Computer-Assisted Language Comparison in Practice

Tutorials on Computational Approaches to the History and Diversity of Languages

Lexibench: Towards an Improved Collection of Benchmark Data for Computational Historical Linguistics

Luise Häuser / Johann-Mattis List
Computational Molecular Evolution Group / Chair for Multilingual Computational Linguistics
Heidelberg Institute for Theoretical Studies / University of Passau

Computational approaches in historical linguistics have made great progress during the past two decades. As of now, it is much more common to propose subgroupings based on phylogenetic analyses than on traditional considerations using shared innovations. We have also seen a drastic increase in openly available datasets that share cognate judgments for various language families. Thanks to new standardization efforts providing facilitated access to several dozen comparative wordlists, it seems about time to work on on improved benchmarks of manually annotated cognates in computational historical linguistics. In this study, a first effort of this kind is undertaken, by presenting Lexibench, a preliminary gold standard for computational historical linguistics. Lexibench builds on the Lexibank repository to extract 63 multilingual wordlists, all manually annotated for cognacy, that can be used to assess the quality of cognate detection and phylogenetic reconstruction methods in computational historical linguistics.

1 Introduction

Benchmark datasets play an important role in the evaluation of computational methods in computer science in general and computational historical linguistics in specific. Specifically when it comes to the typical tasks that mimic individual steps of the workflow underlying the traditional comparative method (Meillet 1925), such as phonetic alignment, cognate detection, and phonological reconstruction (see List 2014), benchmark datasets are of great importance, since they help us to assess how well the automated methods that have been proposed so far work in comparison to each other and in comparison to experts annotating the data manually.

Computer-Assisted Language Comparison in Practice
Volume 8, Number 1
URL: https://calc.hypotheses.org/8227 DOI: 10.15475/calcip.2025.1.2

Published on 26/02/2025

In the following, we will describe initial efforts to make use of recently published standardized data collections in order to establish a new benchmark database consisting of annotated multilingual wordlists that can be used to test and compare existing and new approaches in computational historical linguistics. This endeavor, the Lexibench collection of benchmark data for computational historical linguistics, should be considered as work in progress, since we are currently still testing the best ways to create the database, discussing what datasets to include, which statistics to compute, and how to make sure that datasets conform to our expectations. In the following, we will give a short backgrounds on the role of benchmark datasets in historical linguistics and then describe, how benchmark data for cognate detection are acquired and curated. After providing a few examples on the current state of the collection, we conclude by pointing to future challenges that we hope to address soon.

2 Background

Along with the quantitative turn in historical linguistics, many computational methods have been proposed to handle long-standing problems in historical language comparison. These methods include computational approaches to phonetic alignment (Kondrak 2000, Prokić et al. 2009, List 2014, Kilani 2020), automated methods for cognate detection (List 2014, Jäger et al. 2017, Dellert 2017), and computational methods to phylogenetic reconstruction applied to smaller and larger language families (Gray and Atkinson 2003, Sagart et al. 2019, Dhakal et al. 2024).

While computational methods have improved remarkably over the past two decades, dedicated benchmarks — that is, gold standard datasets — that would help scholars to test novel and existing algorithms rigorously, have been rarely proposed so far. Thus, while the benchmark data for phonetic alignments by by List and Prokić (2014) could be mentioned as a rare exception, the data has not been further modified or extended since it was published in 2014. While several smaller benchmarks have been compiled for individual studies on cognate detection (List et al. 2017, Blum and List 2023), phonological reconstruction (List et al. 2022a), automated workflows for phylogenetic reconstruction (Rama et al. 2018, Häuser et al. 2024), or reflex prediction (List et al. 2022b), there have been no efforts to compile and curate benchmark datasets independently of individual studies that would test novel algorithms.

Standardized lexical data collections have been growing in size during the past decade, with the Lexibank repository (List et al. 2022c, Blum et al. 2025) providing standardized multilingual wordlists for more than 2000 language varieties. It therefore seems that it is time to harvest the available data to establish a first set of multilingual datasets with manual cognate annotation that can be used as a benchmark to test the quality of new and existing methods in historical language comparison that address problems on cognate detection and phylogenetic reconstruction.

