The Journal of Open Source Software

glottospace: R package for language mapping and geospatial analysis of linguistic and cultural data

Sietze Norder [1,2], **Laura Becker** [3], **Hedvig Skirgård** [4], **Leonardo Arias** [2,5], **Alena Witzlack-Makarevich** [6], **and Rik van Gijn** [2]

1 Copernicus Institute of Sustainable Development, Utrecht University, Utrecht, The Netherlands **2** Leiden University Centre for Linguistics, Leiden University, Leiden, The Netherlands **3** Department of General Linguistics, University of Freiburg, Freiburg im Breisgau, Germany **4** Department of Linguistic and Cultural Evolution, Max-Planck-Institute for Evolutionary Anthropology, Leipzig, Germany **5** Department of Evolutionary Genetics, Max-Planck-Institute for Evolutionary Anthropology, Leipzig, Germany **6** Hebrew University of Jerusalem, Jerusalem, Israel

DOI: 10.21105/joss.04303

Software

- Review
- Repository
- Archive

Editor: Andrew Stewart

Reviewers:

- @fauxneticien
- @skalyan91

Submitted: 30 March 2022
Published: 12 September 2022

Summary

The glottospace R package facilitates the geospatial analysis of linguistic and cultural data. The aim of the package is to provide a streamlined workflow for geolinguistic analysis, including data entry, data import, cleaning, exploration, language mapping and visualisation, and export. Glottospace is also intended as an R interface to global linguistic and cultural databases such as Glottolog, WALS, and D-PLACE, contributing to improved reproducibility of data analysis.

Statement of need

Several databases exist that capture aspects of linguistic and cultural diversity globally. However, there is a lack of interfaces to access and manipulate these databases, specifically within the R environment (R Core Team, 2022). For example, it is not straightfoward to link these databases with data collected by researchers at smaller scales or particular sets of languages. While linguistic R packages have been developed for specific purposes (e.g., lingtypR (Becker, 2022), qlcData (Cysouw, 2018), lingtypology (Moroz, 2017), and glottoTrees (Round, 2021)), there is currently no easy-to-use package that automates the most common tasks related to analysing, visualising, and mapping (geo)linguistic data. **glottospace** aims to fill this gap by offering a set of functions that, in essence, provide four things:

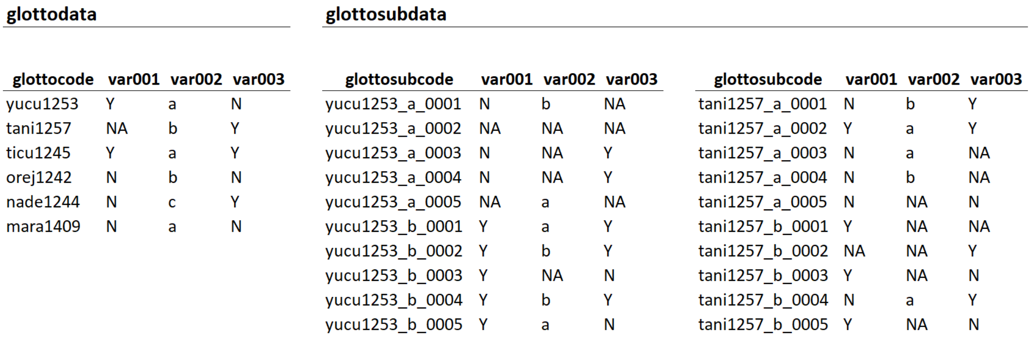
1. Simplified access to global linguistic and cultural databases
2. Standardized data structures for data collection, import, cleaning, and checking
3. Functionalities to merge data from various linguistic and cultural datasets
4. Streamlined workflows for geolinguistic data analysis and visualisation

We will now describe each of these elements in more detail.

