A Global Lexical Dataset (GLED) with cognate annotation and phonological alignment

Release 20220127 Lemmas 275340 Languages 6447 Families 238 Cognatesets 91608 Tokens 2056987

DOI 10.5281/zenodo.5911154

This repository comprises a dataset developed from a subset of ASJP, in which all lemmas are presented in a broad phonological transcription, automatically annotated for cognacy, and phonologically aligned. Per-family NEXUS files with binary annotation of presence/absence of cognate sets are also available. The dataset is intended to facilitate prototyping studies and methods in quantitative historical linguistics.

Statistics

The 20220127 release comprises:

- Entries: 275340
- Doculects: 6447
- Families: 238
- Cognate sets: 91608
- Tokens: 2056987
- Mean cognate set size: 3.01

Contents

The dataset is offered as a single textual tabular file, supported by Frictionless metadata, to simplify its usage. It is released with the full pipeline for processing, allowing to replicate the data and generate future versions. A complementary version following the CLDF standard (Forkel et al. 2018) will be available on the next releases.

The main file released by this project is `data/gled.{releasedate}.tsv`. This tabular source is accompanied by a dataset schema description following the Frictionless standard in `gled.yaml`, but the latter is not necessary if you open the main file as a tabular source within a programming language or a spreadsheet program. In most environment for analysis and development, it should be enough to read the data as a tabular (`"CSV"`) file, specifying tabulations (`"\t"`) as delimiters. The encoding is UTF-8.

Field names are all in uppercase strict ASCII. The file is sorted in ascending order following the value of fields `FAMILY`, `COGSET`, and `ID`. Concepts are linked to the Concepticon (List et al. 2021) reference catalog for comparative concepts, and language varieties are linked to Glottolog (Hammarström et al. 2021)

Field name	**Type**	**Description**
ID	String	A unique identifier for the lemma, as used in ASJP. E.g.: `ADYGHE-34-2`, `KABARDIAN-34-1`.
DOCULECT	String	The name of the doculect (“language”), in uppercase. E.g.: `ADYGHE`, `KABARDIAN`.
DOCULECT_DATE	Integer	The year associated with the doculect; empty fields should be assumed as living languages. E.g.: `1992`.
GLOTTOCODE	String	The languoid associated to the doculect in the Glottolog catalog. E.g.: `adyg1241`, `kaba1278`.
GLOTTOLOG_NAME	String	The language name associated with the `GLOTTOCODE` languoid in the Glottolog catalog. Please note that the mapping between ASJP and Glottolog doculects is not guaranteed to be bijective. E.g.: `Adyghe`, `Kabardian`.
FAMILY	String	The language family for the `DOCULECT`, as specified in ASJP (note that Glottolog’s classification might disagree). E.g.: `Abkhaz-Adyge`, `Dravidian`.
CONCEPT	String	The normalized gloss for the lemma’s concept, as specified in the Concepticon project. E.g.: `HORN (ANATOMY)`, `KNEE`.
IPA	String	A sequence of normalized CLTS BIPA graphemes (i.e., phonemes), separated by spaces. E.g.: `b ʒ ə`, `b ʒ ɐ q’ʷ ə`.
ALIGNMENT	String	A sequence of BIPA graphemes and dashes (representing gaps) expressing the lemma’s alignment in its cognate set. E.g.: `b ʒ ə - -`, `b ʒ ɐ q’ʷ ə`.
COGSET	String	A label identifying the cognate set to which the lemma belongs, also carrying information on linguistic family and concept. All in lowercase, with the cognate set index expressed by trailing digits. E.g.: `abkhaz_adyge_horn_47`, `abkhaz_adyge_knee_49`.

The `nexus/` directory carries individual per-family NEXUS files encoding the presence or absence of each applicable cognate set. All the characters have an ascertainment correction.

The software pipeline for downloading, processing, and releasing new versions of the dataset is available in the `pipeline/` directory. Please note that, due to the processing time necessary for the core step of automatic cognate detection, the entire process can take days on a normal desktop or laptop computer.

Background

The Automated Similarity Judgement Program (ASJP) was a collaborative project in quantitative comparative linguistics concerning the collection and transcription of lexical data for most languages of the world. The database supporting the project, popularly known as the “ASJP database” or just “ASJP”, is a set of basic vocabulary items, mostly comprising 40 comparative concepts, for over half of the world’s languages. Lexemes are transcribed with a custom orthography called “ASJPcode”, providing what we can regard as a broad phonological transcription.

