Are Sounds Sound for the Reconstruction of Language Trees? Comparing Lexical Cognates and Sound Correspondences in Bayesian Phylogenetic Inference

Abstract

In traditional studies on language evolution, scholars often emphasize the important role which sound laws and sound correspondences play in the task of reconstructing the phylogeny of a language family. However computational approaches have largely ignored the potential importance of sound laws and sound correspondences. Most computational studies still employ lexical cognates as the major data when it comes to phylogenetic reconstruction in linguistics, although there are a few studies in which authors praise the benefits of comparing words at the level of sound sequences. Building on (a) ten diverse datasets from different language families, and (b) state-of-the-art methods for automated cognate and sound correspondence detection, we test, for the first time, the performance of sound-based versus cognate-based approaches to phylogenetic reconstruction. Our results show that phylogenies reconstructed from lexical cognates tend to come closer to gold standard phylogenies than phylogenies reconstructed from sound correspondences.

Keywords: sound correspondences, phylogenetic reconstruction lexical cognates

1. Introduction

Although controversially discussed in the beginning (Holm, 2007), quantitative approaches to phylogenetic reconstruction based on Bayesian phylogenetic inference frameworks have now become broadly accepted in the field of comparative linguistics. This is reflected in more and more computer-based phylogenies that have been proposed for the world's largest language families—Dravidian (Kolipakam et al., 2018), Sino-Tibetan (Sagart et al., 2019) and Indo-European (Bouckaert et al., 2012)—and even fully automated workflows have shown to be quite robust (Rama et al., 2018). While rarely practiced in the pre-computational past of historical linguistics, computing detailed phylogenies has now become one of the key tasks of studies on language evolution.

Although traditional scholars have started to accept computational language phylogenies as a new tool deserving its place in the large tool chain of comparative linguistics, scholars still express a lot of skepticism against the idea of most of the language phylogenies that have been proposed so far. One of the major reasons usually mentioned in this context is that phylogenetic approaches are usually based on cognate sets (sets of historically related words) identified in semantically aligned word lists. Since these *cognate sets* reflect *lexical data* only, many scholars mistrust them, given that lexical data are assumed to be much less stable than other aspects of languages (Campbell and Poser, 2008).

In classical historical linguistics, the data used for subgrouping are traditionally composed of small collections of so-called *shared innovations* (Dyen, 1953). What counts as a shared innovation has itself never been well-defined in the literature, but the largest amount of data used by scholars is traditionally taken from sound correspondences or supposed sound change processes (compare, for example the data in Anttila 1972, 305). Although it is controversially debated in the field (Ringe et al., 2002; Dybo and Starostin, 2008), many classical linguists still emphasize that sound correspondences are largely superior to lexical data when it comes to subgrouping.

There have only been a few attempts to test how well quantitative approaches to phylogenetic reconstruction perform when using sound correspondences instead of lexical cognates (Chacon and List, 2015). The main reason is that coding data to compute phylogenies from sound change data is very tedious even for a dataset with 20 languages. For this reason, there have only been a few attempts to test how well quantitative approaches to phylogenetic reconstruction perform when using sound correspondences instead of lexical cognates.

Building on the state-of-the-art methods for automatic cognate detection and phonetic alignment in historical linguistics (List et al., 2016), combined with novel approaches for the inference of sound correspondence patterns in multilingual datasets (List, 2019) and customized solutions for phylogenetic reconstruction (Rama and List, 2019) that

have evolved into a new Python library for phylogenetic inference, we have been able to create a new workflow for phylogenetic reconstruction based on sound correspondence patterns. With a new collection of ten gold standard datasets, we test the workflow and compare it with alternative workflows based on lexical data alone. Our results indicate that sound correspondence patterns are far less suitable for the purpose of computer-based phylogenetic reconstruction than expected.

2. Background

Much of the previous work on phylogenetic reconstruction using Bayesian phylogenetic inference (Kolipakam et al., 2018; Sagart et al., 2019; Rama et al., 2018) is based on cognate sets encoded as binary vectors, where the presence or absence of a language in a cognate set is coded as **0** or **1**, and phylogenetic trees are inferred by assuming that cognate sets evolve along a phylogenetic tree in the form of gain and loss processes (see Figure 1).