Existing global databases of linguistic and cultural diversity, such as Glottolog (Hammarström et al., 2021), WALS (Dryer & Haspelmath, 2013), and D-PLACE (Kirby et al., 2016), are structured according to the cross-linguistic data format (Forkel et al., 2018), allowing for the integration of different databases. One way in which languages can be matched across databases is by using glottocodes, i.e., unique identifiers of languages, dialects and language families (Forkel & Hammarström, 2021). These glottocodes often have geospatial coordinates associated with them, allowing for geospatial analysis and visualisation. With **glottospace**, users can easily access the most recent version of these databases. Researchers can query those databases, use them as a benchmark, or supplement their own data with additional information (like geospatial coordinates, language family).

The **glottospace** package can generate empty data structures to facilitate data entry, or convert existing databases (for example stored in an Excel or CSV file) into two standardized data structures (Figure 1):

- glottodata: a single data table (and optionally, metadata tables)
  - one row for each glottocode
  - any number of columns with linguistic/cultural features
- glottosubdata: multiple data tables (and optionally, metadata tables)
  - one table for each glottocode
    - one row for each glottosubcode
    - any number of columns with linguistic/cultural features



glottodata

glottocode	var001	var002	var003
yucu1253	Y	a	N
tani1257	NA	b	Y
ticu1245	Y	a	Y
orej1242	N	b	N
nade1244	N	c	Y
mara1409	N	a	N

glottosubdata

glottosubcode	var001	var002	var003
yucu1253_a_0001	N	b	NA
yucu1253_a_0002	NA	NA	NA
yucu1253_a_0003	N	NA	Y
yucu1253_a_0004	N	NA	Y
yucu1253_a_0005	NA	a	NA
yucu1253_b_0001	Y	a	Y
yucu1253_b_0002	Y	b	Y
yucu1253_b_0003	Y	NA	N
yucu1253_b_0004	Y	b	Y
yucu1253_b_0005	Y	a	N

glottosubcode	var001	var002	var003
tani1257_a_0001	N	b	Y
tani1257_a_0002	Y	a	Y
tani1257_a_0003	N	a	NA
tani1257_a_0004	N	b	NA
tani1257_a_0005	N	NA	N
tani1257_b_0001	Y	NA	NA
tani1257_b_0002	NA	NA	Y
tani1257_b_0003	Y	NA	N
tani1257_b_0004	N	a	Y
tani1257_b_0005	Y	NA	N

Figure 1: Examples of glottodata (left) and glottosubdata (right) without metadata tables.

The glottodata structure is appropriate when one wants to assign one or more features to each language in the dataset (as, e.g., in WALS). The glottosubdata structure allows for assigning more complex structures (inventories) to each language in the dataset, which may vary in size from one language to another, such as phoneme inventories (Phoible; (Moran & McCloy, 2019) construction/morphological inventories (AUTOTYP; (Bickel et al., 2022), and subordination strategies in SAILS (Gijn, 2016).

Although all metadata tables that can be generated for glotto(sub)data are optional, some of these tables can greatly facilitate the automation and reproducibility of further analysis. Examples of metadata tables that can be generated by **glottospace** are:

- structure table: specify the type, possible levels, weight, group, and subgroup for each variable
- description table: provide a description, reference, and remark for each variable
- references table: provide a reference and page number for each glottocode (or variable)
- remarks table: provide additional remarks for each glottocode (or variable)
- contributors table: add the names of people who contributed the data
- sample: specify a subset of glottocodes that should be used in further analysis
- readme: add information about the person responsible for the data, including contact details and how to cite the data.