The database was first used to estimate dates of language evolution, with a method comparable to glottochronology but accounting for lexical and phonological similarities as measured with an edit distance metric. Despite its known limitations, because of the massive and varied volume of data, the database has since been employed for examining other matters, such as phonological diversity (Wichmann, Rama, and Holman 2011) and sound-meaning associations (Blasi et al. 2016).

The adoption of ASJPcode, followed to ease and speed the transcription, places limits on its re-usage by linguists, and the global-scale level of data makes it impossible to produce a complete expert-annotation of cognacy for all lexemes. Automatic and computer-assisted approaches at cognate detection have been explored, such as in Jäger (2013) and Jäger (2018), but they keep the ASJPcode transcription and focus on identifying phylogenetic signal. The most popular method for automatic cognate detection, LexStat (List 2012), can be used with ASJPcode, but the standard implementation (List and Forkel 2021) cannot be used for an overall cognate detection because of the high number of data-points. In preliminary experiments, even with 256 Gb of RAM the methods were ultimately unable to process large families such Afro-Asiatic and Indo-European.

Aiming at providing a dataset that is easy to use and conforming to FAIR principles of data management (Wilkinson et al. 2016), I am releasing a dataset derived from ASJP, where are all lemmas are given in a broad IPA transcription and annotated for cognacy, with the resulting cognate sets phonologically aligned. I have also prepared the collection to remove spurious cognate sets and to obtain a more feasible volume of data-points, producing a single dataset that is suitable for testing hypotheses on language evolution. It allows to prototype studies and benchmark methods that can later apply to higher quality datasets, such as those provided by the Lexibank project (List et al. under review). Remember to consider all the limitations of this data before making any claims in terms of language evolution or relationship.

Methodology

I took the ASJP dataset (Wichmann et al. 2020a) as available (Wichmann et al. 2020b) via the Lexibank project (List et al. under review) in CLDF format (Forkel et al. 2018), and mapped the source ASJPcode (Brown at al. 2008) to a broad IPA transcription through orthographic profiles (Moran and Cysouw 2018) and CLTS (Anderson et al. 2018), which I had previously prepared. I removed from the dataset languages that did not fit the original design (e.g., artificial languages, reconstructions, isolates, duplicates, etc.; these might be included in future releases). I ran per-family cognate detection using a custom extension of Lexstat (List 2012), available in the released pipeline, which partitions the work into partially overlapping subsets and then joins the results with methods of community detection (Csárdi 2006). At last, I produced phonological alignments of the resulting cognate sets using LingPy (List and Forkel 2021) and organized the data into a tabular resource.

Known limitations

Despite the dataset providing an accessible entry-point for research in quantitative historical linguistics, it suffers from several limitations both inherited from ASJP and induced by the data manipulation. Limitations of the first kind have been amply discussed in the literature, including in the papers presenting the ASJP project, and don't need to be addressed here.

Limitations caused by the data manipulation fall into three categories, mirroring the major steps of processing. First, the reconstruction from ASJPcode to IPA, despite aiming for rather broad transcriptions, is most times only approximate, as by design a single orthographic profile was employed. While entries were highlighted for review using internal tools and occasionally hard-coded in their transcription, in many situations these are less precise than what a manual review would achieve. Second, the lower quality of the phonological data increased the error rate of automatic cognate detection, already impaired in the case of large language families by a supplementary round of result aggregation and community detection. Manual inspection of several concepts and families highlighted noticeable inaccuracies, which were not amended to preserve reproducibility, to avert any human bias, and to ensure a global comparability. Third, automatic alignment is likewise subject to errors, even more when an alignment includes lemmas which don't appear to fit to their cognate set.

Changelog

Release 20220127: - First public release.

Community guidelines

While the author can be contacted directly for support, it is recommended that third parties use GitHub standard features, such as issues and pull requests, to contribute, report problems, or seek support.

Contributing guidelines, including a code of conduct, can be found in the `CONTRIBUTING.md` file.

License

As the original ASJP data, this dataset is released under the terms of the [Creative Commons Attribution 4.0 International (CC BY 4.0)] license. You are free to share and adapt the data, as long as you give appropriate credit, provide a link to the license, indicate if changes were made, and don't establish additional restrictions on the derivative work.