The binary-state coding is the most frequently applied coding technique, and it is being used in this study as well. Once assembled, binary state data can be modeled with binary state Continuous Time Markov Chain model (*binary-CTMC*, Bouckaert et al. 2012), which allow gain and loss processes to occur an arbitrary number of times.

The *major contributions* of this study are: (1) We provide an automated workflow that allows to infer cognates and correspondence patterns and analyze them with the help of Bayesian phylogenetic inference methods, implemented in a new software package, (2) we show how the quality of phylogenetic reconstruction approaches based on sound correspondences can be compared to phylogenetic reconstruction based on lexical data, and in this way (3) we put the debate about the usefulness of sound-based as opposed to cognate-based phylogenies to the test.

As an early example for sound-based approaches to phylogenetic reconstruction, Hruschka et al. (2015) apply a CTMC model that allows transitions between a fixed number of sounds for detecting the important sound changes in a dataset consisting of etymologies across Turkic languages. The authors do not infer phylogenies from their data but rather use an established phylogeny (which are not readily available for many language families of the world) to infer branch lengths and transition probabilities between sounds in their data in order to detect sound changes at different time points in a time-calibrated family tree of Turkic.

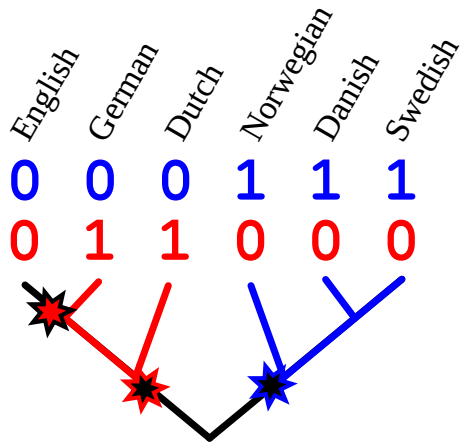
Wheeler and Whiteley (2015) start from typical word lists (that would otherwise be used in phylogenetic reconstruction based on lexical data) and apply a parsimony-based algorithm that aligns

Language	Concept	Form	Cog-Set
English	"big"	big	1
German	"big"	groß	2
Dutch	"big"	groot	2
Norwegian	"big"	stor	3
Danish	"big"	stor	3
Swedish	"big"	stor	3

(A) multi-state matrix

Concept		"big"		
Cog-Set		1	2	3
English	big	0	0	0
German	groß	0	1	0
Dutch	groot	0	1	0
Norweg.	stor	0	0	1
Danish	stor	0	0	1
Swedish	stor	0	0	1

(B) binary-state matrix



(C) evolutionary scenario (binary-state)

Figure 1: Gain-loss processes derived from binary cognate vectors. A shows a wordlist in which cognate words are coded in multi-state fashion. B shows the corresponding binary coding. C shows how gain and loss processes are modeled on a phylogenetic tree.

words regardless if they are cognate or not, reconstructs a hypothetical ancestral word from the alignment, and seeks to infer the phylogeny that allows to explain the sequences by a minimal amount of assumed transitions (Sankoff, 1975). In a later study, Whiteley et al. (2019) apply the same approach to a dataset of Bantu languages. The method by Wheeler and Whiteley (2015) is linguistically flawed, since words are not assigned to cognate sets before aligning them. It is well known that there is a strict difference between reg-

ular sound change processes and processes resulting from lexical replacement (Hall and Klein, 2010) and that even words that are cognate are not necessarily fully *alignable* (Schweikhard and List, 2020, 10).
Chacon and List (2015) start from manually extracted sound correspondence patterns for consonants in a dataset of 21 Tukanoan languages, to which proto-forms had also been added manually. Based on the sound correspondence patterns, they apply—similar to Wheeler and Whiteley (2015)—an algorithm that searches for the tree that provides the most parsimonious scenario for the evolution of the sounds. In contrast to Wheeler and Whiteley (2015), however, they added specific constraints for the transitions from one sound to another sound, which were based on expert judgments for the Tukanoan language family. The approach by Chacon and List (2015), finally, requires an enormous amount of preprocessing that runs the risk of leading to circular results, since proto-forms and major directions of sound change processes are required to be known in advance. While all approaches have their individual shortcomings, one of the largest shortcomings lies in the fact that it is very difficult to apply them. This is also witnessed by the fact that no additional studies have been carried out by other teams, although all the methods have been proposed years ago.