The glottodata and glottosubdata structures are optimized to be linked with the aforementioned global databases, and allow for streamlined workflows for the analysis and visualisation of linguistic and cultural data. For example, with one line of code, glottospace users can map a set of languages as either points or polygons, and colour them by a particular feature. However, language mapping and visualisation is just one aspect of the package's functionality, the aim of **glottospace** is to streamline entire workflows, facilitating common tasks such as:

- read user data from a local path
- convert data into glottodata or glottosubdata structures
- perform quality checks and data cleaning (e.g., missing values, inconsistencies, and undefined glottocodes)
- analyse languages and cultures based on relevant features (e.g., lexicon, phonemes, gender roles, and subsistence strategies)
- create different kinds of maps and visualisations for a set of languages
- export visualisations, maps, and datasets to be used in publications (Figure 2)
- improve reproducibility in data analysis

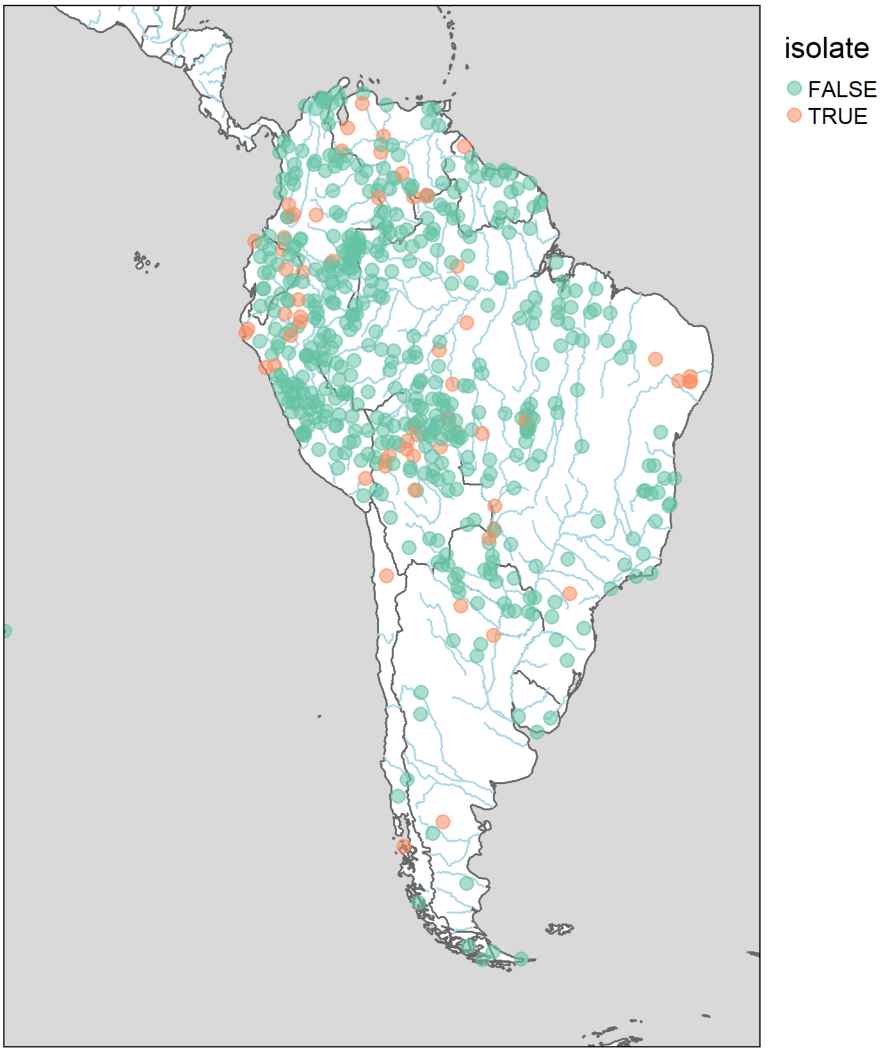


Figure 2: Isolate languages in South America, with major rivers in the background. This visualisation is generated with **glottospace** using one line of code. Although other map projections are supported, the default projection is the equal-area Eckert IV projection (following (McNew et al., 2019)).

To enable this functionality, **glottospace** builds on a combination of spatial and non-spatial packages, including sf (Pebesma, 2018), tmap (Tennekes, 2018), rnaturalearth (South, 2017), ggplot2 (Wickham, 2016), vegan (Oksanen et al., 2020), and dplyr (Wickham et al., 2021). The package is currently used by researchers and students in the field of comparative and areal linguistics as well as language typology and cultural anthropology. Furthermore, we are planning to use the package for classroom teaching.

