The Journal of Open Source Software

Lexedata: A toolbox to edit CLDF lexical datasets

Gereon A. Kaiping [1], **Melvin S. Steiger** [2], **and Natalia Chousou-Polydouri** [23]

1 Department of Geography, Universität Zürich, CH **2** Department of Comparative Linguistics, Universität Zürich, CH **3** Center for the Interdisciplinary Study of Language Evolution, Universität Zürich, CH

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor:

Reviewers:

- @xrotwang
- @peterdekker

Submitted: 01 February 2022
Published: unpublished

Summary

Lexedata is a collection of tools to support the editing process of comparative lexical data. Wordlists are a comparatively easily collected type of language documentation that is nonetheless quite data-rich and useful for the systematic comparison of languages (List et al., 2021). They are an important resource in comparative and historical linguistics, including their use as raw data for language phylogenetics (Gray et al., 2009; Grollemund et al., 2015).

The `lexedata` package uses the "Cross-Linguistic Data Format" (CLDF, Forkel et al. (2021), Forkel et al. (2018)) as main data format for a relational database containing forms, languages, concepts, and etymological relationships. The CLDF specification builds on top of the CSV for the Web (CSVW, Pollock et al. (2015)) specs by the W3C, and as such consists of one or more comma-separated value (CSV) files that get their semantics from a metadata file in JSON format.

Implemented in Python as a set of command line tools, Lexedata provides various helper functions to address issues that frequently arise when working with comparative wordlists for multiple languages, as shown in Figure 1. These include import and export functions to formats more familiar to linguists, and integrity checks for bulk edit and import functionality it provides. For example, there are scripts for importing data from MS Excel sheets of various common formats into CLDF, checking for homophones, manipulating etymological judgements, or exporting coded datasets for use in phylogenetic software.

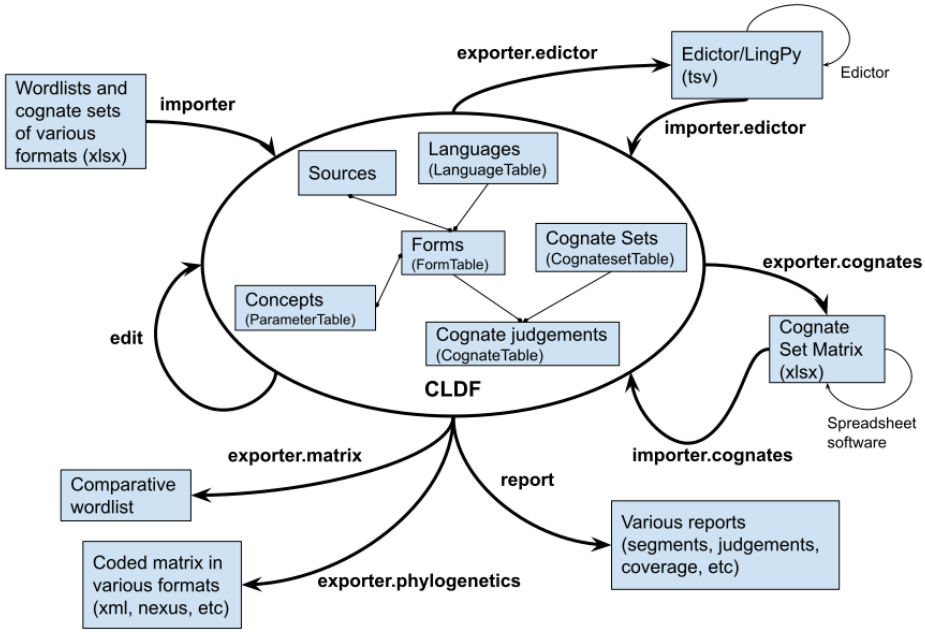

Figure 1: Overview over the functionality in Lexedata.

Statement of Need

Maintaining the integrity of CLDF as a relational database is difficult using general CSV editing tools. This holds in particular for the usual dataset size of hundreds of languages and concepts, and formats unfamiliar to most linguists. Dedicated relational database software, which simplifies the maintenance of the data constraints, would set an even bigger hurdle to researchers, even to those who are reasonably computer-savvy.

The major existing tool for curating lexical datasets in other formats and providing them as CLDF for interoperability is cldfbench (Forkel & List, 2020). However, cldfbench assumes that the data curator is not necessarily in a position to edit the dataset. As such, it provides a very flexible interface to transform and curate CLDF datasets, at the cost of making this accessible through an API which requires writing Python code.

Given that the majority of comparative linguists are unfamiliar with programming, Lexedata is designed to not need any programming skills. In contrast with cldfbench, Lexedata is written for the purpose of not only curating, but also collecting and editing the dataset. It therefore imposes additional constraints on the dataset which are very useful in editing tasks, but not strictly required by CLDF.

There are two major existing tools for editing lexical datasets, LingPy (List & Forkel, 2021) and Edictor (List, 2017). Edictor is a browser-based graphical user interface tool to edit cognate annotations, while LingPy is a Python library focused on automating manipulations of lexical datasets, such as automatic cognate detection. Both of these pre-date the CLDF format, and while their common data format inspired some features of CLDF, it has some differences. Lexedata provides export and import functionality for this TSV-based format to and from CLDF. In addition, Lexedata exposes a major LingPy functionality, the Automatic Cognate Detection (ACD, List et al. (2017)) using Lexstat (List, 2012), to work directly on CLDF

datasets. This avoids both memory issues arising from LingPy's approach to load the entire dataset into memory and the need to convert between CLDF and LingPy.

