Journal of Open Humanities Data

# Glottography: An Open-Source Geolinguistic Data Platform for Mapping the World's Languages

DISCUSSION PAPER

PETER RANACHER
ROBERT FORKEL
NOUR EFRAT-KOWALSKY
MATTHIAS URBAN
ANTONIA HEHLI
MICHA FRANZ
GREGORY BILAND
AARON KREIENBÜHL
ALBA HERMIDA RODRÍGUEZ
MATHEUS C. B. C. AZEVEDO
JAMES GIEBLER
TAKUYA TAKAHASHI
NICO NEUREITER
RIK VAN GIJN
MEELI ROOSE
OUTI VESAKOSKI
ROBERT WEIBEL
GEREON KAIPING
SIETZE NORDER

*Author affiliations can be found in the back matter of this article

ubiquity press

CORRESPONDING AUTHOR:
**Sietze Norder**
Copernicus Institute of Sustainable Development, Faculty of Geosciences, Utrecht University, Utrecht, The Netherlands
s.j.norder@uu.nl

## ABSTRACT

Maps depicting the geographic location of languages are essential tools for linguistic research. Although many language maps are available in the scientific literature, most encode spatial information as static images, often on paper. In contrast, geographic databases store languages as georeferenced digital data, allowing integration with other datasets, quantitative geographic analyses, and mapping. At present, there is no open-access platform providing digital language areas. To address this limitation, we introduce *Glottography*, a free and open geolinguistic data platform for mapping the world's languages. *Glottography* represents the speaker areas of the world's languages as georeferenced spatial polygons, enriched with relevant metadata, including Glottocodes that link each polygon to a unique identifier in Glottolog, a database cataloguing the world's dialects, languages, and language families. *Glottography* currently includes more than 13,000 language areas of 5,300 distinct languages, digitised from 29 source publications. For each source, the platform provides the data in its raw, unmodified form and aggregated at the levels of languages and language families, according to the classification in Glottolog. *Glottography* is accessible through *Rglottography*, an R package, and is accompanied by detailed tutorials for usage and data acquisition that encourage users to contribute new geodata to the platform. Being the first open data source of its kind, *Glottography* enables computational analyses that explore the origins, distribution, and drivers of global linguistic diversity.

KEYWORDS:
language mapping; speaker areas; Glottolog; language diversity; open geospatial data; Cross-Linguistic Data Format

TO CITE THIS ARTICLE:
Ranacher, P., Forkel, R., Efrat-Kowalsky, N., Urban, M., Hehli, A., Franz, M., Biland, G., Kreienbühl, A., Rodríguez, A. H., Azevedo, M. C. B. C., Giebler, J., Takahashi, T., Neureiter, N., van Gijn, R., Roose, M., Vesakoski, O., Weibel, R., Kaiping, G., & Norder, S. (2026). Glottography: An Open-Source Geolinguistic Data Platform for Mapping the World's Languages. *Journal of Open Humanities Data*, 12: 47, pp. 1–16. DOI: https://doi.org/10.5334/johd.459

DOI: 10.5334/johd.459

## 1 CONTEXT AND MOTIVATION

Humans speak, sign, and write in over 7,000 languages customarily grouped into more than 400 language families. These languages vary widely in their geographic range. Some are confined to a handful of villages and are spoken by communities of just a few hundred — or even only a few dozen — speakers. Others have global reach, spanning continents and climate zones, and are spoken by millions.

Numerous print publications feature excellent language maps, such as the *Atlas of the World's Languages* (Asher & Moseley, 2007), which depicts the geographic locations of over 6,000 languages across 150 maps. Language maps are indispensable resources for linguistic research: they capture the distribution of human languages across geographic space, showing the spatial location and extent of dialects, language branches, and entire language families. Maps document language diversity and bear witness to its rapid decline. While traditional language maps provide immense cartographic and scientific value, they are not easily accessible for linguistic or geographic analysis, since all spatial information is encoded solely within static paper images. In contrast, geographic databases store languages as digital spatial data, enabling their integration with other datasets, quantitative geographic analysis, and mapping. In this way, digital spatial data transform maps from illustrations into analytical tools, opening new possibilities for understanding how languages relate to each other and the physical world they inhabit.

Several resources offer digital language maps, often focusing on specific regions such as the Caucasus (Dahl & Veselinova, 2005), North America (Haynie & Gavin, 2019), Australia (AIATSIS, 1996), the North Pacific (Alaskool, 1998), and on minority languages in cities like New York (Perlin et al., 2022). While these resources are valuable in their own right, they vary in quality and often lack geographic source data or detailed metadata. Notable exceptions include the digital edition of Wurm & Hattori's Language Atlas of the Pacific Area (Forkel & Hammarström, 2024), the Uralic Language Atlas (Rantanen et al., 2022) and the digitised Atlas of the World's Languages (Ranacher et al., 2025), which are available as proper geospatial data, i.e. polygons or point geometries that place languages in a geographic coordinate reference system and provide appropriate linguistic metadata. This work builds on these previous efforts, particularly the digitised Atlas of the World's Languages, with the aim of harmonising data collection across sources and providing a common standard and platform for linguistic geodata on language areas.

There have also been previous attempts to create worldwide digital language mapping platforms. The two major ones are Glottolog and Ethnologue. Glottolog (Hammarström et al., 2025) is a digital catalogue of the world's languages, language families, and dialects. Its Glottocode system has become a *de facto* standard for identifying languages, especially those with limited documentation. While Glottolog provides geographic information for many of the world's languages, this is not its primary purpose. Glottolog represents languages as digital point locations — single latitude and longitude coordinates — but it does not provide language polygons — areas with defined boundaries indicating the geographic region where a language is spoken.

In contrast, Ethnologue (Eberhard et al., 2024) provides language area polygons through its World Language Mapping System. Ethnologue employs the ISO 639-3 standard for language identification, which it also maintains. However, the World Language Mapping System is proprietary and behind a paywall, hampering scientific reproducibility (Matacic, 2020). This stands in stark contrast to the principles of open science upheld by Glottolog, which provides free access to its data under a Creative Commons licence. Open science advocates for scientific knowledge to be freely accessible and for the entire scientific process to be transparent, reproducible, and accountable. Transparency and accountability are especially important in language mapping. Language areas are inherently subjective representations of reality, making it necessary to be transparent about how they are created. Moreover, language areas convey complex social and cultural meanings beyond mere linguistic presence, requiring map creators to be accountable for their content.

Unlike physical features such as rivers or mountains, which are directly tied to geographic space, language areas are socially constructed (Toan, 2024). Mapping them involves interpretive choices and remains partly subjective, reflecting both empirical observations of language use in a region and the mapper's judgment in aggregating these observations into a coherent language area. While maps always reflect the perspectives of their creators, this is especially true for socially constructed spaces. Yet most language maps depict areas with sharp

boundaries, implicitly suggesting that such areas are precisely measurable. This is problematic even when maps are used for strictly scientific purposes, as they can be misleading by implying a level of precision that does not exist. However, the impact of maps extends beyond merely depicting linguistic landscapes for research; they are powerful political instruments that convey authority. In many contexts, language signifies group membership and is pivotal in shaping community identity and, in some cases, nation-building. Consequently, language maps are perceived as visual representations of a community's sphere of influence in geographical space. This is exemplified by the Map of Indigenous Australia, which is always published with an explicit disclaimer stating that it "is not suitable for native title or other land claims" (AIATSIS, 1996). In this respect, maps can serve as political instruments for both empowerment and oppression. When used for empowerment, maps enable speaker communities to define their homeland and assert their right to territorial self-determination. For example, Native Land Digital (2025) provides a platform where indigenous communities can represent themselves and their histories, including a map that displays language and territorial areas. Conversely, when used for oppression, maps may act as ideological tools to reinforce or even expand a nation's territorial claims (Mankoff, 2022).