3 Lexibench

The major idea of Lexibench (Häuser and List 2025) is to build upon the Lexibank repository with its large collection of standardized wordlists to establish a benchmark dataset that can be used to test existing and new methods in computational historical linguistics, including methods for automated cognate detection, methods for phonetic alignment analysis and correspondence pattern identification, and more integrated methods in which computational workflows for phylogenetic reconstruction are explored.

Our starting point are those datasets in LexiBank that are supplemented by manually annotated cognates (assigned to the CogCore part of Lexibank). From this selection, use an automated workflow to extract multilingual wordlists (§ 3.1), unify the representation of cognate sets (§ 3.2), split the data by language family (§ 3.3), and then filtering the data based on basic wordlist properties (§ 3.4).

3.1 Data Extraction

We follow the workflow that was first designed for the handling of the data in Lexibank by using a base list that offers the most recent versions of all datasets that we want to include in Lexibench. Based on this list (provided as a TSV file), users can then download the data in Cross-Linguistic Data Formats (CLDF, Forkel et al. 2018). CLDF has the advantage of standardizing language names (by linking languages to Glottolog, Hammarström et al. 2024), concept glosses (by mapping glosses to Concepticon, List et al. 2025a), and by providing standardized phonetic transcriptions using the Cross-Linguistic Transcription Systems, a subset of the IPA, that defines existing sounds explicitly in a generative manner (List et al. 2024, Anderson et al. 2018). As a result, Lexibench data can be automatically derived from Lexibank data, with minimal data curation steps required.

3.2 Unifying the Representation of Cognate Sets

Although cognate set representation is standardized in CLDF, the individual cognate sets provided in Lexibank are pretty diverse. In some cases, cognate sets are assigned for individual concept slots. This means, that the identifiers only work within the scope of one concept slot. Ignoring this annotation can lead to confusion and erroneous results. For example, when treating cognate set identifiers with a local scope –restricted to individual concept slots — as global, one may easily end up falsely grouping words into the same cognate set when their cognate set identifiers recur across different concepts.For this reason, we apply an automated curation step when creating the Lexibench collection. All cognate sets are modeled as global cognate set identifiers, but cognates that span different concepts are not allowed at this stage. In the future, we plan to extend this representation by adding datasets in which cognates spanning several different concepts ([as tested, for example, by Wu et al. 2020) are also annotated.

3.3 Splitting Datasets by Language Families

In most approaches, we want to identify cognate sets within individual language families. Some datasets with annotated cognate sets, however, provide data on several language families, including at times annotated borrowings or supposed deep genetic relations. In these cases, we prefer — at least at this stage — to split the data into individual subsets consisting of only one family per subset. If we manage to keep Lexibench evolving in the future, we may consider adding specific datasets that help to identify borrowings across language families (as illustrated by List and Forkel 2022, and Miller and List 2023), but for now, our focus is on individual language families.

3.4 Filtering Data Based on Wordlist Properties

In order to make sure that our wordlist extraction is reasonable, and that the selection contains datasets that are apt for basic phylogenetic approaches, we compute basic statistics for each dataset. These include not only the number of language varieties (we require that the data contains at least 5 different varieties), or the number of different concepts (with a minimum of 100 concepts per dataset), but also important characteristics like the average coverage of a wordlist (List et al. 2018), reflecting the percentage of concepts for which any pair of languages in a given dataset has entries. Average coverage is supposed to be important for phylogenetic reconstruction (Sagart et al. 2019, and also crucial for automated methods for cognate detection, since it indicates the amount of word pairs that can be compared for individual language pairs. In our first demo version of LexiBench, we set the threshold for the average coverage to 0.5.

3.5 Implementation

The major workflow by which data in LexiBench are assembled works in a fully automated manner. The initial selection of datasets form Lexibank is hand-curated and can be modified. The selection of subsets of languages, however, is automated using a Python script that reads in all data, computes basic statistics, and converts the data from the Cross-Linguistic Data Formats to a tab-separated wordlist in the format required by LingPy (List and Forkel 2024 and EDICTOR (List et al. 2025b, List and van Dam 2024).

The Lexibench repository provides a Python script and additional illustrations that explain how users can replicate the database creation and modify it according to their own needs.

