



A Global Lexical Database (GLED) for Computational Historical Linguistics

DATA PAPER

TIAGO TRESOLDI 

 ubiquity press

ABSTRACT

This work presents a lexical database with cognate annotation and phonological alignment for over 6,500 documented language varieties. The database includes per-family and global phylogenetic resources and offers a pre-computed global tree for language variety distance from normalized trees obtained with Bayesian Markov Chain Monte Carlo (MCMC) inference. Lexical data is provided in a single tabular file for convenience of usage, and resources are built adhering to best practices and state-of-the-art algorithms for historical linguistics. The database is a convenient source for research prototypes, method development, and analysis bootstrap. All resources are freely available for download for all interested researchers.

CORRESPONDING AUTHOR:

Tiago Tresoldi

Department of Linguistics and
Philology, Uppsala University,
Uppsala, Sweden

tiago.tresoldi@lingfil.uu.se

KEYWORDS:

lexical dataset; cognate
coding; phonetic alignment;
comparative method; linguistic
phylogenetics; computational
historical linguistics

TO CITE THIS ARTICLE:

Tresoldi, T. (2023). A Global
Lexical Database (GLED) for
Computational Historical
Linguistics. *Journal of Open
Humanities Data*, 9: 2, pp. 1–7.
DOI: [https://doi.org/10.5334/
johd.96](https://doi.org/10.5334/johd.96)

(1) OVERVIEW

REPOSITORY LOCATION

<https://doi.org/10.5281/zenodo.7368116>

CONTEXT

The Global Lexical Database (GLED) is a resource for computational historical linguistics encompassing a dataset of basic vocabulary for most known natural languages, with accompanying information on machine-detected cognates and phonological alignments, along with per-family and global phylogenetic resources. The latest release holds 262,859 entries for 6,572 doculects (documented language varieties, see Nordhoff & Hammarström, 2011) in 344 families (Figure 1) and is available under the CC-BY licence. The database's key component, a lexical dataset ultimately derived from the word lists of the Automated Similarity Judgement Program (ASJP), carries lemmas for between 30 and 40 comparative concepts for each doculect, all rendered with a broad phonetic transcription. The average concept coverage per doculect is 90.3%, and the average mutual pairwise coverage between doculects is 82.2%. Table 1 details the distribution of concept counts across doculects, and Table 2 lists the concepts along with their coverage.

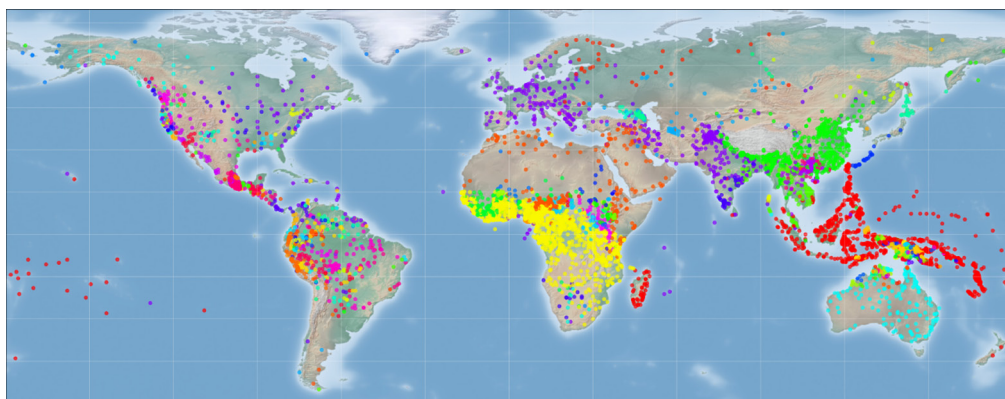


Figure 1 Location of the doculects included in the dataset, using information from Hammarström et al. (2022); colours are automatically assigned to differentiate language families.

