Investigating the potential of ancestral state reconstruction algorithms in historical linguistics

Gerhard Jäger *(gerhard.jaeger@uni-tuebingen.de)* Johann-Mattis List *(mattis.list@lingpy.org)*

***Abstract*—Current efforts in computational historical linguistics are predominantly concerned with phylogenetic inference. However, methods for ancestral state reconstruction have been only sporadically applied. This is surprising since reconstruction is considered essential both in evolutionary biology and in classical historical linguistics. In contradistinction to phylogenetic algorithms, automatic reconstruction methods presuppose phylogenic information in order to explain what has evolved when and where.**

Here we report a pilot study on the potential of reconstruction algorithms in historical linguistics. Based on an explicit family tree, we apply different algorithms to wordlist data in order to infer how the words evolved along the phylogeny, and which words were used without change of meaning in the ancestral languages.

I. Introduction

Phylogenetic reconstruction plays leading role in quantitative approaches to historical linguistics, and many algorithms, workflows, and software packages for the reconstruction of phylogenetic trees and networks have been proposed in the last two decades. While tree- and network-building methods play a leading role in modern historical linguistic research and more and more scholars tend to use them, methods for ancestral state reconstruction (ASR) have been only sporadically tested and applied [1], [2]. This is surprising, firstly, since the application of ASR is quite common in the discipline of evolutionary biology which usually serves as a pool of inspiration for quantitative endeavours in historical linguistics, and secondly, since ASR plays a major role in traditional historical linguistics [3].

While tree-building methods seek to find branching diagrams which explain how a language family has evolved, ASR methods use the branching diagrams in order to explain what has evolved concretely (see the example in Fig. 1). In traditional historical linguistics, the search for the concrete is best reflected in linguistic reconstruction, i.e. the reconstruction of proto-forms of an unattested ancestral language, but also in semantic reconstruction, i.e. the attempt to find the original meaning of a given proto-form.

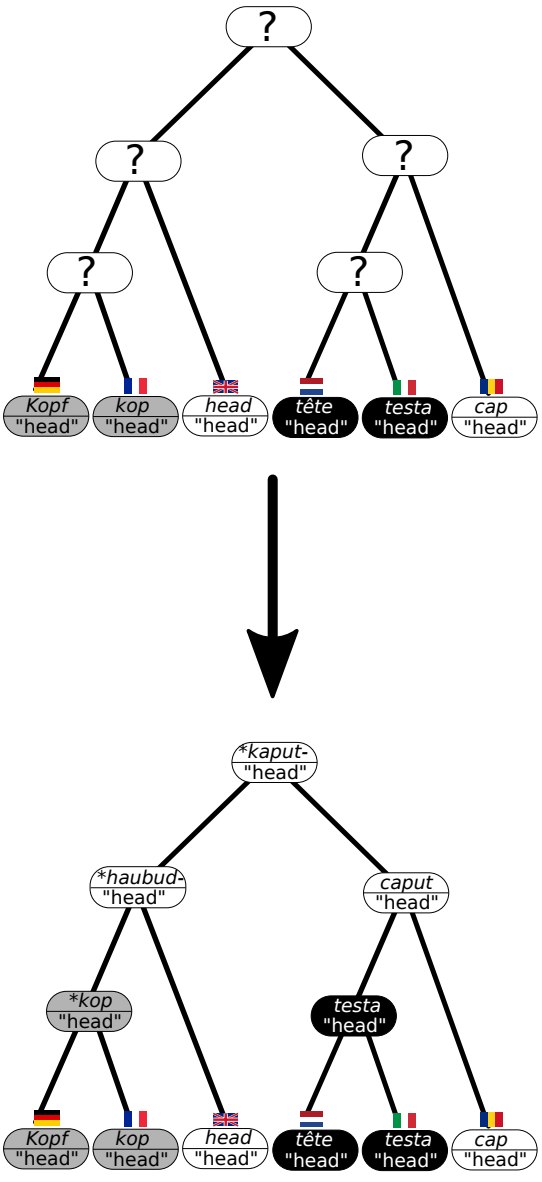

Fig. 1. Ancestral state reconstruction for words meaning ‘head’.