4 Overview on Lexibench

4.1 Basic Statistics

Lexibench currently assembles data from 65 different datasets from the Lexibank repository (Version 2.0, Blum et al. 2025). After applying the automated data curation and selection procedure, by which datasets are split into individual language families, and basic thresholds for the number of languages, the number of concepts, and the average coverage of comparative word pairs have been applied, this leaves us with a total of 63 individual datasets from 27 different language families. The largest dataset in the sample consists of more than 400 language varieties (abvdoceanic), and the average number of languages varieties is 34. The largest concept list consists of 994 entries (yanglalo), with an average of 204 concepts per wordlist. The average coverage of all datasets is considerably high with 0.86, and the datasets have about 1944 cognate sets on average.

4.2 Cognate Detection Baseline

In order to illustrate and test the usefulness of a benchmark database like Lexibench, we applied two cognate detection methods to parts of the data in this first version of Lexibench. The SCA approach uses the Sound Class based phonetic Alignment algorithm (List 2012a) in order to compute phonetic distances between word pairs, which are in turn clustered into cognate sets (see List et al. 2017 for a detailed description). The LexStat approach (List 2012b) adds an additional layer of complexity to this workflow by computing pairwise sound correspondence probabilities before computing phonetic distances, thus accounting to some degree for regular sound correspondences (see also List et al. 2017). While the SCA approach was applied to all data, the LexStat approach was only applied to those datasets consisting of maximally 50 language varieites, given the increased computation time for larger datasets with the LexStat approach. Applying a base threshold of a distance of 0.45 for the SCA approach and of 0.55 for the LexStat approach, we can compute precision, recall, and F-Score, using B-Cubed scores to evaluate cognate detection performance (Amigo et al. 2009).

The results for this test are shown in Table 1. As can be seen from this table, both methods perform reasonably well, reaching F-Scores of more than 80%, with the LexStat approach outperforming the SCA method by 3 points.

Method	Precision	Recall	F-Score
SCA	0.89	0.81	0.85
LexStat	0.92	0.84	0.88

Table1: Results of the cognate detection test on 60 wordlist.

This shows on the one hand that all datasets are in a state that they can be directly analyzed with the help of software packages like LingPy. On the other hand, it also shows that the SCA approach in all its computational simplicity provides a good approximation of expert judgments and can therefore be considered a useful baseline for future developments in the task of automated cognate detection.

5 Outlook

The Lexibench repository provides one of the largest collections of manually annotated cognate sets in multilingual wordlists that has been compiled so far. It may therefore prove useful for all those who want to develop new methods for computational historical linguistics, including methods for phonetic alignment, cognate detection, or borrowing detection. In its current form, however, the repository is still work in progress and should be used with a certain care. While we are confident that errors in the data have been minimized, we cannot promise that there won’t be no errors at all with the data. In the future, we hope we can improve the data further by computing more statistics on the existing wordlists, testing more methods for cognate detection, establishing a pool of baseline results, and by improving the workflow for data creation.

References

Amigó, Enrique and Gonzalo, Julio and Artiles, Javier and Verdejo, Felisa (2009): A comparison of extrinsic clustering evaluation metrics based on formal constraints. Information Retrieval 12.4. 461-486.

Anderson, Cormac and Tresoldi, Tiago and Chacon, Thiago Costa and Fehn, Anne-Maria and Walworth, Mary and Forkel, Robert and List, Johann-Mattis (2018): A Cross-Linguistic Database of Phonetic Transcription Systems. Yearbook of the Poznań Linguistic Meeting 4.1. 21-53. https://doi.org/10.2478/yplm-2018-0002

Blum, Frederic and List, Johann-Mattis (2023): Trimming phonetic alignments improves the inference of sound correspondence patterns from multilingual wordlists. In: Proceedings of the 5th Workshop on Computational Typology and Multilingual NLP. Association for Computational Linguistics 52-64. https://aclanthology.org/2023.sigtyp-1.6

Blum, Frederic, Carlos Barrientos, Johannes Englisch, Robert Forkel, Russell D. Gray, Simon J. Greenhill, Christoph Rzymski, and Johann-Mattis List (2025): Lexibank²: Precomputed Features for Large-Scale Lexical Data. Geneva: Zenodo. https://doi.org/10.5281/zenodo.14800315

Dellert, Johannes (2017): Information-theoretical causal inference of lexical flow . PhD. Eberhard-Karls Universität: Tübingen.

