses a collection of Papunesian lects. For the estimator, he takes (the integral of) a logspline density estimate based on the collection of all PMI-distances between these lects, *as if all these distances were independent*.

In Step 3, Jäger applies the test at a (nominal) significance level of $\alpha = 0.0001$ to the set of all available pairs of African lects. However, looking at the resulting contingency table (p. 160), of 532 + 1254726 actually unrelated African pairs, 532 of them are predicted to be related – a higher ratio (≈ 0.00042) than the nominal significance level.

Applying Steps 4 and 5, he claims to be able to reject the hypothesis of any Romance lects being unrelated with a high significance.

A heuristic justification for the procedure is that the estimate $\widehat{F}$ should be close to the actual distribution function of the PMI-distances of unrelated pairs of lects, and that if it were *exactly equal*, then the test in Step 2 would be exact against the hypothesis that $d(L, L')$ is from this actual distribution. While conceptually appealing, this method, as implemented, has some serious flaws.

The p-values will be at best approximate, since the estimator $\widehat{F}$ is not the true distribution function. That being said, if the empirical distribution function were used for the estimator $\widehat{F}$ the test *would be equivalent to a standard permutation test* if the observation being tested is included in the sample or a correction term $1/n$, for n the sample size, is added. Performing Jäger's procedure with this change (and leaving everything else unchanged), we obtained larger but comparable p-values. *If they had been based on an independent sample of distances* our nominal significance level would have been correct.

However, the pairs of similarities are not independent, which must be taken into account when using them in statistical tests. In essence, the similarity of A–B and B–C put bounds on the similarity A–C. For n input lects, there are $n(n-1)/2$ pairs, of which at most $n/2$ – a much smaller number – can be truly independent. Using a Holm–Bonferroni correction it would not be possible to obtain a rejection based on such an independent sample, since each individual p-value would be at least 2/2053 (where 2054 is the number of Papunesian lects in the data).

An issue is that the similarity distribution between unrelated lects depends on areal effects. Comparing the PMI distances between unrelated lects in different world-regions we find that the distributions differ (see Figure 2). The differences are especially strong in the left tails of the distributions, which are of importance for the statistical test outlined above. Observe, for example, that around 20% of *non-related* European lect pairs would be incorrectly classified as related

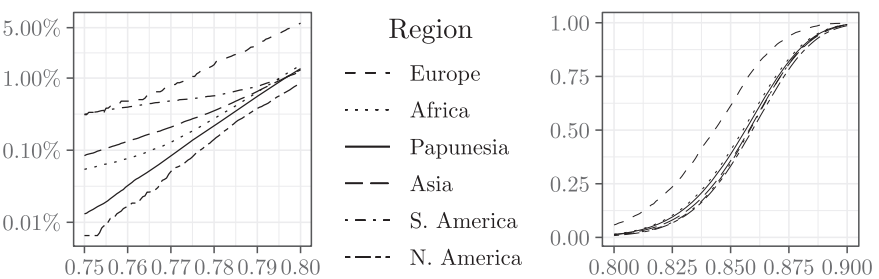

Figure 2: On the right, estimated cumulative distribution functions of PMI distances between *unrelated* lects in different world regions, estimated using non-independent samples. On the left, the left tails of the distributions.

at a nominal significance level of 5%, if any of the other regions were used as reference (presumably due to language contact). This issue could be alleviated by comparing the PMI distances of European lects (like the Romance lects) to a similarity distribution between lects in an area where strong areal effects are known to exist.

In general, it is not obvious that the hypothesis being tested (that no distances within the proposed family are "large") is actually one of interest. For example, the overall hypothesis would be false for a family containing two very distant relatives, so far away from each other as to have no discernible cognates, while being linked by a long chain of closely related languages. Even correcting for the technical issues outlined before, such a family would still be rejected as long as the test relies on rejecting non-relatedness for *all* pairs of lects in the family.

In this sense, being a language family is a strictly weaker condition than all pairs of languages in the family "being close", it suffices that an appropriate subset of all the distances be "small". On a positive note, therefore, one should be able to find stronger tests for the hypothesis that the languages form a family.