3. Materials and Methods

3.1. Materials

- describe datasets (table)
- describe preprocessing
- describe binarization

3.2. Methods

Phylogenetic inference was carried out using the software *MrBayes* (Ronquist and Huelsenbeck, 2003), version 3.2.7. We used the following prior settings for all datasets:

- Dirichlet$(1.0, 1.0)$ prior for base frequencies
- gamma-distributed rates, approximated by 4 discrete categories, with a standard exponential prior distribution over the shape of the gamma distribution,
- uniform prior over tree topologies, and
- relaxed clock model of branch lengths using the *independent gamma rates* model, with an exponential distribution with rate 10.0 as prior for the variance of the gamma distribution.

Every 1,000th generation, the current state of the Markov chain was sampled. MCMC chains were stopped when the average standard deviation of split frequencies was below 0.01 after discarding the first 25% of the samples. From the remaing 75% of the recorded samples from all four chains, 1,000 trees were drawn at random and used for further evaluation.
For each dataset, phylogenetic inference was performed on three character matrices:

- **cognate classes:** Each cognate class is one binary character.
- **sound correspondences:** Each sound correspondence pattern is one binary character.
- **combined:** The character matrices for both character types are combined.

We considered three hypotheses:

- **Hypothesis 1:** Phylogenetic inference based on sound correspondences is more accurate than phylogenetic inference based on cognate classes.
- **Hypothesis 2:** Phylogenetic inference based on sound correspondences is more accurate than phylogenetic inference based on cognate classes and sound correspondences combined.
- **Hypothesis 3:** Both character types capture the same signal.

If one of the two individual character types provides the best results, this would be evidence for Hypothesis 1 or Hypothesis 2. If the combined dataset provides the best results, this would be evidence for Hypothesis 3.
To evaluate the quality of the inferred phylogenies, we used the classifications from Glottolog (Hammarström et al., 2019). The degree of consistency of a binary-branching inferred phylogeny and a (possible polytomous) Glottolog tree was measured as the *generalized quartet distance* (GQD), as proposed in (Pompei et al., 2011).[1]

4. Results

The results of the evaluation are shown in Table 4. As can be seen from the table, phylogenetic inference from cognate class characters outperforms the other two options for eight out of ten datasets. On average, cognate classes evidently provide the best result. This provides strong evidence for Hypothesis 1, and against the other two hypotheses.

[1] The GQD is a generalization of the well-known *quartet distance* (Estabrook et al., 1985) that allows to compare binary-branching trees with polytomous trees. The GQD is defined as the number of quartets that are not shared between the two trees, divided by the number of all possible quartets. The GQD is a number between 0 and 1, where 0 means that the two trees are identical, and 1 means that the two trees are completely different.

dataset	cognate classes	sound correspondences	combined
constenlachibchan	**0.269**	0.355	0.352
crossandean	**0.148**	0.561	0.494
dravlex	**0.323**	0.409	0.326
felekesemitic	**0.083**	0.119	0.103
hattorijaponic	0.585	**0.388**	0.511
houchinese	**0.240**	0.428	0.441
leekoreanic	**0.215**	0.298	0.253
robinsonap	0.424	**0.348**	0.359
walworthpolynesian	**0.179**	0.328	0.331
zhivlovobugrian	**0.330**	0.356	0.356
median	**0.254**	0.356	0.354

Table 1: Generalized quartet distances (posterior medians) for the three character types. The best result for each dataset is highlighted in bold.

5. Implementation

Mattis + Gerhard

6. Discussion and Conclusion

7. Supplementary Material

Raimo Anttila. 1972. *An introduction to historical and comparative linguistics*. Macmillan, New York.

Remco Bouckaert, Philippe Lemey, Michael Dunn, Simon J Greenhill, Alexander V Alekseyenko, Alexei J Drummond, Russell D Gray, Marc A Suchard, and Quentin D Atkinson. 2012. Mapping the origins and expansion of the Indo-European language family. *Science*, 337(6097):957–960.