In conclusion, there is a need for a geolinguistic data platform that maps the world's languages in geographic space and adheres to the principles of open science. The platform should depict languages as areas rather than points and unambiguously identify these with Glottocodes, making them readily accessible for linguistic and geographic analysis. It should acknowledge the plurality and subjectivity of perspectives concerning their locations, and provide relevant metadata and scientific references for full accountability. Finally, the platform should recognise that maps are political instruments and encourage active community participation in the mapping process, particularly from communities whose territories and languages are under threat.

We present *Glottography*, a geolinguistic data platform for mapping the speaker areas of the world's languages. *Glottography* represents the geographic locations of languages as polygons, along with relevant metadata, including Glottocodes that uniquely identify each language. All speaker areas are digitised from scientific literature, with *Glottography* providing references to the corresponding source maps to ensure full accountability. To capture the uncertainty associated with language areas, the platform includes multiple polygons for most languages, sourced from different references. *Glottography* provides the data via GitHub, where users can comment on the quality of specific entries and suggest improvements or changes. Finally, tutorials guide users in contributing their own geodata to *Glottography*, enabling them to do so with proper support and guidance.

## 2 DATASET DESCRIPTION

### REPOSITORY LOCATION

All *Glottography* datasets are openly available through the Zenodo *Glottography* community at https://zenodo.org/communities/glottography (Accessed: 2026-02-19). Individual DOIs for each dataset are provided in the *DOI* column of Table 1. The datasets are maintained via the *Glottography* organisation on GitHub,[1] where users can track changes over time, report issues, suggest improvements, and contribute updates. The *Rglottography* package (Ranacher, 2026b) provides a convenient interface for downloading *Glottography* datasets and importing them directly into the R programming environment. The package is developed and maintained on GitHub through the *Rglottography*[2] repository and archived through Zenodo to ensure long-term accessibility and versioned releases.

### REPOSITORY NAME

The *Glottography* community on Zenodo (for official releases) and the *Glottography* organisation on GitHub (for maintenance and feedback).

### OBJECT NAME

All dataset names are listed in the *Dataset name* column of Table 1.

1 https://github.com/Glottography (Accessed: 2026-02-19).

2 https://github.com/Glottography/Rglottography (Accessed: 2026-02-19).

### FORMAT NAMES AND VERSIONS

Each dataset is provided in Cross-Linguistic Data Format (CLDF), a standard for historical and typological language data (Forkel et al., 2018). Each source is assigned its own repository, structured as follows:

- The *etc* folder contains CSV files with attribute data and BibTeX files referencing the source publications.
- The *raw* folder contains the speaker area polygons in GeoJSON format.
- The *cldf* folder stores the CLDF datasets, which aggregate the speaker area polygons according to the classification in the source publication, as well as at the Glottolog language and language family levels.

### DATASET CREATORS

All datasets were created by the *Glottography* consortium, whose members are the authors of this publication. In addition, the *Source publication* column of Table 1 lists the authors of the original sources who provided the primary maps and data.

### LANGUAGE

English.

### LICENSE



Table 1 lists all currently available *Glottography* datasets together with their source publication, linguistic and geographic coverage, time period covered, number of languages and dataset name and DOI. The (spatial) overlap column indicates whether two or more language areas in the source publication can overlap geographically and share the same space on the map, with the value *partly* indicating that some, but not all, maps in the source contain overlapping polygons.

## 3 METHODS

*Glottography* provides digital polygons representing the areas of languages, sourced from the scientific literature. Only sources that were citable, uniquely identifiable, and accompanied by complete bibliographic metadata were considered. The initial collection was populated with sources recommended by collaborators and coauthors to ensure broad global coverage and later supplemented with publications addressing underrepresented regions. A typical workflow includes the following steps, largely following the approach outlined in Ranacher et al. (2025):

1. Georeferencing the map and placing it in a coordinate reference system (CRS).
2. Digitising the map and converting the language areas into digital polygons.
3. Recording language attributes and metadata from the source publication.
4. Linking the language areas with Glottocodes, unique identifiers for languages maintained by Glottolog.
5. Curating the digitised polygons, attributes, and metadata, and converting them into a Cross-Linguistic Data Format (CLDF) dataset ready for upload to *Glottography*.

Language maps typically come in four formats, each requiring different georeferencing and digitising steps:

i) Physical images in printed publications were scanned at $\geq 400$ dots per inch (DPI), exported as Tagged Image File Format (TIFF) files, then georeferenced and digitised in QGIS (2025), an open-source geographic information system (GIS). Source publications: *asher2007world*.

ii) Digital raster images, typically embedded in PDFs, were copied directly or captured via high-resolution screenshots, saved as TIFF files, and then georeferenced and digitised in QGIS. Source publications: all remaining sources not listed in (i), (iii), or (iv).