As a final note, in utilising Glottolog, Jäger violates his "only raw data as input"-desideratum (p. 156). The procedure is ultimately circular since the Glottolog classification is largely based on basic vocabulary comparison. The reliance on Glottolog could have been avoided by seeking the similarity distribution for unrelated lects elsewhere, for example by comparing words of different meaning, as pioneered in linguistics by Oswalt (1970) (and espoused by Jäger 2013: 248–257) or by careful approximation of symbol frequencies and phonotactics (see e.g. Ratcliffe 2012 and earlier papers cited therein).

3.2 Pairwise string comparison

Jäger (Section 3.2) presents a versatile string comparison scheme. At the outset, only symbols that are the same count as similar, but in subsequent steps, non-identical symbols that are aligned around sufficiently many similar symbols count as successively more similar. Clouding the elegance of this procedure is the use of thresholds θ_{PMI} and parameters d and e, regarding which the reader is referred to Jäger (2013), where these depend on another threshold θ_{ERC}. In Jäger (2013: 268–270) these are set with some amount of arbitrariness (θ_{ERC}) or with a limited search (θ_{PMI}, d, e).[1] There is a neglected theoretically more elegant

1 The search limitation may be important. It seems that if there is one pair of related languages with one very long identical cognate one can set θ_{PMI} very high and get nonsensical delete and substitution costs.

solution to the problem addressed in Ellison (2007) but it is formulated specifically for the case of two input languages. While the thresholds used by Jäger have good practical value, a general theoretical solution to the problem would be desirable and within reach.

3.3 Cognate clustering

Jäger (Section 3.3) distills the comparatively rich body of work on cognate clustering into a graph-based clustering algorithm whose edges are pairwise string similarities. There are one or two mild unexplained choices in the exclusion of edges with a probability value of less than half versus the inclusion of those with more than a half, and the choice of the Label Propagation algorithm.

The section on cognate clustering is concluded with the claim that "Based on evaluations against manually assembled cognacy judgments for different but similar data (Jäger and Sofroniev, 2016; Jäger et al., 2017), we can expect an average *F*-score of 60–80% for automatic cognate detection". This claim is somewhat misleading in this context because the numbers refer to cognate clustering using a supervised approach, different from the unsupervised approach used in the paper. Also the scores found in the source papers has a much wider range than 60–80%. In fact, cognate detection from scratch is highly dependent on the depth of cognates in a given dataset (cf. List et al. 2017) – finding deep cognates accurately is much more difficult than finding shallow ones – so it is not that one detection method has a more-or-less constant accuracy for all datasets. Here, in the position paper by Jäger, 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).

3.4 Phylogenetic inference

3.4.1 General remarks

We thank Jäger for giving a coherent description of the complex terminology and the diversity of phylogenetic methods that have been used in linguistics. One point that we would like to add is that while phylogenetic methods have been used in linguistics for almost twenty years, their overall robustness (for linguistic data) has not yet been coherently proven. Rama (2018) shows that commonly used Bayesian phylogenetic methods are very sensitive to the tree priors used and the number of poorly attested languages that are included in the analysis. So while we agree that further exploration of phylogenetic methods is warranted,

better assessments of the accuracy and robustness of these methods are urgently needed.

3.4.2 Application to the case study

Jäger sets up a biologically inspired (Section 3.4) phylogenetic inference model fitting cognate gain and loss through time in the branches of a tree (now not uncommon in linguistics, see e.g. Dunn 2014).

Jäger adopts the typical binarising character coding scheme, whereby each cognate class for each meaning constitutes a character and its presence/absence is indicated for a given language (Nicholls and Gray 2008: 553–554). In other words, if a meaning has *n* cognate classes present in the input dataset, *n* binary characters will be associated with this meaning. This approach has the advantage of being relatively easy to implement. However, the meanings with many cognate classes, i.e. the unstable meanings, will occupy a proportionately larger chunk of the probability mass which is quite possibly not the desired weighting scheme. Furthermore, it allows for ancestral languages in the inference process to have any number between 0 and the (potentially large) number of cognate classes *n* present for a given meaning. While this is logically possible, languages typically have only one or a few words per meaning in basic vocabulary datasets (e.g. Gudschinsky 1956: 179, Nicholls and Gray 2008: 564–565, Chang et al. 2015: 215 – the ASJP Romance dataset used by Jäger has 1.107 words per meaning, near identical to the 1.105 of the ASJP dataset). This means that there is an incongruence between the model and the data that is fed into it, though little is known about the practical implications (cf. Nicholls and Gray 2008: 564–565).