Author and citation

The dataset is developed by Tiago Tresoldi (tiago.tresoldi@lingfil.uu.se). During its development, the author received funding from Cultural Evolution of Texts project (Riksbankens Jubileumsfond, grant agreement ID: MXM19-1087:1).

If you use this dataset, please cite it as:

> Tresoldi T. 2022. A Global Lexical Dataset (GLED) with cognate annotation and phonological alignments. [Data set]. Zenodo. doi: 10.5281/zenodo.5911132

In BibTeX:

```
@misc{Tresoldi2022gled,
  year      = {2022},
  author    = {Tiago Tresoldi},
  title     = {A Global Lexical Dataset (GLED) with cognate annotation and
phonological alignments},
  publisher = {Zenodo},
  doi       = {10.5281/zenodo.5911132}
}
```

References

Anderson C, Tresoldi T, Chacon TC, Fehn AM, Walworth M, Forkel R, and List JM. 2018. A Cross-Linguistic Database of Phonetic Transcription Systems. *Yearbook of the Poznań Linguistic Meeting 4.1*. 21-53. doi: 10.2478/yplm-2018-0002

Blasi DE, Wichmann S, Hammarström H, Stadler PF, and Christiansen MH. 2016. “Sound–meaning association biases evidenced across thousands of languages.” *P Natl Acad Sci USA* 113.39: 10818-10823. doi: 10.1073/pnas.1605782113

Brown CH, Holman EW, Wichmann S, and Velupillai V. 2008. Automated classification of the world’s languages: a description of the method and preliminary results. *Language Typology and Universals*, vol. 61, no. 4, pp. 285-308. doi: 10.1524/stuf.2008.0026

Csárdi G, and Nepusz T. 2006. The igraph software package for complex network research. *InterJournal Complex Systems*, 1695. doi: 10.5281/zenodo.3630268

Forkel R, List JM, Greenhill SJ, Rzymski C, Bank S, Cysouw M, Hammarström H, Haspelmath M, Kaiping GA, and Gray RD. 2018. Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Sci Data* 5, 180205. doi: 10.1038/sdata.2018.205

Hammarström H, Forkel R, Haspelmath M, and Bank S. 2021. *Glottolog 4.5*. Leipzig: Max Planck Institute for Evolutionary Anthropology. doi: 10.5281/zenodo.5772642

Jäger G. 2013. Phylogenetic Inference from Word Lists Using Weighted Alignment with Empirically Determined Weights. *Language Dynamics and Change*, 3(2), 245-291. doi: 10.1163/22105832-13030204

Jäger G. 2018. Global-scale phylogenetic linguistic inference from lexical resources. *Scientific Data* 5, 180189. doi: 10.1038/sdata.2018.189

List JM. 2012. “LexStat: Automatic detection of cognates in multilingual wordlists.” *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*. p. 117-125.

List JM, and Forkel R. 2021. *LingPy. A Python library for historical linguistics*. Version 2.6.9. doi: 10.5281/zenodo.597082

List JM, Rzymski C, Greenhill S,Schweikhard N,Pianykh K, Tjuka A, Hundt C, and Forkel R (eds.). 2021. CLLD Concepticon 2.5.0 [Data set]. *Zenodo*. doi: 10.5281/zenodo.4911605

List JM, Forkel R, Greenhill SJ, Rzymski C, Englisch J, and Gray RD. Forthcoming. *Lexibank: A public repository of standardized wordlists with computed phonological and lexical features*. doi: 10.21203/rs.3.rs-870835/v1

Moran S, and Cysouw M. 2018. *The Unicode cookbook for linguists: Managing writing systems using orthography profiles*. Translation and Multilingual Natural Language Processing 10. Berlin: Language Science Press. doi: 10.5281/zenodo.1296780

Wichmann S, Rama T, and Holman E. 2011. “Phonological diversity, word length, and population sizes across languages: The ASJP evidence”. *Linguistic Typology*, vol. 15, no. 2, pp. 177-197. doi: 10.1515/lity.2011.013

Wichmann S, Holman EW, and Brown CH (eds.). 2020. *The ASJP Database*. Version 19. Available at: https://asjp.clld.org/

Wichmann S, Holman EW, Brown CH, Forkel R, and Tresoldi T. 2020. CLDF dataset derived from Wichmann et al.’s “ASJP Database” v19 from 2020. (v19.1) [Data set]. *Zenodo*. doi: 10.5281/zenodo.3843469

Wilkinson M, Dumontier M, Aalbersberg I et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018. doi: 10.1038/sdata.2016.18