| SOURCE PUBLICATION | LINGUISTIC AND GEOGRAPHIC COVERAGE | TIME PERIOD | OVERLAP | NO. LANGUAGES | DATASET NAME | DOI |
|---|---|---|---|---|---|---|
| Allen et al. (2016) | central California | before TOC | no | 11 | allen2016resource | https://doi.org/10.5281/zenodo.17333165 |
| Asher & Moseley (2007) | global | contemporary | partly | 4064 | asher2007world | https://doi.org/10.5281/zenodo.15287258 |
| | global | TOC | partly | 4503 | | |
| Bouckaert et al. (2012) | Indo-European in Europe & southern Asia | contemp. & ancient | yes | 70 | bouckaert2012indoeuropean | https://doi.org/10.5281/zenodo.17333413 |
| Bowern (2021) | Australia | traditional | no | 326 | bowern2021australia | https://doi.org/10.5281/zenodo.17334090 |
| Bowern & Atkinson (2012) | Pama-Nyungan in Australia | traditional | no | 7 | bowern2012pama-nyungan | https://doi.org/10.5281/zenodo.17333460 |
| Carling & Gippert (2025) | global | contemporary | yes | 800 | carling2025diacl | https://doi.org/10.5281/zenodo.17334192 |
| Dedio et al. (2019) | Indo-European on British Isles & in northern Europe | 800–1900 AD | yes | 31 | dedio2019britain | https://doi.org/10.5281/zenodo.17334236 |
| Denevan (1966) | northeastern Bolivia | around 1700 AD | no | 5 | denevan1966aboriginal | https://doi.org/10.5281/zenodo.17334281 |
| Edwards (2020) | Timor island | contemporary | no | 39 | edwards2020metathesis | https://doi.org/10.5281/zenodo.17338066 |
| Eriksen (2011) | Amazonia | TOC | partly | 102 | eriksen2011nature | https://doi.org/10.5281/zenodo.17339139 |
| Figueira (1982) | Argentina | contemporary | no | 11 | figueira1982atlastotalargentina | https://doi.org/10.5281/zenodo.17339172 |
| Goddard (1999) | North America | before TOC | no | 286 | goddard1999native | https://doi.org/10.5281/zenodo.17339338 |
| Grierson (1903) | India | contemporary | no | 112 | grierson1903lsi | https://doi.org/10.5281/zenodo.17340138 |
| Haynie & Gavin (2019) | North America | before TOC | no | 350 | haynie2019modern | https://doi.org/10.5281/zenodo.17340247 |
| Hochstetler et al. (2004) | Dogon in West Africa | contemporary | no | 14 | hochstetler2004sociolinguistic | https://doi.org/10.5281/zenodo.17340571 |
| Matsumae et al. (2021) | northeast Asia | contemporary | no | 11 | matsumae2021exploring | https://doi.org/10.5281/zenodo.17340654 |
| Messineo (2011) | Gran Chaco, South America | contemporary | no | 16 | messineo2011aproximacion | https://doi.org/10.5281/zenodo.17340771 |
| Ministerio de Educación de Argentina (2009) | Argentina | contemporary | no | 12 | ministerio2009pueblos | https://doi.org/10.5281/zenodo.17340812 |
| Queixalos & Renault-Lescure (2000) | northern South America | contemporary | no | 206 | queixalos2000linguas | https://doi.org/10.5281/zenodo.17341026 |
| Rantanen et al. (2021) | Northern Europe & northwestern Siberia | contemporary & traditional | yes | 41 | rantanen2021uralic | https://doi.org/10.5281/zenodo.17341268 |
| Schapper (2020) | Papuan on Alor-Pantar | contemporary | no | 18 | schapper2020papuan | https://doi.org/10.5281/zenodo.17341890 |
| Suttles & Suttles (1985) | American Northwest Coast | traditional | no | 14 | suttles1985northwest | https://doi.org/10.5281/zenodo.17341948 |
| Steever (2019) | Dravidian in South Asia | contemporary | no | 98 | steever2019dravidian | https://doi.org/10.5281/zenodo.17341914 |
| Tarble de Scaramelli & Zucchi (1984) | Amazonia | traditional | no | 10 | tarble1984nuevos | https://doi.org/10.5281/zenodo.17341986 |
| Vuillermet (2012) | Amazonia | contemporary | no | 27 | vuillermet2012grammar | https://doi.org/10.5281/zenodo.17342021 |
| Walker & Ribeiro (2011) | Arawakan in South America | contemporary | no | 30 | walker2011bayesian | https://doi.org/10.5281/zenodo.17342060 |
| Wikipedia contributors (2024) | global | contemporary | yes | 158 | wikipedia2024officiallang | https://doi.org/10.5281/zenodo.17342116 |
| Wurm & Hattori (1981) | Australia, Papunesia & southeast Asia | contemporary | no | 1921 | wurm1981pacific | https://doi.org/10.5281/zenodo.17342180 |
| Zucchi (2017) | Arawakan in Amazonia | traditional | no | 10 | zucchi2017arqueologia | https://doi.org/10.5281/zenodo.17342211 |

**Table 1** Summary of *Glottography* datasets at the time of manuscript publication.

TOC: Time of (European) contact.

iii) Digital vector geometries without explicit geographic reference were extracted computationally and georeferenced in QGIS; no digitisation was required. Source publications: *haynie2019modern*.

iv) Digital vector geometries with explicit geographic reference already contained valid spatial polygons, so georeferencing and digitising were not necessary; the polygons were cleaned, reprojected, and standardised as needed. Source publications: *bouckaert2012indoeuropean*, *bowern2021australia*, *carling2025diacl*, *dedio2019britain*, *grierson1903lsi*, *matsumae2021exploring*, *rantanen2021uralic*, *steever2019dravidian*, *wurm1981pacific* and *wikipedia2024officiallang*.

The *wikipedia2024officiallang* dataset of official languages by country and territory from Wikipedia (2024) constitutes a special case: the source provides textual listings of official languages, which we mapped to country polygon geometries from Natural Earth (2024).

A series of tutorials (Ranacher, 2026a) helps *Glottography* users contribute their own data to the platform. The tutorials are available via GitHub[3] and archived on Zonodo. The following sections provide a step-by-step summary of the workflow, including links to the relevant tutorials.

## 3.1 GEOREFERENCING SOURCE MAPS

Georeferencing[4] assigns geographic coordinates to either a language map image (formats i and ii) or vector geometries without spatial reference (format iii), enabling accurate alignment in a GIS. It was performed using the `Georeferencer` plugin in QGIS, where the unreferenced map (image or vector) is displayed alongside a reference basemap with landforms or administrative boundaries. Shared, recognisable features such as coastal bends or river estuaries were used to place control points across both layers. These control points aligned the two layers by shifting and warping the language map (language geometries) from its original position to its correct spatial location. If noticeable distortions appeared in specific regions, additional control points were added locally. If the CRS of the source map was known, we used it as the target CRS for the georeferenced map; otherwise, we applied a generic global CRS, e.g., Web Mercator (EPSG:3857) or WGS84 (EPSG:4326).

## 3.2 DIGITISING LANGUAGE POLYGONS

Digitising[5] involves tracing the outlines of language areas on the georeferenced map image and converting them into polygon geometries, and was needed for formats i and ii. The georeferenced language maps were exported as GeoTIFF and digitised using the `Advanced Digitising` toolbar in QGIS, following one of two approaches. Polygons were either traced from scratch using the *Add Feature* tool or derived by cutting them from existing geometries of the Earth's landmasses and major islands (Natural Earth, 2024) using the *Split Features* tool. Adding features is generally faster and involves a simpler workflow in QGIS, whereas cutting from existing geometries ensures that language areas align with landmasses, coastlines, and major lake shores. Adding features was preferred for regional inland language maps (e.g., for *hochstetler2004sociolinguistic*), while splitting features was better suited to maps covering continental-scale or coastal language areas (e.g., for *asher2007world*). In both cases, we digitised the polygons in the CRS of the georeferenced map and later reprojected them to a common CRS (WGS84, EPSG:4326).

For format iv), where spatial polygons were already available, any apparent geometric issues — such as invalid geometries that failed to close — were corrected, but the polygons were otherwise left unchanged. Finally, all polygons from a single source representing the same language were aggregated. Aggregation resulted in a single polygon geometry if the polygons formed one contiguous area, or a MultiPolygon geometry — a combined geometry consisting of multiple disjoint polygons — if they were spatially separated. All (Multi)Polygons from a single source were saved in a single GeoJSON file.

3 https://glottography.github.io/tutorials (Accessed: 2026-02-19).

4 A georeferencing tutorial is available at https://glottography.github.io/tutorials/georeferencing/ (Accessed: 2026-02-19).

5 A digitising tutorial is available at https://glottography.github.io/tutorials/digitising/ (Accessed: 2026-02-19).

### 3.3 ATTRIBUTES AND METADATA

*Glottography* records relevant attributes and metadata[6] to uniquely identify and describe each source publication, language map, and speaker area polygon depicted in each map. A citation key in the format `authorYYYYtopic` identifies the source publication. The citation key points to a scientific reference stored in BibTeX format, which includes bibliographic metadata, such as the author(s), year, type of publication, and title. References were retrieved from official publisher websites, Glottolog, or Google Scholar, with manual corrections made to address any missing or inaccurate entries.

For each digitised speaker area polygon, we recorded the following attributes: the language *name* as it appears on the map; a unique numeric *id* to uniquely identify the polygon; the *year* the language area refers to, which can be the date indicated on the source map or, if not explicitly given, the publication year of the source; the *full map name(s)* used to uniquely identify the language map(s) in the source publication; if a language area spans multiple maps, all map names are listed, separated by a vertical bar; the *Glottocode*, a unique identifier assigned to the language (see next section); and a *note*, for any additional comments or annotations about the language area, for example if the language area was poorly visible on the map.