NUMBER OF CONCEPTS	DOCULECTS	PERCENTAGE OF DOCULECTS
30	330	5.0
31	306	4.7
32	361	5.5
33	401	6.1
34	595	9.1
35	627	9.5
36	786	12.0
37	605	9.2
38	627	9.5
39	736	11.2
40	1198	18.2

Table 1 Number of doculects per number of concepts expressed in absolute and relative terms. Note that the number of entries for a doculect will be higher than the number of concepts in the case of synonyms.

The collection is not as accurate as alternative global (e.g., List et al., 2022a) and family or areal resources (e.g., Matisoff, 2008), which merge different sources, offer more significant concept coverages, and are manually curated for linguistic and data qualities. Such alternatives should be favoured when they encompass all the languages an investigation needs. Nonetheless, GLED constitutes a reliable and convenient source for probing language relationships, prototyping studies, and bootstrapping phylolinguistic analyses (Greenhill et al., 2020). It is likewise designed to support the development of new methods for tasks in computational historical linguistics, including phonological alignment, cognate detection, and sound correspondence inference (List et al., 2018). Finally, the language distances built in the database can be used for adjusted language sampling, as illustrated in Section 4.

CONCEPT GLOSS	DOCULECTS (RATIO)	CONCEPTICON NAME / ID
1pl	5265 (0.801)	WE / 1212
1sg	5379 (0.818)	I / 1209
2sg	5231 (0.795)	THOU / 1215
blood	6426 (0.977)	BLOOD / 946
bone	6351 (0.966)	BONE / 1394
breast	5957 (0.906)	BREAST / 1402
come	6130 (0.932)	COME / 1446
die	6125 (0.931)	DIE / 1494
dog	6430 (0.978)	DOG / 2009
drink	6058 (0.921)	DRINK / 1401
ear	6475 (0.985)	EAR / 1247
eye	6494 (0.988)	EYE / 1248
fire	6417 (0.976)	FIRE / 221
fish	6226 (0.947)	FISH / 227
full	4190 (0.637)	FULL / 1429
hand	5693 (0.866)	HAND / 1277
hear	5898 (0.897)	HEAR / 1408
horn	4317 (0.656)	HORN (ANATOMY) / 1393
knee	5357 (0.815)	KNEE / 1371
leaf	6077 (0.924)	LEAF / 628
liver	5454 (0.829)	LIVER / 1224
louse	5711 (0.868)	LOUSE / 1392
mountain	5321 (0.809)	MOUNTAIN / 639
name	6042 (0.919)	NAME / 1405
new	5711 (0.868)	NEW / 1231
night	6289 (0.956)	NIGHT / 1233
nose	6404 (0.974)	NOSE / 1221
one	6296 (0.958)	ONE / 1493
path	6151 (0.935)	PATH / 2252
person	5552 (0.844)	PERSON / 683
see	6104 (0.928)	SEE / 1409
skin	6182 (0.940)	SKIN / 763
star	6220 (0.946)	STAR / 1430
stone	6290 (0.957)	STONE / 857
sun	5877 (0.894)	SUN / 1343
tongue	6430 (0.978)	TONGUE / 1205
tooth	6399 (0.973)	TOOTH / 1380
tree	5850 (0.890)	TREE / 906
two	6285 (0.956)	TWO / 1498
water	6413 (0.975)	WATER / 948

Table 2 Absolute and relative doculect coverage per concept, along with the Concepticon mapping for each concept.

(2) METHOD

The dataset provided by Jäger (2018), derived from ASJP (Brown et al., 2008), was used as the lexical source, excluding doculects that did not fit the design (such as artificial languages, reconstructions, and duplicates). The original transcription system, “ASJPcode”, was mapped to a broad transcription consistent with CLTS/BIPA (Anderson et al., 2018) through an orthographic

profile (Moran & Cysouw, 2018). Such a profile was based on the one produced by the author for including ASJP in the Lexibank project. Decisions followed the non-exhaustive examples of phonological mapping and tokenization given in the original ASJP paper and the phonemic transcriptions of the ASJP word lists provided by other datasets.