Here we report initial tests of the potential of ASR algorithms in historical linguistics. Based on an explicit model of external language change as represented by a family tree, we apply different ASR algorithms to wordlist data in order to infer how the words evolved along the tree, and which words were used without change of meaning in the ancestral languages.

II. Materials and Methods

In order to investigate the power of Parsimony methods, we used specific samples of two large lexicostatistical databases, the Indo-European Lexical Cognate Database ([4]; http://ielex.mpi.nl/), and the Austronesian Basic Vocabulary Database (ABVD [5]; http://language.psy.auckland.ac.nz/austronesian/). This data is structured in wordlist form, that is, for a given set of meanings (207 in IELex and 210 in ABVD), the translations into different languages are given, and annotated for cognacy. For the pilot study, we used all 153 doculects present in IELex and a sample of 100 doculects from ABVD. The data for both samples was divided into one training and one test set. For both samples, the proto-forms for the oldest proto-language in the sample (Proto-Indo-European and Proto-Austronesian) was available and used as a gold standard in our investigations.

ASR relies on a (rooted) phylogenetic tree. To obtain such trees, we performed Bayesian phylogenetic inference on the full (binarized) data from IELex and the data for all cognate classes for the 100-doculect sample from ABVD (using the software *Beast*; [6]; http://beast.bio.ed.ac.uk/beast) and obtained a summary tree (using *TreeAnnotator*; http://beast.bio.ed.ac.uk/treeannotator).[1] Additionally, we sampled 1,000 trees from the posterior distribution for both families.

Both the summary tree and the trees from the posterior sample are binary branching. We also considered multifurcating trees by collapsing all branches with a length below a certain threshold which was identified manually on the training sets.

ASR was performed (1) on the summary tree, (2) its multifurcating version, (3) on all trees from the posterior sample, and (4) on their multifurcating version. In (3) and (4), a cognate class was reconstructed for the proto-language iff it was reconstructed for at least 50% of all trees in the sample.

[1]We imposed the constraints that Anatolian branches off first for IELex and that Malayo-Polynesian forms a monophyletic clade for ABVD.

Furthermore we compared three different methods for ancestral state reconstruction: (a) Sankoff parsimony [7] with binary state characters, (b) Sankoff parsimony with multistate characters, and (c) weighted parsimony for binary state characters based on the minimal lateral networks (MLN) method [8].

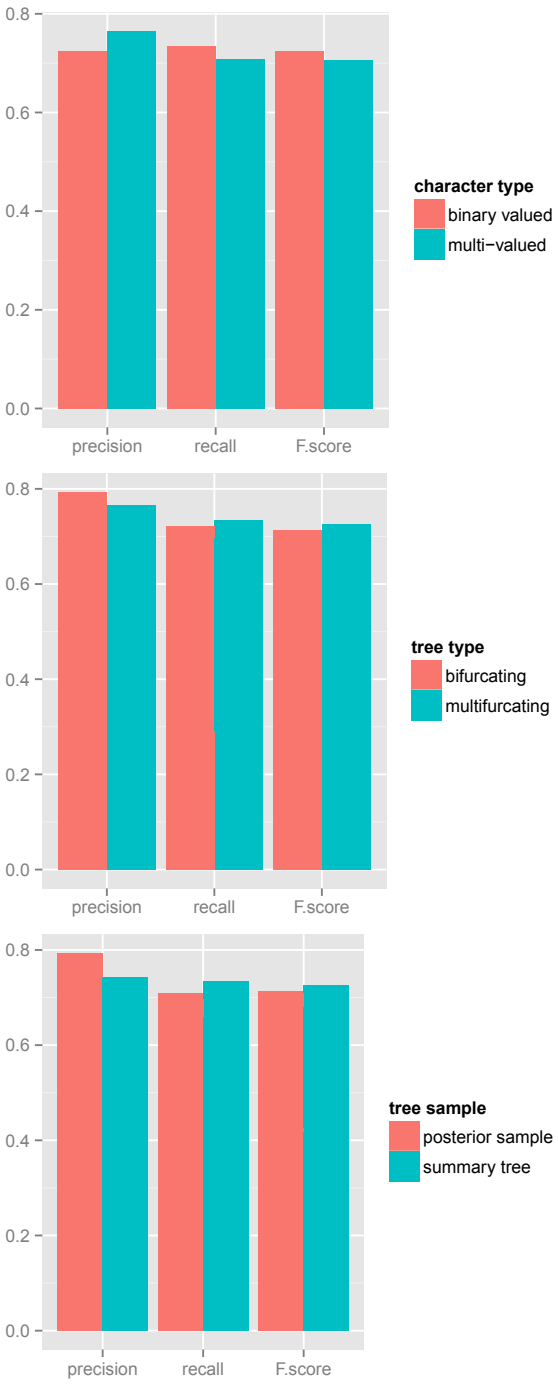

Fig. 2. Comparison of evaluations (continued in Fig. 3).