Another strategy for character coding corresponds more closely (in fact, exactly) to the one-word-per-meaning behaviour of typical input data: if all cognate class labels representing the same concept are instead coded as states of one single character (as done, e.g. in part (i) of the analysis in Kitchen et al. 2009: 2705, S2, S4). This strategy comes with its own set of problems – not least the fact that it does not allow *any* multiple simultaneous words per meaning, which languages do, albeit only a little – and the practical and theoretical difficulties associated with estimating individual within-character state transitions and potential ancestral states not attested in the input data.

Clearly, important questions surrounding cognate character representation need to be elucidated.

A novel component, not used by other linguists, is the inclusion of soundclass-concept characters by Jäger. The soundclass-concept characters are interesting in that they, as Jäger mentions, “also capture sound shifts” (p. 169).

This approach is somewhat imperfect in that it does not only capture sound shifts. It will also compare and possibly correlate sounds of non-cognate words. As such, it introduces noise that should cancel out in the grand scheme but may have negative local effects. The exact workings in the present case study, or in a 100- or even 200-wordlist remain opaque. The characters reflecting genuine sound shifts introduce a series of interesting weighting questions both within the soundclass characters and the soundclass characters vis-a-vis the lexical cognacy characters.

If, for a given language in the data matrix, five concepts are listed as containing /g/, and three concepts listed as containing /f/, and both sounds arose through a single regular sound change each, these two events will be weighted differently solely based on the random relative prevalence in the data matrix of words going through the sound change. This could be a desired weighting, if one views the word list as representative of the language as a whole. Then the more prevalent sounds resulting from regular sound change can be seen as providing a more secure attestation of the sound change, and thus validating its higher weighting.

Jäger's maximum clade credibility tree (p. 170) is based on an analysis using both soundclass characters and lexical cognacy characters. If each soundclass character is weighted equally to a lexical cognacy character, this gives more importance to the former (since there are more of them in number), which may or may not be the desired weighting scheme.

Our aim here is not to push any of these weighting schemes over any other, but to raise the issue arising (but left open) from Jäger's experiment

3.5 Ancestral state reconstruction, multiple sequence alignment and proto-form reconstruction

Given a posterior distribution of trees, Jäger (Section 3.5–3.7) projects proto-forms to the proto-node. First ancestral state reconstruction is used to get probabilities for cognate classes being present at proto-nodes. The ancestral state reconstruction methodology closely follows that of the phylogenetic inference and produces a most likely cognate class for each meaning class at each proto-node. The remaining problem is then to pinpoint an actual form at the root for each meaning given a set of n relevant modern forms. Jäger chooses a multiple sequence alignment approach maximising the sum of a function of the pairwise alignments. To find this maximum a heuristic algorithm, progressive alignment, is applied which takes the n strings and a tree on them as input. Since the tree, which is binary, is given as input, it is opaque to us why the progressive alignment is not used directly to search for a minimally different proto-form.

3.6 Result

The automatically reconstructed forms for proto-Romance are given by Jäger (pp. 175–178) for inspection, allowing an immediate intuitive appreciation of the results. The formal evaluation against attested Latin forms shows that the automatic reconstruction is closer to Latin than any modern language, and thereby proves that the procedure makes significant progress towards history. This is what Jäger aimed to demonstrate and successfully achieved.