Per-family automatic cognate attribution was performed with LexStat (List, 2012) for small and medium families (i.e., less than 18,000 items) and the SVM technique (Jäger, 2018) for large ones. Phonological alignments of the ensuing cognate sets were compiled with LingPy (List & Forkel, 2021). Finally, the data was organized in a singular tabular resource; entries were sorted, in order, by family, concept, language, and form (Table 3).

LANGUAGE	CODE	FAMILY	CONCEPT	FORM	ALIGNMENT	COGSET
Aché	ache1246	Tupian	DOG	bɛgi	b e g i	16
Amundava	amun1246	Tupian	DOG	ɲɛɲwɛɾɛ	ɲ e ɲ w - ɛ r ɛ	17
Avá Canoeiro	avac1239	Tupian	DOG	ʃɛwɛɾɛ	j e - w - ɛ r ɛ	17
Paraguayan Guarani	para1311	Tupian	DOG	ɖʒɛgwɛ	ɖ ʒ e g w - ɛ - -	17
Kaiwá	kaiw1246	Tupian	DOG	ʃɛgwɛ	j e g w - ɛ - -	17
Eastern Bolivian Guarani	east2555	Tupian	DOG	jeimbɛ	j e - i m b ɛ	19
Tapieté	tapi1253	Tupian	DOG	ɲɛʔambɛ	ɲ e ʔ ə m b ɛ	19
Cinta Larga	cint1239	Tupian	DOG	ɛwəli	e w ə l i	20
Gavião Do Jiparaná	gavi1246	Tupian	DOG	ɛvələ	e v ə l ə	20

Table 3 A modified snippet from the lexical dataset, showing the most critical columns for a subset of Tupian words for the concept “dog”. The data includes a unique language name, a Glottocode (when available), the family name, a concept gloss derived from the Concepticon catalog, the phonological transcription of the word, the phonological alignment of the word in its cognate set (with hyphens indicating gaps), and a cognate set index.

Per-family distance matrices based on the proportion of shared cognates were obtained from this dataset (Figure 2), and unrooted trees were constructed with the Neighbor-Joining method (Saitou & Nei, 1987). Models for inferring phylogenetic trees were produced with a patched version of BEASTling (Maurits et al., 2017) and monophyletically constrained using Glottolog 4.6 (Hammarström et al., 2022). Bayesian MCMC analyses were carried out with BEAST2 (Bouckaert et al., 2019), and summary Maximum Clade Credibility (MCC) trees were obtained with TreeAnnotator (Heled & Bouckaert, 2013). Finally, custom scripts were employed to normalize distances and join these trees, along with the language isolates, into a single unrooted tree (Figure 3). It must be underlined that the latter is in absolutely no manner proposed as supporting “Proto-Human” hypotheses but merely as a convenient resource for measuring language distance.

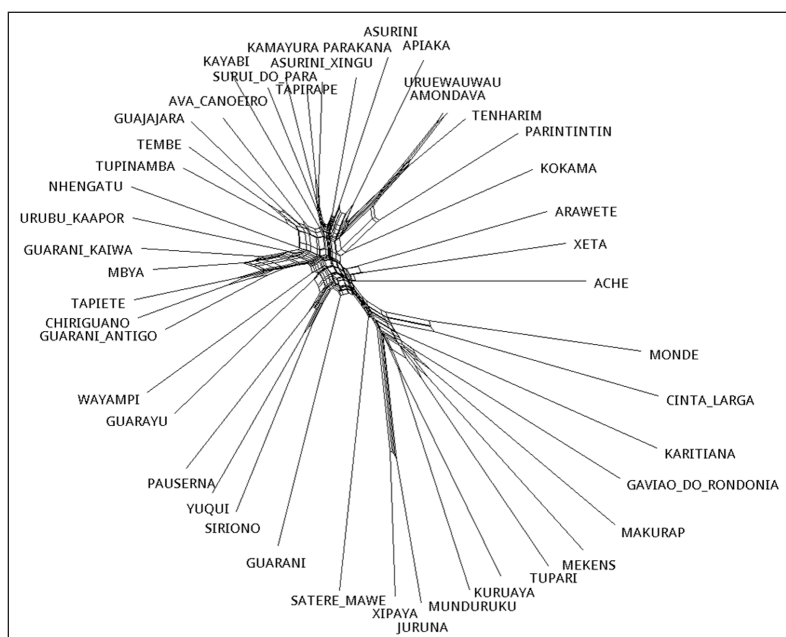


Figure 2 A neighbour-net for the Tupian languages in the dataset, plotted with SplitsTree v4 (Huson & Bryant, 2006).