Lyle Campbell and William J. Poser. 2008. *Language Classification: History and Method*. Cambridge University Press.

Thiago Costa Chacon and Johann-Mattis List. 2015. Improved computational models of sound change shed light on the history of the Tukanoan languages. *Journal of Language Relationship*, 13(3):177–204.

Anna Dybo and George S Starostin. 2008. In defense of the comparative method, or the end of the Vovin controversy. In I. S. Smirnov, editor, *Aspekty komparativistiki*, pages 119–258. RGGU, Moscow.

Isidore Dyen. 1953. [Review] Malgache et maanjan: Une comparaison linguistique by Otto Chr. Dahl. *Language*, 29(4):577–590.

George F Estabrook, FR McMorris, and Christopher A Meacham. 1985. Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units. *Systematic Zoology*, 34(2):193–200.

David Hall and Dan Klein. 2010. Finding cognate groups using phylogenies. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1030–1039. Association for Computational Linguistics.

Harald Hammarström, Martin Haspelmath, and Robert Forkel. 2019. *Glottolog. Version 4.1*. Max Planck Institute for the Science of Human History, Jena.

Hans J. Holm. 2007. The new arboretum of Indo-European trees. *Journal of Quantitative Linguistics*, 14(2-3):167–214.

Daniel J Hruschka, Simon Branford, Eric D Smith, Jon Wilkins, Andrew Meade, Mark Pagel, and Tanmoy Bhattacharya. 2015. Detecting regular sound changes in linguistics as events of concerted evolution. *Current Biology*, 25(1):1–9.

Vishnupriya Kolipakam, Fiona M. Jordan, Michael Dunn, Simon J. Greenhill, Remco Bouckaert, Russell D. Gray, and Annemarie Verkerk. 2018. A Bayesian phylogenetic study of the Dravidian language family. *Royal Society Open Science*, 5(171504):1–17.

Johann-Mattis List. 2019. Automatic inference of sound correspondence patterns across multiple languages. *Computational Linguistics*, 1(45):137–161.

Johann-Mattis List, Philippe Lopez, and Eric Bapteste. 2016. Using sequence similarity networks to identify partial cognates in multilingual wordlists. In *Proceedings of the Association*

of Computational Linguistics 2016 (Volume 2: Short Papers), pages 599–605, Berlin. Association of Computational Linguistics.

Simone Pompei, Vittorio Loreto, and Francesca Tria. 2011. On the accuracy of language trees. *PloS one*, 6(6):e20109.

Taraka Rama and Johann-Mattis List. 2019. An automated framework for fast cognate detection and Bayesian phylogenetic inference in computational historical linguistics. In *57th Annual Meeting of the Association for Computational Linguistics*, page 6225–6235. Association for Computational Linguistics.

Taraka Rama, Johann-Mattis List, Johannes Wahle, and Gerhard Jäger. 2018. Are automatic methods for cognate detection good enough for phylogenetic reconstruction in historical linguistics? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 393–400.

Donald Ringe, Tandy Warnow, and Ann Taylor. 2002. Indo-European and computational cladistics. *Transactions of the Philological Society*, 100(1):59–129.

Frederik Ronquist and John P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19(12):1572–1574.

Laurent Sagart, Guillaume Jacques, Yunfan Lai, Robin Ryder, Valentin Thouzeau, Simon J. Greenhill, and Johann-Mattis List. 2019. Dated language phylogenies shed light on the ancestry of sino-tibetan. *Proceedings of the National Academy of Science of the United States of America*, 116:10317–10322.

David Sankoff. 1975. Minimal mutation trees of sequences. *SIAM Journal on Applied Mathematics*, 28(1):35–42.

Nathanael E. Schweikhard and Johann-Mattis List. 2020. Developing an annotation framework for word formation processes in comparative linguistics. *SKASE Journal of Theoretical Linguistics*, 17(1):2–26.

W. C. Wheeler and Peter M. Whiteley. 2015. Historical linguistics as a sequence optimization problem: the evolution and biogeography of uto-aztecan languages. *Cladistics*, 31(2):113–125.

Peter M. Whiteley, Ming Xue, and Ward C. Wheeler. 2019. Revising the bantu tree. *Cladistics*, 35:329–348.