ABVD

algorithm	*characters*	*furcating*	*treeSample*	*precision*	*recall*	*F-score*
MLN	binary	multifurcating	summary tree	0.440	0.722	0.547
MLN	binary	multifurcating	posterior sample	0.757	0.354	0.483
MLN	binary	bifurcating	summary tree	0.481	0.405	0.440
Sankoff	multi	multifurcating	summary tree	0.305	0.709	0.426
Sankoff	binary	multifurcating	summary tree	0.341	0.557	0.423
Sankoff	multi	multifurcating	posterior sample	0.295	0.696	0.415
Sankoff	multi	bifurcating	summary tree	0.295	0.696	0.415
Sankoff	multi	bifurcating	posterior sample	0.279	0.671	0.394
MLN	binary	bifurcating	posterior sample	0.537	0.304	0.388
Sankoff	binary	multifurcating	posterior sample	0.205	0.570	0.301
Sankoff	binary	bifurcating	posterior sample	0.205	0.570	0.301
Sankoff	binary	bifurcating	summary tree	0.175	0.570	0.268

TABLE I
RESULTS, ORDERED BY DESCENDING F-SCORES, FOR ABVD.

IELex

algorithm	*characters*	*furcating*	*treeSample*	*precision*	*recall*	*F-score*
Sankoff	binary	multifurcating	summary tree	0.716	0.734	0.725
Sankoff	binary	bifurcating	posterior sample	0.718	0.709	0.713
Sankoff	binary	bifurcating	summary tree	0.704	0.722	0.713
Sankoff	binary	multifurcating	posterior sample	0.724	0.696	0.710
Sankoff	multi	multifurcating	posterior sample	0.765	0.658	0.707
MLN	binary	multifurcating	posterior sample	0.758	0.633	0.690
Sankoff	multi	bifurcating	posterior sample	0.746	0.633	0.685
Sankoff	multi	multifurcating	summary tree	0.735	0.633	0.680
Sankoff	multi	bifurcating	summary tree	0.721	0.620	0.667
MLN	binary	multifurcating	summary tree	0.584	0.658	0.619
MLN	binary	bifurcating	posterior sample	0.793	0.291	0.426
MLN	binary	bifurcating	summary tree	0.742	0.291	0.418

TABLE II
RESULTS, ORDERED BY DESCENDING F-SCORES, FOR IELEX.

III. RESULTS

All methods were applied to the two test sets, using the four samples of reference trees mentioned above. The list of cognate classes present in the proto-language according to expert assessments were used as gold-standard. It provides a binary classification of all cognate classes (1: present/ 0: not present in the proto-language). The performance of the methods compared was evaluated by calculating precision, recall and F-score. The full results are given in Table I (ABVD) and II (IELex). Figs. 2 and 3 give aggregated values for the various binary choices of algorithms, data and guide trees.[2]

Results are generally better for IELex than for ABVD. A possible explanation for this discrepancy might be that we used only a sample for 100 doculects for ABVD (out of 700 doculects in the database), while the full database was used for IELex.

As for the other binary choices, we generally observe a trade-off between precision and recall. Therefore it is not possible to single out an optimal ASR method on the basis of our results.

[2]As the MLN algorithm always uses binarized characters, its results were excluded for this choice.