### 3.4 LINKING TO GLOTTOLOG

*Glottography* uses Glottocodes[7] to uniquely identify each language polygon. Glottocodes are standardised, unambiguous identifiers for language varieties, maintained by Glottolog (Hammarström et al., 2025). They provide a consistent way to reference languages, dialects, and language families. Although some publications, such as Haynie & Gavin (2019), already included Glottocodes, most other publications lacked them. For these, we assigned Glottocodes to the language areas following the approach in Ranacher et al. (2025). We used the Python 3 `guess_glottocode` package (Ranacher, 2025), which filters candidate languages based on spatial proximity and then identifies the most suitable candidate using large language models (LLMs) and web crawling, or, if necessary, manual annotation.

### 3.5 DATA CURATION

Data curation[8] aggregates the raw speaker area polygons according to the classification in the source publication, or by language or language family as defined in Glottolog, and exports them in CLDF format with all polygons in the WGS84 CRS (EPSG:4326). We used the `pyglottography` Python 3 package to create three sets of vector geometries in GeoJSON format, each enriched with Glottocodes at different levels of aggregation:

- *Features*: Speaker areas retaining the classification from the source publication.
- *Language areas*: Speaker areas aggregated at the language level according to Glottolog's classification.
- *Family areas*: Speaker areas aggregated at the top-level language families according to Glottolog's classification.

### 3.6 ERROR CORRECTION

We addressed potential errors arising during data collection and performed additional quality checks, both automated and manual.[9] For geometry validation, we used `shapely's is_valid` function in Python 3. Name checks compared the language names on the map with those in Glottolog, including any known alternative names. Glottocode checks ensured that each code was present in the Glottolog database and conformed to the correct format. While we corrected errors introduced during data processing, we generally did not alter the original

6 A tutorial on recording attributes and metadata is available at https://glottography.github.io/tutorials/metadata/ (Accessed: 2026-02-19).

7 A tutorial on assigning Glottocodes is available at https://glottography.github.io/tutorials/glottocodes/ (Accessed: 2026-02-19).

8 A data curation tutorial is available at https://glottography.github.io/tutorials/glottocodes/ (Accessed: 2026-02-19).

9 An error correction tutorial is available at https://glottography.github.io/tutorials/correction/ (Accessed: 2026-02-19).

data except for clear mistakes. Disputable mappings were preserved, as each source was treated as a valid, subjective interpretation of a language's geographic area, consistent with our data collection policy.

In a project on the scale of *Glottography*, error correction will likely be an ongoing endeavour. To address this, we implemented a workflow designed to make error identification and resolution straightforward.

## 4 RESULTS AND DISCUSSION

*Glottography* currently includes speaker areas from 29 source publications. It comprises 17,114 features, which retain the classifications of the source publications and can therefore be mapped to Glottolog entries at the dialect, language, or subfamily level. The features cover 7,562 distinct Glottolog entries, indicating that many entries are associated with multiple speaker areas from different sources. The features are aggregated into 13,303 language areas according to Glottolog's classification, covering 5,338 distinct languages. Finally, the features are aggregated into 1,390 top-level family areas, covering 394 distinct Glottolog families. Three sources provide global coverage, with the dataset name in brackets corresponding to Table 1: the *Atlas of the World's Languages* (*asher2007world*), the DiACL/TITUS Polygon Archive (*carling2025diacl*) and the *Official Languages by Country and Territory* from Wikipedia (*wikipedia2024officiallang*). All other sources focus on specific geo-linguistic macro-areas, as defined by Hammarström & Donohue (2014): ten on South America, seven on Eurasia, four on North America, three each on Australia and Papunesia, and one on Africa.

### 4.1 COVERAGE

We assessed the geographic density of language polygons in the *Glottography* data using an equal-area global hexagonal grid (CRS: EPSG:8857; Šavrič et al. 2019) (Figure 1). For each grid cell, we counted the number of language polygons that intersected it, considering only unique languages. Density is generally highest in regions known for high language diversity, such as Papua New Guinea, West Africa, and the upper Amazon in Bolivia and Peru. *Glottography* incorporates languages from multiple sources and time periods, including extinct languages; therefore, polygon density should not be interpreted as a direct measure of actual language diversity.

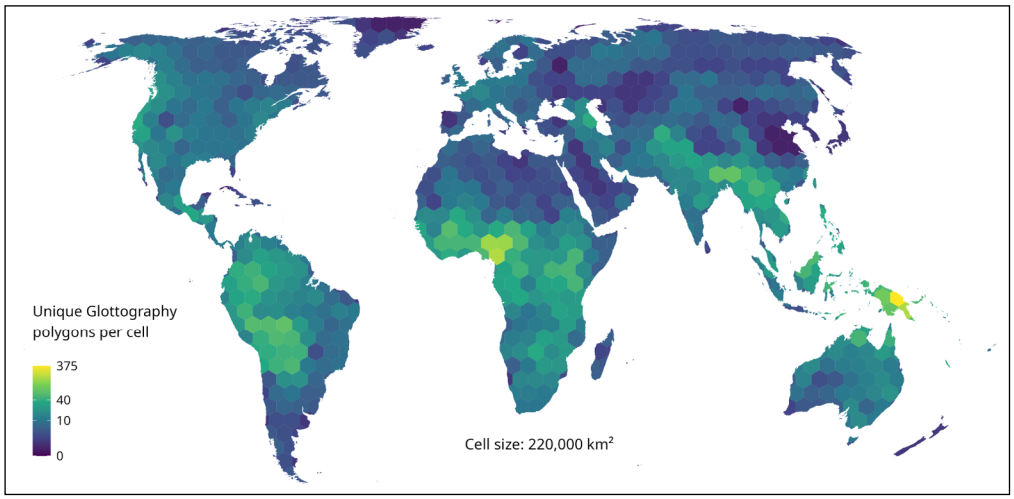


**Figure 1** Geographic language polygon density in *Glottography*. The map shows the number of unique language polygons per grid cell. The colour gradient is pseudo-logarithmic, emphasising differences in order of magnitude.

We also evaluated the coverage across the 25 major language families by number of languages (Figure 2). Most families are well represented, but there are several notable exceptions with low coverage, including Indo-European, Dravidian, Otomanguean, Tai-Kadai, Pidgin, and Hmong-Mien. This highlights immediate future directions for extending coverage and filling data gaps.

### 4.2 COMPARISON WITH ETHNOLOGUE AND GLOTTOLOG

We compared *Glottography's* coverage with two established reference datasets: the paywalled polygons provided by Ethnologue (Eberhard et al., 2024) and the language point coordinates provided by Glottolog (Hammarström et al., 2025). Ethnologue includes a total of 7,651