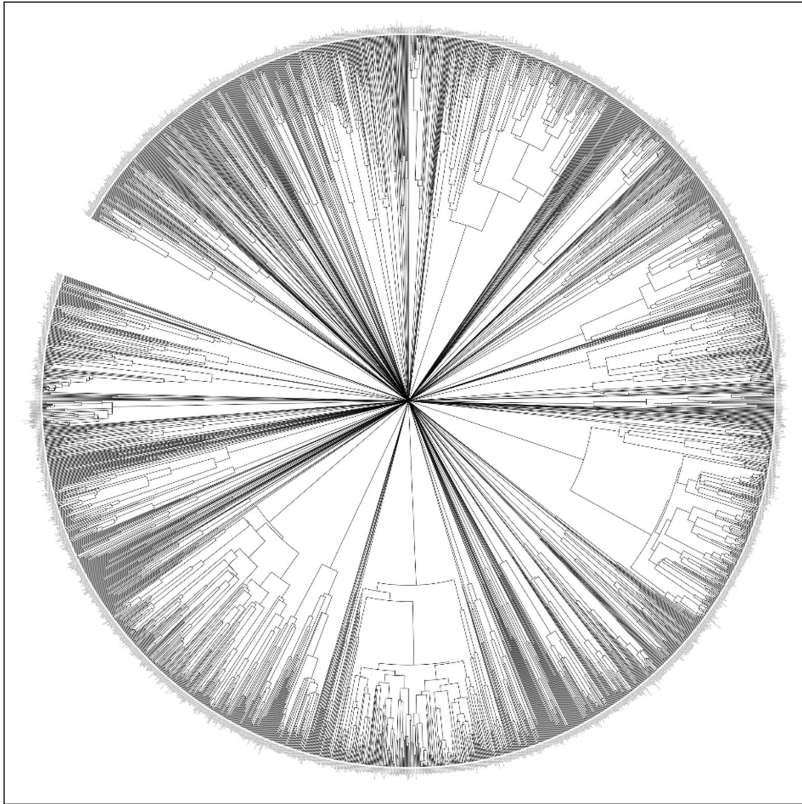


Figure 3 The “global” language tree from the combined Bayesian MCMC phylogenetic inferences, plotted with iTOL (Letunic & Bork, 2021).

The complete pipeline is accessible via the public GitHub repository at <https://github.com/tresoldi/gled> and takes approximately three days to be processed in a typical laptop (i5 processor, 8GB RAM, Fedora Linux 37). It will expedite planned forthcoming releases aggregating sources for languages missing in ASJP, such as recently documented isolates, and employing alternative methods for computational tasks, such as new methods of cognate detection.

(3) DATASET DESCRIPTION

OBJECT NAME

gled

FORMAT NAMES AND VERSIONS

The dataset has the following components:

- A TSV file (“gled.tsv”) with columns for (a) unique entry ID, (b) language ID (as provided in ASJP), (c) language name (provided by Glottolog, ASJP, or the author), (d) Glottocode when available, (e) Glottolog name when available, (f) family name, (g) concept gloss, (h) Concepticon ID (List et al., 2022b), (i) ASJP original form, (j) reconstructed form, (k) broad IPA transcription, (l) alignment, (m) cognate set ID, and (n) cognate set ID as an integer
- A YAML file (“gled.resource.yaml”) with the metadata as per the FrictionlessData project
- NEXUS files (“nexus/*.nex”) for families with more than one language
- Distance Matrices (“phylo/*.dst”) for families with more than one language, based on the percentage of shared cognates
- NJ trees in Newick notation (“phylo/*.tree”) for families with more than one language, based on the corresponding distance matrix
- Bayesian MCMC per-family (“trees/*.tree”) and global (“trees/global.tree”) trees in Newick notation