Dhakal, Dubi Nanda and List, Johann-Mattis and Roberts, Seán G. (2024): A phylogenetic study of South-Western Tibetic. Journal of Language Evolution 9.2. 14-28. https://doi.org/10.1093/jole/lzae008

Forkel, Robert and List, Johann-Mattis and Greenhill, Simon J. and Rzymski, Christoph and Bank, Sebastian and Cysouw, Michael and Hammarström, Harald and Haspelmath, Martin and Kaiping, Gereon A. and Gray, Russell D. (2018): Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. Scientific Data 5.180205. 1-10. https://doi.org/10.1038/sdata.2018.205

Hammarström, Harald and Haspelmath, Martin and Forkel, Robert and Bank, Sebastian (2024): Glottolog [Dataset, Version 5.1]. Leipzig:Max Planck Institute for Evolutionary Anthropology. https://glottolog.org

Gray, Russell D. and Atkinson, Quentin D. (2003): Language-tree divergence times support the Anatolian theory of Indo-European origin. Nature 426.6965. 435-439.

Häuser, Luise, Gerhard Jäger, Taraka Rama, Johann-Mattis List, and Alexandros Stamatakis (2024): Are sounds sound for phylogenetic reconstruction? In: Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP. Association for Computational Linguistics. 78-87. https://aclanthology.org/2024.sigtyp-1.11

Häuser, Luise and List, Johann-Mattis (2025): Lexibench. An improved collection of benchmark data for computational historical linguistics [Dataset, Version 0.1]. Passau: MCL Chair at the University of Passau. https://doi.org/10.5281/zenodo.14916463

Jäger, Gerhard, Johann-Mattis List, and Pavel Sofroniev (2017): Using support vector machines and state-of-the-art algorithms for phonetic alignment to identify cognates in multi-lingual wordlists. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Long Papers. 1204-1215. https://aclweb.org/anthology/E/E17/E17-1113

Kilani, Marwan (2020): FAAL: a Feature-based Aligning ALgorithm. Language Dynamics and Change 11.1. 30-76. https://doi.org/10.1163/22105832-01001300

Kondrak, Grzegorz (2000): A new algorithm for the alignment of phonetic sequences. In: Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference. 288-295.

List, Johann-Mattis (2014): Sequence comparison in historical linguistics. Düsseldorf: Düsseldorf University Press. https://sequencecomparison.github.io

List, Johann-Mattis (2012a): SCA: Phonetic alignment based on sound classes. In: Slavkovik, Marija and Lassiter, Dan (eds.): New directions in logic, language, and computation. Berlin and Heidelberg:Springer. 32-51. https://doi.org/10.1007/978-3-642-31467-4_3

List, Johann-Mattis (2012b): LexStat. Automatic detection of cognates in multilingual wordlists. In: Proceedings of the EACL 2012 Joint Workshop of Visualization of Linguistic Patterns and Uncovering Language History from Multilingual Resources. 117-125. https://aclanthology.org/W12-0216/

List, Johann-Mattis and van Dam, Kellen (2024): Computer-Assisted Language Comparison with EDICTOR 3 [Invited Paper]. Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change. https://aclanthology.org/2024.lchange-1.1

List, Johann-Mattis and Forkel, Robert (2023): LingPy. A Python library for quantitative tasks in historical linguistics [Software Library, Version 2.6.13]. Passau: MCL Chair at the University of Passau. https://pypi.org/project/lingpy

List, Johann-Mattis and Prokić, Jelena (2014): A benchmark database of phonetic alignments in historical linguistics and dialectology. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation. 288-294. https://aclanthology.org/L16-1379

List, Johann-Mattis and Tjuka, Annika and Blum, Frederic and Kučerová, Alžběta and Barrientos Ugarte, Carlos and Rzymski, Christoph and Greenhill, Simon J. and Robert Forkel (2025a): CLLD Concepticon [Dataset, Version 3.3.0]. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://concepticon.clld.org