DOI: 10.5334/johd.459

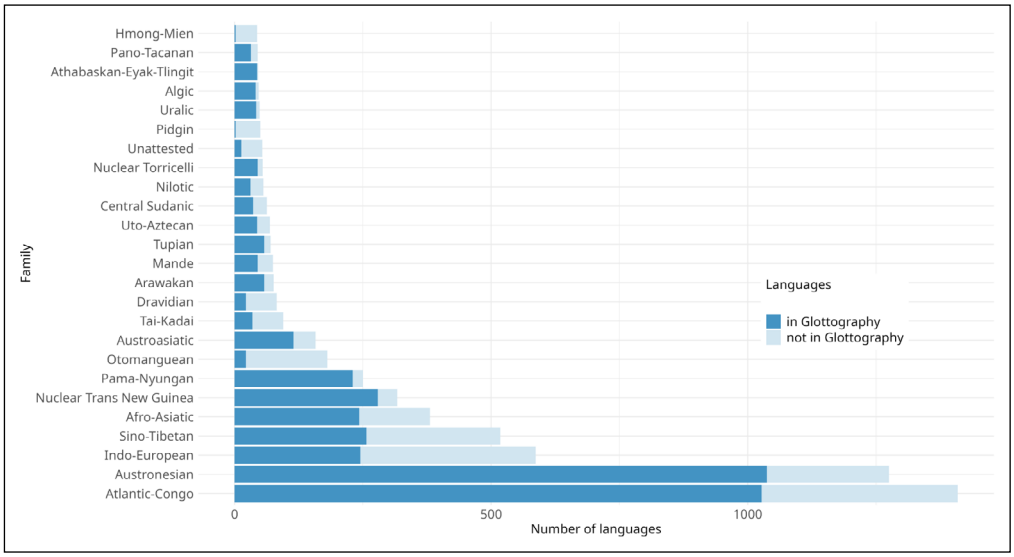


**Figure 2** Number of languages included in and missing from *Glottography* for the 25 largest language families (by number of languages) in Glottolog.

unique polygons, compared to 5,338 language polygons in *Glottography*. To evaluate regional differences between the two datasets, we intersected the hexagonal grid with polygons from both sources, counted the number of unique languages in each grid cell, and computed the coverage difference (*Glottography* – Ethnologue) (Figure 3). Because Ethnologue uses the ISO 639-3 standard for language identification rather than Glottocodes, there is no fully compatible way to filter language polygons across both datasets. We compared the language polygons in *Glottography* with all entries in Ethnologue. Since some Ethnologue polygons are not classified as languages by Glottolog, this comparison tends to favour Ethnologue. For example, Serbian, Croatian, and Bosnian are treated as dialects in Glottolog and are therefore aggregated under the single language Serbian-Croatian-Bosnian in *Glottography*, whereas Ethnologue treats them as three distinct languages.

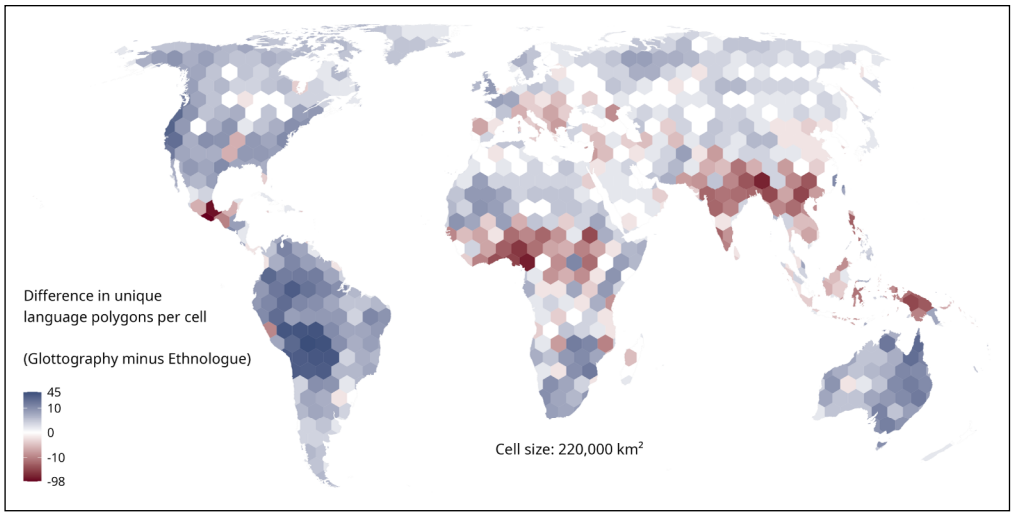


**Figure 3** Comparison of coverage in *Glottography* and Ethnologue. The map shows the difference in the number of unique language polygons per grid cell (*Glottography* minus Ethnologue). Positive values indicate greater coverage in *Glottography* (blue), negative values greater coverage in Ethnologue (red). Differences near zero (white) indicate similar coverage across datasets. The colour gradient is pseudo-logarithmic, highlighting differences in order of magnitude.

Overall, *Glottography's* coverage closely matches that of Ethnologue. It even exceeds Ethnologue across large parts of the upper Amazon in Peru and Bolivia, the northwestern United States and Canada, and northern Australia. In southern Africa and northern Asia, *Glottography* generally matches, and in some areas slightly exceeds, Ethnologue's coverage. The weakest regions are Mesoamerica, Papua New Guinea, West Africa, India, and Southeast Asia, where *Glottography* currently falls short of Ethnologue.

Because Glottolog provides only point coordinates rather than polygons, coverage per grid cell cannot be directly compared between the two sources. Instead, we indicate which of the languages with point coordinates in Glottolog have a corresponding polygon in *Glottography* (Figure 4). The results are consistent with those obtained when comparing *Glottography* to Ethnologue. Overall, *Glottography* includes polygons for 65% of the 8,300 languages with point coordinates in Glottolog. Notable gaps remain in Mesoamerica, Europe, Papua New Guinea, West Africa, India, and Southeast Asia, where *Glottography* lacks Glottolog languages.

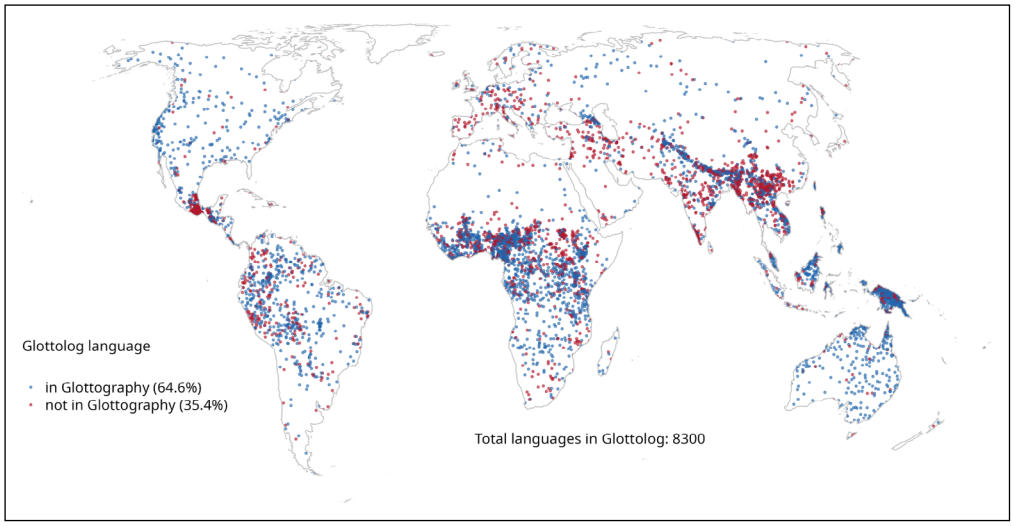


**Figure 4** Languages in Glottolog with a corresponding polygon in *Glottography* (blue) and without one (red). Point locations are based on Glottolog.

## 4.3 EXEMPLARY LANGUAGES IN GLOTTOGRAPHY

We argue that language maps should reflect the plurality of perspectives regarding a language's location. To address this, *Glottography* incorporates multiple polygons per language, each sourced from different references.