LANGUAGE

English

PUBLICATION DATE

2022-11-27

(4) REUSE POTENTIAL

Provided that its limits in proportion and strictness, arising from ASJP and examined in Brown et al. (2008) and Jäger (2018), are considered, the dataset provides many opportunities for reuse in empirical historical linguistics focused on lexical and phonetic data. Furthermore, as the doculects are linked to Glottolog, it is viable to integrate the data with other global-level resources, such as the World Loanword Database (Haspelmath & Tadmor, 2009), the World Atlas of Language Structures (Haspelmath et al., 2005), and Phoible (Moran & McCloy, 2019).

The distance matrices and phylogenetic trees offer a convenient starting point for comparing the results of different and more advanced analyses, notably with under-studied and under-resourced language families for which no distance matrix or phylogenetic tree with branch lengths is available. Table 4 illustrates such distances, showing values from the trees inferred without (NJ) and with (B) a molecular clock. Such distances can be managed to perform weighted random sampling at global, family, and sub-family levels, addressing issues such as sample bias and autocorrelation in cross-linguistic analyses.

LANGUAGE (GLOTTOCODE)	NJ	B	NB
Norwegian Bokmål (norw1259)	0.21	0.11	0.02
Danish (dani1285)	0.24	0.02	0.01
Dutch (dutc1256)	0.41	1.40	0.35
English (stan1293)	0.42	1.40	0.35
Italian (ital1282)	0.84	1.60	0.40
Hindi (hind1269)	0.90	1.95	0.48
Hittite (hitt1242)	0.90	1.97	0.49
Basque (basq1248)	∞	4.00	1.00

Table 4 Distance between Swedish (swed1254) and other languages, as computed using the Neighbour Joining trees (NJ, from zero to infinite), the Bayesian trees (B, from zero to 4.0), and the normalized Bayesian trees (NB, from zero to 1.0).

FUNDING STATEMENT

The database was developed in the “Cultural Evolution of Texts” project, with funding from the Riksbankens Jubileumsfond (grant agreement ID: MXM19-1087:1).


COMPETING INTERESTS

The author has no competing interests to declare.

AUTHOR CONTRIBUTIONS

Tiago Tresoldi: conceptualization, data curation, methodology, project administration, software, visualization, writing – original draft, writing – review & editing.

AUTHOR AFFILIATION

Tiago Tresoldi  orcid.org/0000-0002-2863-1467
Department of Linguistics and Philology, Uppsala University, Uppsala, Sweden