Acknowledgements

Sietze Norder, Leonardo Arias, and Rik van Gijn were supported by the European Research Council under the EU H2020 and Research and Innovation program (SAPPHIRE grant 818854). The authors would like to thank Simon Greenhill, Robert Forkel and Harald Hammarström for brainstorming about linguistic data formats and the functionality of the package. Finally, the authors are very grateful to Nay San, Siva Kalyan, and Andrew Stewart for reviewing and editing.

References

Becker, L. (2022). *lingtypR: Easy data manipulation for typologists*. https://gitlab.com/laurabecker/lingtypr

Bickel, B., Nichols, J., Zakharko, T., Witzlack-Makarevich, Alena Hildebrandt, K., Reßler, M., Bierkandt, L., Zúñiga, F., & Lowe, J. B. (2022). *The AUTOTYP database*. https://doi.org/10.5281/zenodo.5931509

Cysouw, M. (2018). *qlcData: Processing data for quantitative language comparison (QLC)*.

Dryer, M., & Haspelmath, M. (2013). *Dryer, Matthew S. & Haspelmath, Martin (eds.) The World Atlas of Language Structures Online. Leipzig: Max Planck Institute for Evolutionary Anthropology.*

Forkel, R., & Hammarström, H. (2021). Glottocodes: Identifiers linking families, languages and dialects to comprehensive reference information. *Semantic Web*, *1*, 1–8. https://doi.org/10.3233/sw-212843

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*, 1–10. https://doi.org/10.1038/sdata.2018.205

Gijn, R. van. (2016). *Construction-based subordination data (SUB). In Muysken, Pieter et al. (eds.) South American Indian Language Structures (SAILS) Online. Jena: Max Planck Institute for the Science of Human History. (Available at https://sails.clld.org).*

Hammarström, H., Forkel, R., Haspelmath, M., & Bank, S. (2021). *Glottolog 4.5. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at https://glottolog.org).*

Kirby, K. R., Gray, R. D., Greenhill, S. J., Jordan, F. M., Gomes-Ng, S., Bibiko, H.-J., Blasi, D. E., Botero, C. A., Bowern, C., Ember, C. R., Leehr, D., Low, B. S., McCarter, J., Divale, W., & Gavin, M. C. (2016). D-PLACE: A global database of cultural, linguistic and environmental diversity. *PLoS ONE*, *11*(7), 1–14. https://doi.org/10.1371/journal.pone.0158391

McNew, G., Derungs, C., & Moran, S. (2019). Towards faithfully visualizing global linguistic diversity. *LREC 2018 - 11th International Conference on Language Resources and Evaluation*, 805–809. ISBN: 9791095546009

Moran, S., & McCloy, D. (2019). *PHOIBLE 2.0. Jena: Max Planck Institute for the Science of Human History*. https://phoible.org

Moroz, G. (2017). *Lingtypology: Easy mapping for linguistic typology*. https://CRAN.R-project.org/package=lingtypology

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E., & Wagner, H. (2020). *vegan: Community Ecology Package*. https://cran.r-project.org/package=vegan

Pebesma, E. (2018). Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, *10*(1), 439–446. https://doi.org/10.32614/RJ-2018-009

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://doi.org/10.1007/978-3-540-74686-7

Round, E. R. (2021). *glottoTrees: Phylogenetic trees in linguistics*. https://github.com/erichround/glottoTrees

South, A. (2017). *rnaturalearth: World Map Data from Natural Earth. R package version 0.1.0.* https://cran.r-project.org/package=rnaturalearth

Tennekes, M. (2018). tmap: Thematic Maps in R. *Journal of Statistical Software*, *84*(6), 1–39. 10.18637/jss.v084.i06

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wickham, H., Romain, F., Henry, L., & Müller, K. (2021). *dplyr: A grammar of data manipulation. R package version 1.0.7.*