# 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) five diverse datasets from five 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 tophylogenetic 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. Since coding data to compute phylogenies from sound change data is very tedious—specifically when working with more than just a few languages—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 five 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. 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 (details are given in Section **??**). While linguists tend to prefer these models intuitively, and there has been some debate about the limits of binary-state coding (Atkinson and Gray, 2006; Pagel and Meade, 2006; List, 2016). It is well known that a multitude of processes can lead to *lexical replacement*, including the typical transition through a synonymous phase in which a concept can be expressed by two or more word forms, and various derivation processes.

An alternative coding technique is to treat each concept in a wordlist as a single character and to allow for each character to have a range of different states. In contrast, it is not clear what the *transition* between multiple states is supposed to reflect when modeling character evolution on multi-state data with CTMC models. While multi-state models are rarely applied to lexical data, they have an immediate appeal for phonological data on sound correspondences and sound change processes, since it is well known that the sounds reflected in particular correspondence patterns can be numerous across larger language families.

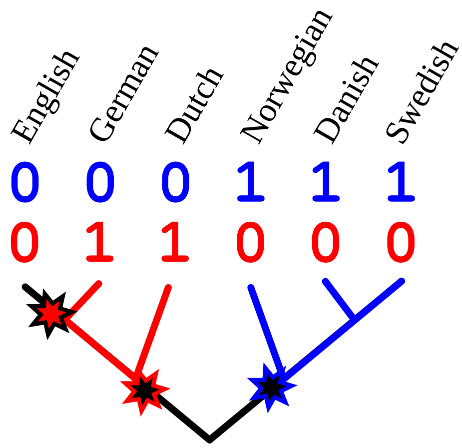
The *major contributions* of this study are: (1) we provide an automated workflow that allows to infer cognates and correspondence patterns and analyze them with the help of Bayesian phylogenetic inference methods, implemented in a new

| **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.

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 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 regular 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

mostly gerhard on trees, etc.

## 3.3. Evaluation

mostly gerhard on trees

## 3.4. Implementation

Mattis + Gerhard

# 4. Results

Gerhard + mattis comments

# 5. Discussion and Conclusion

# 6. Supplementary Material


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

Quentin D Atkinson and Russell D Gray. 2006. How old is the Indo-European language family? illumination or more moths to the flame. *Phylogenetic methods and the prehistory of languages*, 91:109.

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.

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.

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. 2016. Beyond cognacy: Historical relations between words and their implication for phylogenetic reconstruction. *Journal of Language Evolution*, 1(2):119–136.

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.

Mark Pagel and Andrew Meade. 2006. Estimating rates of lexical replacement on phylogenetic trees of languages. In P. Forster and C. Renfrew, editors, *Phylogenetic methods and the prehistory of languages*, pages 173–182. McDonald institute Monographs.

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.

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.

# A. Appendix: How to Produce the `.pdf`

**Submissions may be of three types:**

- Regular long papers - up to eight (8) pages maximum,* presenting substantial, original, completed, and unpublished work.
- Short papers - up to four (4) pages,[1] describing a small, focused contribution, negative results, system demonstrations, etc.
- Position papers - up to eight (8) pages,* discussing key hot topics, challenges and open issues, and cross-fertilization between computational linguistics and other disciplines.

Upon acceptance, final versions of long papers will be given one additional page – up to nine (9) pages of content plus unlimited pages for acknowledgments and references – so that reviewers' comments can be considered. Final versions of short papers may have up to five (5) pages, plus unlimited pages for acknowledgments and references. All figures and tables that are part of the main text must fit within these page limits for long and short papers.

Papers must be of original, previously-unpublished work. Papers must be **anonymized to support double-blind reviewing**. Submissions, thus, must not include authors' names and affiliations. The submissions should also avoid links to non-anonymized repositories: the code should be either submitted as supplementary material in the final version of the paper or as a link to an anonymized repository (e.g., Anonymous GitHub or Anonym Share). Papers that do not conform to these requirements will be rejected without review.

# B. Final Paper

Each final paper should be submitted online. The fully justified text should be formatted according to LREC-COLING2024 style as indicated for the Full Paper submission.

As indicated above, the font for the main body of the text should be Times New Roman 10 pt with interlinear spacing of 11 pt. Papers must be between 4 and 8 pages long, including figures (plus more pages for references if needed), regardless of the presentation mode (oral or poster).

## B.1. General Instructions for the Final Paper

The unprotected PDF files will appear in the online proceedings directly as received. **Do not print the page number**.

[1]Excluding any number of additional pages for references, ethical consideration, conflict-of-interest, as well as data and code availability statements.

| Output | natbib command | Old command |
|---|---|---|

Table 1: Citation commands supported by the style file. The style is based on the natbib package and supports all natbib citation commands. It also supports commands defined in previous style files for compatibility.

# C. Page Numbering

**Please do not include page numbers in your Paper.** The definitive page numbering of papers published in the proceedings will be decided by the Editorial Committee.

# D. Headings / Level 1 Headings

Level 1 Headings should be capitalised in the same way as the main title, and centered within the column. The font used is Arial 12 pt bold. There should also be a space of 12 pt between the title and the preceding section and 3 pt between the title and the following text.

## D.1. Level 2 Headings

The format of Level 2 Headings is the same as for Level 1 Headings, with the font Arial 11 pt, and the heading is justified to the left of the column. There should also be a space of 6 pt between the title and the preceding section and 3 pt between the title and the following text.