We illustrate examples of language polygons from multiple sources for four languages (Figure 5): Shona (sub-Saharan Africa), Bengali (South Asia), Algonquin (North America), and Bulgarian (Europe). For Shona, *asher2007world* includes most of Zimbabwe, except for the southwest, which it assigns to Ndebele, and the areas of Zambia surrounding Lake Cahora Bassa. *carling2025diacl* largely follows *asher2007world* but additionally includes parts of Mozambique within the Shona area. The Shona polygon in *wikipedia2024officiallang* corresponds to the national territory of Zimbabwe.

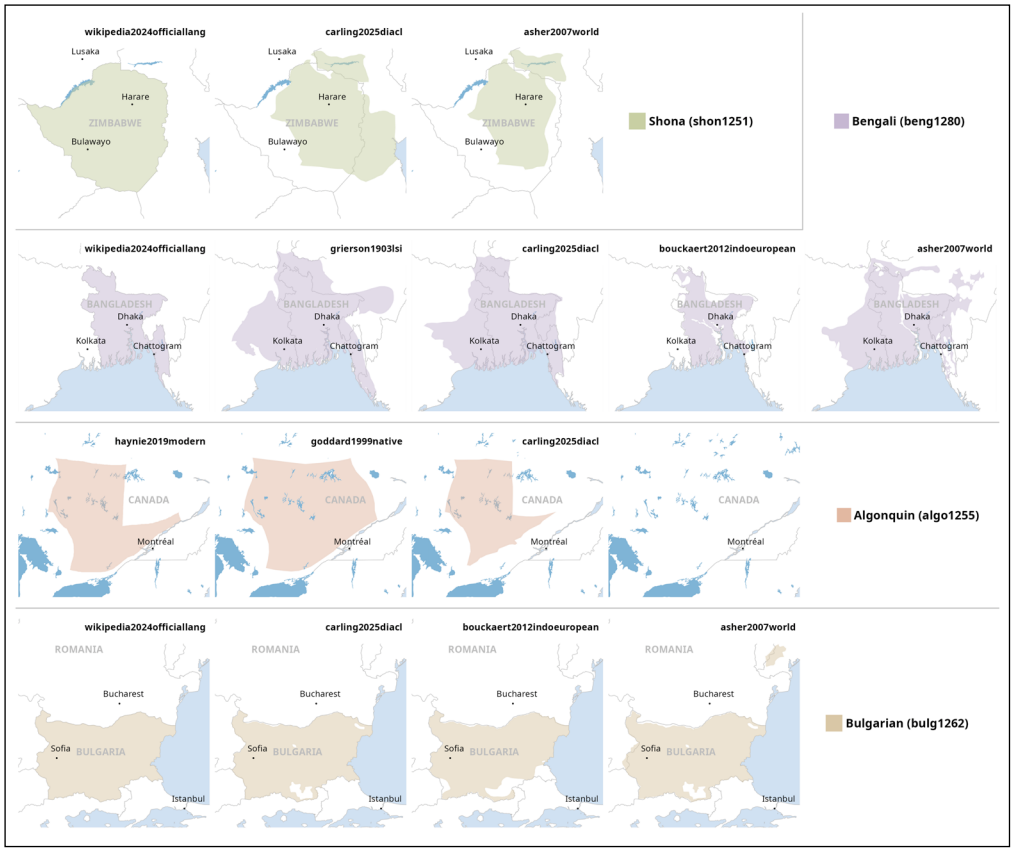


**Figure 5** *Glottography* language polygons for Shona, Bengali, Algonquin, and Bulgarian from multiple sources. The contemporary and traditional polygons in the Atlas of the World's Languages are treated as distinct sources. As these polygons are identical outside the Americas and Australia, only one is included.

There are four sources for Algonquin: *haynie2019modern*, *goddard1999native*, *asher2007world*, and *carling2025diacl*. They largely agree that the language was traditionally spoken northwest of present-day Montreal, within the provinces of Ontario and Quebec. *goddard1999native* differs in that it also includes a large area north of Montreal that is absent from the other sources.

For Bengali, *asher2007world* includes most of Bangladesh, as well as parts of India: to the west (West Bengal) and to the east (Tripura, Meghalaya, and Assam). *carling2025diacl* covers Bangladesh, West Bengal, and Tripura in India. *bouckaert2012indoeuropean* includes most of Bangladesh except for parts of Rangpur, Rajshahi, and Mymensingh in the north. *grierson1903lsi* covers Bangladesh, Tripura, Meghalaya, and Assam in India, and even the border area with Myanmar south of Chattogram. The Bengali polygon in *wikipedia2024officiallang* corresponds to the national territory of Bangladesh.

The Bulgarian language area in *asher2007world* includes parts of Bessarabia along the border between Moldova and Ukraine, but excludes several regions in central, southern, and northeastern Bulgaria, which are instead assigned to Gagauz. *bouckaert2012indoeuropean* similarly excludes these areas from Bulgarian, additionally assigns western Bulgaria to Macedonian, but omits Bessarabia. *carling2025diacl* largely follows the polygons in *asher2007world*, omitting the enclaves and likewise excluding Bessarabia. Finally, the polygon for Bulgarian in *wikipedia2024officiallang* corresponds to the national territory of Bulgaria.

At present, most languages in *Glottography* are represented by only one or two sources, while languages with more sources are relatively rare. As several publications focus on specific languages within a region—for example, exclusively Indo-European languages in Northwestern Europe—they contribute additional sources for some languages.

### 4.4 AGGREGATION AND ITS IMPLICATIONS

*Glottography* provides speaker areas at three levels of aggregation: as *features*, *languages*, and top-level *families*, as defined by Glottolog. Here, we briefly discuss aggregation and some of its implications.

The features always retain the classification of the original publication. For example, *asher2007world* views Danu and Intha in Myanmar as separate languages; hence, the features include separate speaker area polygons for both. In contrast, Glottolog views Danu (danu1251) and Intha (inth1239) as dialects of the Danu-Intha language (inth1238). When aggregating at the language level, the polygons for Danu and Intha are combined into a single Danu-Intha language area, following Glottlog's classification.

Wikipedia lists Mongolian (mong1331) as the official language of Mongolia, whereas Glottolog treats it as a subfamily of Mongolic-Khitan, with daughter languages Halh Mongolian, Oirad-Kalmyk-Darkhat, and Peripheral Mongolian. When aggregating *wikipedia2024officiallang* at the language level, Mongolian is therefore excluded, since mapping it unambiguously to a single daughter language is not possible. More generally, polygons corresponding to Glottolog (sub-) families are omitted at the language level; this applies, for example, to Uzbek and Azerbaijani in *wikipedia2024officiallang* and to Malagasy in *asher2007world*. A special case is the traditional speaker areas of Australia in *asher2007world*, where multiple languages can share one area. We mapped these to the closest common subgroup in Glottolog at the feature level (see also Ranacher et al. 2025).

The choice of whether to use feature-, language-, or family-level speaker areas depends on the specific use case. We recommend using feature polygons when the classification in the source is deemed appropriate, or when the classification itself is not crucial to the task at hand. For example, when creating a map for Danu, it may not matter that *asher2007world* treats it as a language, whereas Glottolog classifies it as a dialect. When consistent classification is important, we instead recommend using language- and family-level polygons, which follow Glottolog's classification of languages and (top-level) families. In all cases, we advise readers to consult the relevant Glottolog entry and inspect the corresponding polygon geometry to ensure that speaker areas at a given level of aggregation are appropriate for their use case.

## 5 IMPLICATIONS/APPLICATIONS

*Glottography* provides data on the current and past spatial distribution of languages, supporting multiple lines of research on cultural and linguistic evolution. The most straightforward application is to use *Glottography* polygons to create custom, high-resolution maps for individual languages, entire language families, or specific geographic regions. The *Rglottography* package in R (Ranacher, 2026b) provides tutorials that demonstrate how to use *Glottography* data in R to create simple language maps (e.g., Norder et al. 2022).