4 Conclusion

The state-of-the-art as pronounced by Jäger reflects clear achievements on a number of subproblems of Computational Historical Linguistics. With Jäger's case study as a convenient departure point, we highlight important questions remaining to be worked out in proving relatedness, cognate detection, phylogenetic modelling and the dialectology between human, machine and data in the general workflow. Yet perhaps the most conspicuous difference between Computational Historical Linguistics as described by Jäger and the classical comparative method is the recognition of *sound changes* and their role in reconstruction and subgrouping. Sound changes are especially powerful for subgrouping as they are typically directional (see e.g. (Chang et al. 2001) for a reason why), i.e. given a correspondence between two modern sounds *X* and *Y*, regardless of the languages involved, it is much more likely that one of them, say *X*, turned to *Y* than vice versa. This is powerful for subgrouping because the language(s) with the *Y* reflex are then innovative, for which a common ancestor may be the preferred explanation. In the biologically inspired phylogenetic work, relying on cognate correspondences which are not similarly directional, this powerful criterion is not used (cf. Donohue et al. 2012). The sound class concepts introduced by Jäger go in the direction of sound changes, but do not actually make use of the inherent directionality of sound changes. The lack of a clear analogue in biology is obviously the reason, but why not make the best of both worlds?

Acknowledgment: This research was made possible thanks to the financial support of the CHRONOS – Chronology of Roots and Nodes of Family Trees funded by Stiftelsen Marcus och Amalia Wallenbergs Minnesfond 2007.0050 awarded to Gerd Carling (Lund University). We thank Joakim Nivre for useful input on predictive and explanatory models.

References

Chang, S., M. C. Plauché & J. J. Ohala. 2001. Markedness and consonant confusion asymmetries. In Hume, E. and Johnson, K. (eds.), *The role of speech perception in phonology*, 79–101. San Diego CA: Academic Press.

Chang, W., C. Cathcart, D. Hall & A. Garrett. 2015. Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language* 91(1). 194–244.

Donohue, M., T. Denham & S. Oppenheimer. 2012. New methodologies for historical linguistics? Calibrating a lexicon-based methodology for diffusion vs. subgrouping. *Diachronica* 29(4). 505–522, 538–546.

Dunn, M. 2014. Language phylogenies. In C. Bowern & B. Evans (eds.), *The Routledge Handbook of Historical Linguistics*, 190–211. New York: Routledge.

Ellison, T. M. 2007. Bayesian identification of cognates and correspondences. In *Proceedings of Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology*, SigMorPhon 2007, pages 15–22. ACL, Stroudsburg, PA, USA.

Gudschinsky, S. C. 1956. The abc's of lexicostatistics (glottochronology). *Word* 12(2). 175–210.

Jäger, G. 2013. Phylogenetic inference from word lists using weighted alignment with empirically determined weights. *Language Dynamics and Change* 3. 245–291.

Kitchen, A., C. Ehret, S. Assefa & C. J. Mulligan. 2009. Bayesian phylogenetic analysis of semitic languages identifies an early bronze age origin of semitic in the near east. *Proceedings of the Royal Society of London, Series B* 276. 2703–2710.

List, J.-M., S. J. Greenhill & R. D. Gray. 2017. The potential of automatic word comparison for historical linguistics. *PLOS ONE* 12(1). 1–18.

Nicholls, G. K. & R. D. Gray. 2008. Dated ancestral trees from binary trait data and their application to the diversification of languages. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 70(3). 545–566.

Oswalt, R. L. 1970. The detection of remote linguistic relationships. *Computer Studies in the Humanities and Verbal Behavior* 3. 117–129.

Rama, T. 2018. Three tree priors and five datasets: A study of Indo-European phylogenetics. *Language Dynamics and Change* 8. 182–218.

Ratcliffe, R. R. 2012. On calculating the reliability of the comparative method at long and medium distances: Afroasiatic comparative lexica as a test case. *Journal of Historical Linguistics* 2(2). 239–281.

Shmueli, G. 2010. To explain or to predict? *Statistical science* 25(3). 289–310.

Symons, J. & R. Alvarado. 2016. Can we trust big data? Applying philosophy of science to software. *Big Data & Society* 3(2). 1–17.

Winsberg, E. (2010). *Science in the age of computer simulation*. Chicago: University of Chicago Press.