### D.1.1. Level 3 Headings

The format of Level 3 Headings is the same as Level 2, except that the font is Arial 10 pt, and there should be no space left between the heading and the text as in D.1.1. There should also be a space of 6 pt between the title and the preceding section and 3 pt between the title and the following text.

# E. Citing References in the Text

## E.1. Bibliographical References

Table 1 shows the syntax supported by the style files. We encourage you to use the natbib styles. You can use the command `\citet` (cite in text) to get "author (year)" citations, like this citation to a paper by

When several authors are cited, those references should be separated with a semicolon:

## E.2. Language Resource References

### E.2.1. When Citing Language Resources

See Appendix A for details on how to produce this with bibtex.

# F. Figures & Tables

## F.1. Figures

All figures should be centred and clearly distinguishable. They should never be drawn by hand, and the lines must be very dark in order to ensure a high-quality printed version. Figures should be numbered in the text, and have a caption in Arial 10 pt underneath. A space must be left between each figure and its respective caption.
Example of a figure enclosed in a box:

Figure 2: The caption of the figure.

Figure and caption should always appear together on the same page. Large figures can be centered, using a full page.

## F.2. Tables

The instructions for tables are the same as for figures.

| Level | Tools |
|---|---|
| Morphology | Pitrat Analyser |
| Syntax | LFG Analyser (C-Structure) |
| Semantics | LFG F-Structures + Sowa's Conceptual Graphs |

Table 2: The caption of the table

# G. Footnotes

Footnotes are indicated within the text by a number in superscript[2].

# H. Copyrights

The Language Resouces and Evaluation Conference (LREC) Proceedings are published by the European Language Resources Association (ELRA). They are available online from the conference website.
ELRA's policy is to acquire copyright for all LREC contributions. In assigning your copyright, you are not forfeiting your right to use your contribution elsewhere. This you may do without seeking permission and is subject only to normal acknowledgment to the LREC proceedings. The LREC Proceedings are licensed under CC-BY-NC, the Creative Commons Attribution-Non-Commercial 4.0 International License.

# I. Conclusion

Your submission of a finalized contribution for inclusion in the LREC Proceedings automatically assigns the above copyright to ELRA.

# J. Acknowledgements

Place all acknowledgments (including those concerning research grants and funding) in a separate section at the end of the paper.

# K. Optional Supplementary Materials

Appendices or supplementary material (software and data) will be allowed ONLY in the final, camera-ready version, but not during submission, as papers should be reviewed without the need to refer to any supplementary materials.
Each **camera ready** submission can be accompanied by an appendix usually being included in a main PDF paper file, one `.tgz` or `.zip` archive containing software, and one `.tgz` or `.zip` archive containing data.
We encourage the submission of these supplementary materials to improve the reproducibility of results and to enable authors to provide additional information that does not fit in the paper. For example, preprocessing decisions, model parameters, feature templates, lengthy proofs or derivations, pseudocode, sample system inputs/outputs, and other details necessary for the exact replication of the work described in the paper can be put into the appendix. However, the paper submissions must remain fully self-contained, as these supplementary materials are optional, and reviewers are not even asked to review or download them. If the pseudo-code or derivations, or model specifications are an essential part of the contribution, or if they are important for the reviewers to assess the technical correctness of the work, they should be a part of the main paper and not appear in the appendix. Supplementary materials need to be fully anonymized to preserve the double-blind reviewing policy.

## K.1. Appendices

Appendices are material that can be read and include lemmas, formulas, proofs, and tables that are not critical to the reading and understanding of the paper, as in *ACLPUB. It is highly recommended that the appendices should come after the references; the main text and appendices should

[2] Footnotes should be in Arial 9 pt, and appear at the bottom of the same page as their corresponding number. Footnotes should also be separated from the rest of the text by a 5 cm long horizontal line.

be contained in a ‘single’ manuscript file, without being separately maintained. Letter them in sequence and provide an informative title: *Appendix A. Title of Appendix*

### K.2. Extra space for ethical considerations and limitations

Please note that extra space is allowed after the 8th page (4th page for short papers) for an ethics/broader impact statement and a discussion of limitations. At submission time, if you need extra space for these sections, it should be placed after the conclusion so that it is possible to rapidly check that the rest of the paper still fits in 8 pages (4 pages for short papers). Ethical considerations sections, limitations, acknowledgments, and references do not count against these limits. For camera-ready versions, nine pages of content will be allowed for long (5 for short) papers.

## L. Providing References

### L.1. Bibliographical References

Bibliographical references should be listed in alphabetical order at the end of the paper. The title of the section, “Bibliographical References”, should be a Level 1 Heading. The first line of each bibliographical reference should be justified to the left of the column, and the rest of the entry should be indented by 0.35 cm.
The examples provided in Section M (some of which are fictitious references) illustrate the basic format required for papers in conference proceedings, books, journal articles, PhD theses, and books chapters.

### L.2. Language Resource References

Language resource references should be listed in alphabetical order at the end of the paper.

## M. Bibliographical References

## N. Language Resource References

In order to generate a PDF file out of the LaTeX file herein, when citing language resources, the following steps need to be performed:

```
xelatex paper.tex
bibtex paper.aux
bibtex languageresource.aux
xelatex paper.tex
xelatex paper.tex
```