Researchers can leverage *Glottography* data to test hypotheses about the distribution of languages and language families in space and to reevaluate existing claims from a new perspective. For example, previous studies have suggested that language ranges near the poles tend to be larger than those closer to the equator (Collard & Foley, 2002; Gavin & Stepp, 2014; Mace & Pagel, 1995). Such claims are typically based on language densities derived from grid maps that use point locations for languages rather than their full spatial extents.

*Glottography* captures language areas at different points in time, allowing researchers to reconstruct language history in space and to explore the geographic factors that have shaped it (Takahashi et al., 2023). Language polygons can also be incorporated into phylogeographic analyses, which reconstruct the spatial spread of a language family alongside its diversification from a common ancestor. While current phylogeographic models typically rely on point geometries (Bouckaert et al., 2012), incorporating polygons may offer a more realistic representation of diffusion.

## 5.1 LIMITATIONS AND FUTURE WORK

We aim to establish *Glottography* as a free, open, and community-maintained repository for collecting digital areal information about languages, analogous to the role that Glottolog plays in language classification. To achieve this goal, current limitations must be addressed.

While more than 60% of languages in Glottolog have a polygon (Figure 4), coverage remains low in certain regions. We aim to expand the dataset by collecting additional language areas, particularly for geographic regions and time periods with limited coverage. We also seek to engage other researchers in refining the standards for *Glottography* and integrating their language areas of interest into the repository. Ideally, contributions will extend beyond the academic community, allowing also speakers of (minority) languages to map the areas of their own languages. A first step in this direction is the set of tutorials,[10] which provide detailed instructions for collecting and contributing data to *Glottography* and are available to anyone wishing to participate in the project.

*Glottography* treats all scientific sources that meet the formal requirements for inclusion equally, with a strong emphasis on transparency: no source is considered inherently better than another. Currently, the most comprehensive dataset with full global coverage is *asher2007world*. Still, this dataset has limitations. Coverage is relatively sparse in some regions, notably Western Africa. In addition, the classification of linguistic varieties into separate languages, dialects, or subfamilies does not always align with that in Glottolog, for example in the case of Australian Aboriginal languages. For these, other datasets (e.g., *hochstetler2004sociolinguistic*, *bowern2021australia*) likely provide more suitable alternatives. A logical next step would be to compile a *consensus* dataset that selects the most appropriate language polygon for each language entry in Glottolog. However, given that Glottolog currently lists more than 5,000 unique languages, such an undertaking is beyond the scope of this manuscript. Achieving this goal would require substantial community collaboration, together with sustained feedback and guidance from domain experts specialising in particular regions and language families. As a first step in this direction, we encourage the research community to provide feedback via GitHub's *Issues* mechanism, in the form of objections or comments on specific speaker areas, their geometry and metadata.

Language areas are constructed spaces that indicate the presence of a language in a specific geographic region. They are derived from the presence of individuals speaking, signing, or writing a particular language, making them a useful simplification of linguistic reality. One could also imagine that individual utterances tagged with a language and a location, or the prolonged presence of a speaker producing them, could directly signal the presence of a language in a region. While this approach stays closer to discretely measurable observations, it comes at the expense of feasibility, simplicity, and interpretability, and will likely remain an aspirational goal.

10 https://glottography.github.io/tutorials/ (Accessed: 2026-02-19).

### 5.2 CONCLUSION

*Glottography* is a free and open repository providing digital polygons for the world's languages. Currently, it contains over 13,000 language polygons representing more than 5,300 unique languages. The coverage of *Glottography* largely matches that of the only comparable service, the subscription-based World Mapping System offered by Ethnologue. All polygons are available in the Cross-Linguistic Data Format on Zenodo, ready for use in Geographic Information Systems or any spatial programming environment, e.g. through the *Rglottography* package in R. Detailed tutorials encourage community members to contribute their own language polygons to the platform, expanding coverage to previously uncharted regions, languages, and historical epochs.

## ACKNOWLEDGEMENTS

The authors wish to thank Thiago Chacon for recommending additional source publications for South America, and Odri Klaussova and Martijn Romar for their assistance in digitising language maps.

## FUNDING STATEMENT

PR, AH, NEK and MF were funded by the URPP 'Language and Space', University of Zurich. PR and MF were partially funded by the NCCR Evolving Language, Swiss NSF Agreement No. 51NF40_180888. GK and TT were funded by the project 'Out of Asia', Swiss NSF agreement No. CRSII5_183578. MU and JG were funded by the European Union (ERC, LANGUAGE REDUX, 101124345). RVG was funded by the European Union (ERC, SAPPHIRE, 818854) and the Dutch Scientific Organization (NWO Open Competition-L, Disentangling the roles of social and biophysical factors in the evolution of linguistic diversity in South America). MR was funded by the Finnish Society of Sciences and Letters (grant no. 87), the Finnish Cultural Foundation (grant no. 00220881), the Human Diversity consortium (HuDi) under the Profi7 programme of the Research Council of Finland (grant no. 352727).

## COMPETING INTERESTS

The authors have no competing interests to declare.

## AUTHOR CONTRIBUTIONS

Conceptualization: PR, SN, GK; Supervision: PR, GK, SN, RW, OV, RVG; Methodology: PR, GK, SN, AH, AK, MR, GB, NEK, MF, RW; Investigation & data curation: PR, GK, SN, AH, AK, MR, GB, NEK, JG, MF, RVG; Visualisation: PR, MF; Validation: RF, MU, PR, AH, RVG, AHR, MA, SN, MR, MF, NEK, NN, TT; Writing - original draft: PR, RF, SN, MR, OV; Writing - review & editing: PR, RF & MF: Software: RF, PR, AH, NN, MF; Funding acquisition: RW, RVG.

## AUTHOR AFFILIATIONS

**Peter Ranacher** orcid.org/0000-0002-8680-4063
University Research Priority Program (URPP) 'Language and Space', University of Zurich, Zurich, Switzerland; Department of Geography, University of Zurich, Zurich, Switzerland

**Robert Forkel** orcid.org/0000-0003-1081-086X
Department of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

**Nour Efrat-Kowalsky** orcid.org/0000-0002-5929-4897
University Research Priority Program (URPP) 'Language and Space', University of Zurich, Zurich, Switzerland; Department of Geography, University of Zurich, Zurich, Switzerland

**Matthias Urban** orcid.org/0000-0001-7633-7433
Laboratoire "Dynamique du Langage", UMR 5596, CNRS & Université Lumière Lyon 2, Lyon, France

**Antonia Hehli** orcid.org/0009-0008-3972-4280
University Research Priority Program (URPP) 'Language and Space', University of Zurich, Zurich, Switzerland; Department of Geography, University of Zurich, Zurich, Switzerland

**Micha Franz** orcid.org/0009-0005-5114-5087
University Research Priority Program (URPP) 'Language and Space', University of Zurich, Zurich, Switzerland; Department of Geography, University of Zurich, Zurich, Switzerland

**Gregory Biland** orcid.org/0009-0001-3753-9945
University Research Priority Program (URPP) 'Language and Space', University of Zurich, Zurich, Switzerland; Department of Geography, University of Zurich, Zurich, Switzerland

**Aaron Kreienbühl**
University Research Priority Program (URPP) 'Language and Space', University of Zurich, Zurich, Switzerland; Department of Geography, University of Zurich, Zurich, Switzerland