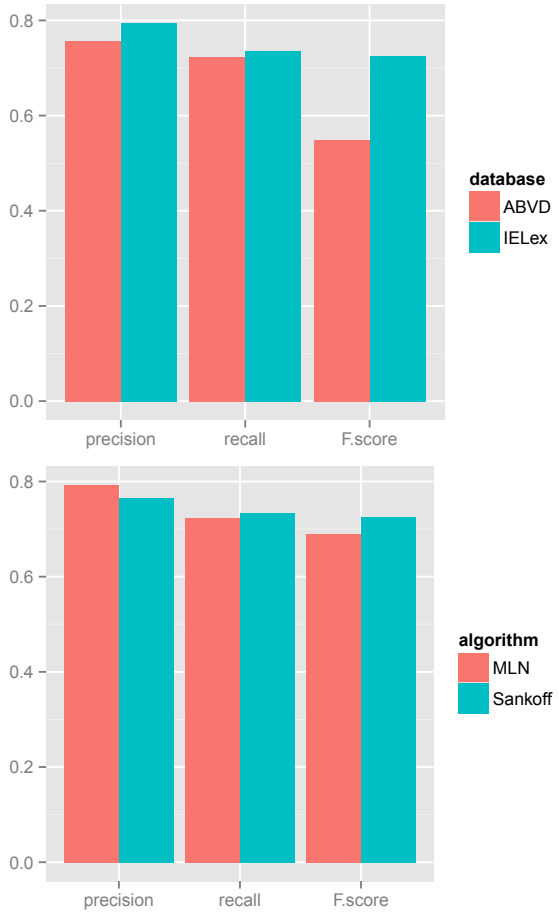

Fig. 3. Comparison of evaluations (continued from Fig. 2).

IV. Discussion

ASR for IELex and ABVD *per se* is of limited use as good reconstructions are available from traditional comparative research. Still, ASR is a first step towards more rewarding goals. Let us conclude with listing three of them. (1) A reconstruction of all changes in cognate classes allows to identify loci of homoplasy. There are only two plausible explanations for homoplasy of cognate classes: (a) parallel semantic change and (b) borrowing. A semi-automatic inspection of homoplasies is a promising route towards identifying pre-historic borrowings and thereby improving phylogenetic inference. (2) ASR affords to quantify differential rates of evolution for different Swadesh concepts. We expect this to be a principled way to assess the stability of concepts. (3) ASR is a precondition for automatically identifying sound change and reconstructing proto-forms.

Acknowledgment

This research was supported by the ERC Advanced Grant 324246 EVOLAEMP and the DFG-KFG 2237 *Words, Bones, Genes, Tools* (GJ) and the DFG research fellowship grant 261553824 *Vertical and lateral aspects of Chinese dialect history* (JML), which is gratefully acknowledged.

References

[1] A. Bouchard-Ct, D. Hall, T. L. Griffiths, and D. Klein, "Automated reconstruction of ancient languages using probabilistic models of sound change," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, no. 11, p. 42244229, 2013.

[2] D. J. Hruschka, S. Branford, E. D. Smith, J. Wilkins, A. Meade, M. Pagel, and T. Bhattacharya, "Detecting regular sound changes in linguistics as events of concerted evolution," *Curr. Biol.*, vol. 25, no. 1, pp. 1–9, Jan 2015.

[3] A. Fox, *Linguistic reconstruction*. Oxford: Oxford University Press, 1995.

[4] R. Bouckaert, P. Lemey, M. Dunn, S. J. Greenhill, A. V. Alekseyenko, A. J. Drummond, R. D. Gray, M. A. Suchard, and Q. D. Atkinson, "Mapping the origins and expansion of the Indo-European language family," *Science*, vol. 337, no. 6097, pp. 957–960, Aug 2012.

[5] S. J. Greenhill, R. Blust, and R. D. Gray, "The austronesian basic vocabulary database: From bioinformatics to lexomics," *Evolutionary Bioinformatics*, vol. 4, pp. 271–283, 2008.

[6] A. J. Drummond, M. A. Suchard, D. Xie, and A. Rambaut, "Bayesian phylogenetics with BEAUti and the BEAST 1.7," *Molecular Biology and Evolution*, vol. 29, no. 8, pp. 1969–1973, Aug 2012.

[7] D. Sankoff, "Minimal mutation trees of sequences," *SIAM Journal on Applied Mathematics*, vol. 28, no. 1, pp. 35–42, January 1975. [Online]. Available: http://www.jstor.org/stable/2100459

[8] J.-M. List, S. Nelson-Sathi, H. Geisler, and W. Martin, "Networks of lexical borrowing and lateral gene transfer in language and genome evolution," *Bioessays*, vol. 36, no. 2, pp. 141–150, 2014.