- Anderson, C., Tresoldi, T., Chacon, T., Fehn, A. M., Walworth, M., Forkel, R., & List, J.-M. (2018). A cross-linguistic database of phonetic transcription systems. In *Yearbook of the Poznan Linguistic Meeting*, 4(1), 21–53. De Gruyter Open. DOI: <https://doi.org/10.2478/yplm-2018-0002>
- Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., ..., & Drummond, A. J. (2019). BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLOS Computational Biology*, 15(4), 1–28. DOI: <https://doi.org/10.1371/journal.pcbi.1006650>
- Brown, C. H., Holman, E. W., Wichmann, S., & Velupillai, V. (2008). Automated classification of the world's languages: a description of the method and preliminary results. *Language Typology and Universals*, 61(4), 285–308. DOI: <https://doi.org/10.1524/stuf.2008.0026>
- Greenhill, S. J., Heggarty, P., & Gray, R. D. (2020). Bayesian Phylolinguistics. In Janda, R. D., Joseph, B. D., & Vance, B. S. (eds.) *The Handbook of Historical Linguistics*, Volume II, 226–253. Wiley-Blackwell: New Jersey. DOI: <https://doi.org/10.1002/9781118732168.ch11>
- Hammarström, H., Forkel, R., Haspelmath, M., & Bank, S. (2022). *Glottolog 4.6*. Leipzig: Max Planck Institute for Evolutionary Anthropology. DOI: <https://doi.org/10.5281/zenodo.6578297>
- Haspelmath, M., Dryer, M. S., Gil, D., & Comrie, B. (2005). *The world atlas of language structures*. OUP Oxford.
- Haspelmath, M., & Tadmor, U. (2009). The loanword typology project and the world loanword database. *Loanwords in the world's languages: A comparative handbook*, 1–34. DOI: <https://doi.org/10.1515/9783110218442.1>
- Heled, J., & Bouckaert, R. R. (2013). Looking for trees in the forest: summary tree from posterior samples. *BMC Evolutionary Biology*, 13(1), 1–11. DOI: <https://doi.org/10.1186/1471-2148-13-221>
- Huson, D. H., & Bryant, D. (2006). Application of Phylogenetic Networks in Evolutionary Studies, *Molecular Biology and Evolution*, 23(2), 254–267. DOI: <https://doi.org/10.1093/molbev/msj030>
- Jäger, G. (2018). Global-scale phylogenetic linguistic inference from lexical resources. *Scientific Data*, 5(1), 1–16. DOI: <https://doi.org/10.1038/sdata.2018.189>
- Letunic I., & Bork P. (2021). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Research*, 49(W1), 293–296. DOI: <https://doi.org/10.1093/nar/gkab301>
- List, J.-M. (2012). LexStat: Automatic detection of cognates in multilingual wordlists. In Butt, M., Carpendale, S., Penn, G., Prokić, J., & Cysouw, M. (eds.), *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, 117–125.
- List, J.-M., & Forkel, R. (2021). *LingPy. A Python library for quantitative tasks in historical linguistics*. Leipzig: Max Planck Institute for Evolutionary Anthropology.
- List, J.-M., Forkel, R., Greenhill, S. J., Rzymiski, C., Englisch, J., & Gray, R. D. (2022a). Lexibank, A public repository of standardized wordlists with computed phonological and lexical features. *Scientific Data*, 9(1), 1–16. DOI: <https://doi.org/10.1038/s41597-022-01432-0>
- List, J.-M., Tjuka, A., Rzymiski, C., Greenhill, S., & Forkel, R. (2022b). *CLLD Concepticon 3.0.0*. Leipzig: Max Planck Institute for Evolutionary Anthropology. DOI: <https://doi.org/10.5281/zenodo.7298023>
- List, J.-M., Walworth, M., Greenhill, S. J., Tresoldi, T., & Forkel, R. (2018). Sequence comparison in computational historical linguistics. *Journal of Language Evolution*, 3(2), 130–144. DOI: <https://doi.org/10.1093/jole/lzy006>
- Matisoff, J. A. (2008). *The Tibeto-Burman reproductive system: Toward an etymological thesaurus*. University of California Press.
- Maurits, L., Forkel, R., Kaiping, G. A., & Atkinson, Q. D. (2017). BEASTling: A software tool for linguistic phylogenetics using BEAST 2. *PLoS One*, 12(8), e0180908. DOI: <https://doi.org/10.1371/journal.pone.0180908>
- Moran, S., & Cysouw, M. (2018). *The Unicode Cookbook for Linguists: Managing writing systems using orthography profiles*. Language Science Press.
- Moran, S., & McCloy, D. (eds.) (2019). *PHOIBLE 2.0*. Jena: Max Planck Institute for the Science of Human History.
- Nordhoff, S., & Hammarström, H. (2011). Glottolog/Langdoc: Defining dialects, languages, and language families as collections of resources. In Kauppinen, T., Pouchard, L. C., Kessler, C. (eds.), *Proceedings of the First International Workshop on Linked Science*. Vol. 783, 1–7. CEUR.
- Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4), 406–425. DOI: <https://doi.org/10.1093/oxfordjournals.molbev.a040454>

TO CITE THIS ARTICLE:

Tresoldi, T. (2023). A Global Lexical Database (GLED) for Computational Historical Linguistics. *Journal of Open Humanities Data*, 9: 2, pp. 1–7. DOI: <https://doi.org/10.5334/johd.96>

Published: 02 February 2022

COPYRIGHT:

© 2023 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.