Lexedata is designed to facilitate adding comments to cognate sets and cognate judgements, through the annotation tools in the Excel format (which naturally extend to comment threads in Google Sheets for collaborative editing), as well as tracking the editing workflow through status columns with customizable messages. Last but not least, to ensure that the user retains a good sense of control and overview, Lexedata includes helpful warning messages that suggest potential solutions and next steps to the user, while it keeps the user informed about batch operations with intermediate info messages and final reports.

In summary, Lexedata addresses the need to curate and edit a lexical dataset in CLDF format without the ability to program, which is still a rare skill among comparative linguists. It allows this without sacrificing the power and familiarity of existing software such as GUI spreadsheed apps or Edictor, by providing user-friendly access to format conversions and bulk editing functionality from simple terminal commands.

Research use

The extensive lexical dataset editing functionality is currently used by projects at UC Berkeley and Universität Zürich for Arawakan and Mawetí-Guaraní languages and at Universiteit Gent for Bantu. Precursor scripts have also been used for Timor-Alor-Pantar and Austronesian languages (Kaiping & Klamer, 2018). The export to phylogenetic alignments, derived from BEASTling (Maurits et al., 2018, 2017), has been used in different language phylogenetics projects that are already under review (Gunnink et al., accepted; Kaiping & Klamer, 2019; Kaiping & Neureiter, (under review)).

Acknowledgement

Development of Lexedata was funded by the Swiss National Science Foundation (SNSF) Sinergia Project "Out of Asia" CRSII5_183578.

References

Forkel, R., Cysouw, M., List, J.-M., Rzymski, C., Greenhill, S. J., & Moran, S. (2021). *Cldf/cldf: CLDF 1.1.2*. Zenodo. https://doi.org/10.5281/zenodo.4804030

Forkel, R., & List, J.-M. (2020). CLDFBench: Give your cross-linguistic data a lift. *LREC*.

Forkel, R., List, J.-M., Greenhill, S. J., Rzymski, C., Bank, S., Cysouw, M., Hammarström, H., Haspelmath, M., Kaiping, G. A., & Gray, R. D. (2018). Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data*, *5*, 180205. https://doi.org/10.1038/sdata.2018.205

Gray, R. D., Drummond, A. J., & Greenhill, S. J. (2009). Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science*, *323*(5913), 479–483. https://doi.org/10.1126/science.1166858

Grollemund, R., Branford, S., Bostoen, K., Meade, A., Venditti, C., & Pagel, M. (2015). Bantu expansion shows that habitat alters the route and pace of human dispersals. *Proceedings of the National Academy of Sciences*, *112*(43), 13296–13301. https://doi.org/10.1073/pnas.1503793112

Gunnink, H., Chousou-Polydouri, N., & Bostoen, K. (accepted). Divergence and contact in southern bantu language and population history: A new phylogeny in cross-disciplinary perspective. *Language Dynamics and Change*.

Kaiping, G. A., & Klamer, M. (2019). *Subgrouping the Timor-Alor-Pantar languages using systematic Bayesian inference*. https://doi.org/10.31235/osf.io/9s5hj

Kaiping, G. A., & Klamer, M. (2018). LexiRumah: An online lexical database of the Lesser Sunda Islands. *PLOS ONE*, *13*(10), e0205250. https://doi.org/10.1371/journal.pone.0205250

Kaiping, G. A., & Neureiter, N. ((under review)). Burst Clock. *Journal of Language Evolution*.

List, J.-M. (2017). A web-based interactive tool for creating, inspecting, editing, and publishing etymological datasets. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. System Demonstrations*, 9–12. http://edictor.digling.org

List, J.-M. (2012). LexStat: Automatic detection of cognates in multilingual wordlists. *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, 117–125. ISBN: 978-1-937284-19-0

List, J.-M., & Forkel, R. (2021). *LingPy. A Python Library for Historical Linguistics* (Version v2.6.8) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.5144474

List, J.-M., Forkel, R., Greenhill, S. J., Rzymski, C., Englisch, J., & Gray, R. D. (2021). *Lexibank: A public repository of standardized wordlists with computed phonological and lexical features*. https://doi.org/10.21203/rs.3.rs-870835/v1

List, J.-M., Greenhill, S. J., & Gray, R. D. (2017). The Potential of Automatic Word Comparison for Historical Linguistics. *PLOS ONE*, *12*(1), e0170046. https://doi.org/10.1371/journal.pone.0170046

Maurits, L., Forkel, R., & Kaiping, G. A. (2018). *BEASTling*.

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. https://doi.org/10.1371/journal.pone.0180908

Pollock, R., Tennison, J., Kellogg, G., & Herman, I. (2015). *Metadata vocabulary for tabular data* ({W3c} Recommendation REC-tabular-metadata-20151217). W3C. https://www.w3.org/TR/2015/REC-tabular-metadata-20151217/