List, Johann-Mattis, van Dam, Kellen Parker, and Frederic Blum (2025b): EDICTOR 3. An Interactive Tool for Computer-Assisted Language Comparison [Software Tool, Version 3.0]. Passau: MCL Chair at the University of Passau. https://edictor.org

List, Johann-Mattis and Anderson, Cormac and Tresoldi, Tiago and Forkel, Robert (2024): Cross-Linguistic Transcription Systems [Dataset, Version 2.3]. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://clts.clld.org

List, Johann-Mattis and Forkel, Robert (2022): Automated identification of borrowings in multilingual wordlists [version 3; peer review: 4 approved]. Open Research Europe 1.79. 1-11. https://doi.org/10.12688/openreseurope.13843.3

List, Johann-Mattis and Forkel, Robert and Greenhill, Simon J. and Rzymski, Christoph and Englisch, Johannes and Gray, Russell D. (2022a): Lexibank, A public repository of standardized wordlists with computed phonological and lexical features. Scientific Data 9.316. 1-31. https://doi.org/10.1038/s41597-022-01432-0

List, Johann-Mattis and Hill, Nathan W. and Forkel, Robert (2022b): A new framework for fast automated phonological reconstruction using trimmed alignments and sound correspondence patterns. In: Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change. Association for Computational Linguistics 89-96. https://aclanthology.org/2022.lchange-1.9

List, Johann-Mattis and Vylomova, Ekatarina and Forkel, Robert and Hill, Nathan and Cotterell, Ryan D. (2022c): The SIGTYP shared task on the prediction of cognate reflexes. In: Proceedings of the 4th Workshop on Computational Typology and Multilingual NLP. Association for Computational Linguistics 52-62. https://aclanthology.org/2022.sigtyp-1.7

List, Johann-Mattis and Greenhill, Simon J. and Anderson, Cormac and Mayer, Thomas and Tresoldi, Tiago and Forkel, Robert (2018): CLICS². An improved database of cross-linguistic colexifications assembling lexical data with help of cross-linguistic data formats. Linguistic Typology 22.2. 277-306. https://doi.org/10.1515/lingty-2018-0010

List, Johann-Mattis and Greenhill, Simon J. and Gray, Russell D. (2017): The potential of automatic word comparison for historical linguistics. PLOS ONE 12.1. 1-18. https://doi.org/10.1371/journal.pone.0170046

Meillet, Antoine (1954): La méthode comparative en linguistique historique [The comparative method in historical linguistics]. Paris: Honoré Champion.

Miller, John E. and List, Johann-Mattis (2023): Detecting lexical borrowings from dominant languages in multilingual wordlists. In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. Short Papers. Association of Computational Linguistics 2591-2597. https://aclanthology.org/2023.eacl-main.190

Prokić, Jelena and Wieling, Martijn and Nerbonne, John (2009): Multiple sequence alignments in linguistics. In: Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education. 18-25.

Rama, Taraka and List, Johann-Mattis and Wahle, Johannes and Jäger, Gerhard (2018): Are automatic methods for cognate detection good enough for phylogenetic reconstruction in historical linguistics? In: Proceedings of the North American Chapter of the Association of Computational Linguistics. 393-400. https://aclweb.org/anthology/N18-2063

Sagart, Laurent and Jacques, Guillaume and Lai, Yunfan and Ryder, Robin and Thouzeau, Valentin and Greenhill, Simon J. and List, Johann-Mattis (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. https://doi.org/10.1073/pnas.1817972116

Wu, Mei-Shin and Schweikhard, Nathanael E. and Bodt, Timotheus A. and Hill, Nathan W. and List, Johann-Mattis (2020): Computer-Assisted Language Comparison. State of the Art. Journal of Open Humanities Data 6.2. 1-14. https://doi.org/10.5334/johd.12

Supplementary Material
Data and code are curated at CodeBerg (https://codeberg.org/lexibank/lexibench, Version 0.1) and archived with Zenodo (https://doi.org/10.5281/zenodo.14916463).

Funding Information
This project has received funding from the European Research Council (ERC) under the European Union's Horizon Europe research and innovation programme (Grant agreement No. 101044282). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.