**Alba Hermida Rodríguez** orcid.org/0000-0002-8038-9866
Faculty of Arts and Philosophy, Ghent University, Ghent, Belgium

**Matheus C. B. C. Azevedo** orcid.org/0009-0002-4489-7174
Department of Modern Languages and Cultures, Radboud University, Nijmegen, The Netherlands

**James Giebler** orcid.org/0009-0005-3040-4617
Laboratoire "Dynamique du Langage", UMR 5596, CNRS & Université Lumière Lyon 2, Lyon, France

**Takuya Takahashi** orcid.org/0000-0002-6813-8212
Department of Geography, University of Zurich, Zurich, Switzerland

**Nico Neureiter** orcid.org/0000-0002-3719-2259
Department of Geography, University of Zurich, Zurich, Switzerland; University Research Priority Program (URPP) 'Language and Space', University of Zurich, Zurich, Switzerland

**Rik van Gijn** orcid.org/0000-0001-9911-2907
Leiden University Centre for Linguistics, Leiden University, Leiden, The Netherlands

**Meeli Roose** orcid.org/0000-0003-2776-9883
Department of Geography and Geology, University of Turku, Turku, Finland

**Outi Vesakoski** orcid.org/0000-0002-7220-3347
School of Languages and Translation Studies, University of Turku, Finland

**Robert Weibel** orcid.org/0000-0002-2425-0077
Department of Geography, University of Zurich, Zurich, Switzerland; University Research Priority Program (URPP) 'Language and Space', University of Zurich, Zurich, Switzerland

**Gereon Kaiping** orcid.org/0000-0002-8155-9089
Department of Geography, University of Zurich, Zurich, Switzerland

**Sietze Norder** orcid.org/0000-0003-4692-4543
Copernicus Institute of Sustainable Development, Faculty of Geosciences, Utrecht University, Utrecht, The Netherlands

## REFERENCES

**AIATSIS.** (1996). *Map of Indigenous Australia.* https://aiatsis.gov.au/explore/map-indigenous-australia. (Accessed: 2025-09-05).

**Alaskool.** (1998). *Alaska Native Languages.* http://www.alaskool.org/language/languageindex.htm. (Accessed: 2025-09-05).

**Allen, M. W., Bettinger, R. L., Codding, B. F., Jones, T. L.,** & **Schwitalla, A. W.** (2016). Resource scarcity drives lethal aggression among prehistoric hunter-gatherers in central California. *Proceedings of the National Academy of Sciences*, *113*(43), 12120–12125. https://doi.org/10.1073/pnas.1607996113

**Asher, R. E.,** & **Moseley, C. J.** (Eds.) (2007). *Atlas of the World's Languages* (2nd ed.). Abingdon: Routledge.

**Bouckaert, R., Lemey, P., Dunn, M., Greenhill, S. J., Alekseyenko, A. V., Drummond, A. J., … Atkinson, Q. D.** (2012). Mapping the origins and expansion of the Indo-European language family. *Science*, *337*(6097), 957–960. https://doi.org/10.1126/science.1219669

**Bowern, C.** (2021). *Files for Australian Language Locations.* Zenodo. (Dataset) https://doi.org/10.5281/zenodo.4898185

**Bowern, C.,** & **Atkinson, Q.** (2012). Computational phylogenetics and the internal structure of Pama-Nyungan. *Language*, *88*, 817–845. https://doi.org/10.1353/lan.2012.0081

**Carling, G.,** & **Gippert, J.** (Eds.) (2025). *DiACL/TITUS Polygon Archive.* Frankfurt am Main: Goethe University. Retrieved from https://diacl.uni-frankfurt.de/GeographicalPresence/Index (Accessed: 2026-02-19).

**Collard, I. F.,** & **Foley, R. A.** (2002). Latitudinal patterns and environmental determinants of recent human cultural diversity: do humans follow biogeographical rules? *Evolutionary ecology research*, *4*(3), 371–383.

**Dahl, Ö.,** & **Veselinova, L.** (2005). Language Map Server. In *Proceedings of the 25th annual esri international user conference.* San Diego, CA: Environmental Systems Research Institute (ESRI). Retrieved from http://proceedings.esri.com/library/userconf/proc05/papers/pap2425.pdf (Accessed: 2026-02-19).

**Dedio, S., Ranacher, P.,** & **Widmer, P.** (2019). Evidence for Britain and Ireland as a linguistic area. *Language*, *95*(3), 498–522. https://doi.org/10.1353/lan.2019.0054

**Denevan, W. M.** (1966). *The aboriginal cultural geography of the Llanos de Mojos of Bolivia* (No. 48). Berkeley: University of California Press. Retrieved from http://www.pueblos-originarios.ucb.edu.bo:4080/digital/106000618.pdf (Accessed: 2026-02-19).

**Eberhard, D. M., Simons, G. F.,** & **Fennig, C. D.** (Eds.) (2024). *Ethnologue: Languages of the World. Twenty-seventh edition.* Dallas, Texas: SIL International. Retrieved from https://www.ethnologue.com (Accessed: 2026-02-19).

**Edwards, O. D.** (2020). *Metathesis and unmetathesis in Amarasi*. Language Science Press. Retrieved from https://langsci-press.org/catalog/book/228 (Accessed: 2026-02-19).

**Eriksen, L.** (2011). *Nature and Culture in Prehistoric Amazonia: Using G.I.S. to reconstruct ancient ethnogenetic processes from archaeology, linguistics, geography, and ethnohistory* (Doctoral Thesis, Human Ecology). Retrieved from https://lup.lub.lu.se/search/files/3626162/1890749.pdf (Accessed: 2026-02-19).

**Figueira, R.** (1982). *Atlas total de la República Argentina* (E. Chiozza, Ed.). Buenos Aires: Centro Editor de América Latina.

**Forkel, R.,** & **Hammarström, H.** (2024). A revised digital edition of Wurm & Hattori's Language Atlas of the Pacific Area. *Scientific Data 2024 11:1*, *11*, 1–13. https://doi.org/10.1038/s41597-024-03816-w

**Forkel, R., List, J.-M., Greenhill, S. J., Rzymski, C., Bank, S., Cysouw, M., … Gray, R. D.** (2018). Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Scientific data*, *5*(1), 1–10. https://doi.org/10.1038/sdata.2018.205

**Gavin, M. C.,** & **Stepp, J. R.** (2014). Rapoport's rule revisited: geographical distributions of human languages. *PloS one*, *9*(9), e107623. https://doi.org/10.1371/journal.pone.0107623

**Goddard, I.** (1999). *Native Languages and Language Families of North America* (Revised and enlarged edition, with additions and corrections ed.). Lincoln: University of Nebraska Press. Retrieved from https://www.nebraskapress.unl.edu/nebraska/9780803292697/ (Accessed: 2026-02-19).

**Grierson, G. A.** (1903). *Linguistic Survey of India*. Calcutta: Office of the Superintendent of Government Printing. Retrieved from https://dsal.uchicago.edu/books/lsi/ (Accessed: 2026-02-19).

**Hammarström, H.,** & **Donohue, M.** (2014). Some Principles on the Use of Macro-Areas in Typological Comparison. *Language Dynamics and Change*, *4*(1), 167–187. https://doi.org/10.1163/22105832-00401001

**Hammarström, H., Forkel, R., Haspelmath, M.,** & **Bank, S.** (2025). *Glottolog 5.2.* Leipzig: Max Planck Institute for Evolutionary Anthropology. Retrieved from http://glottolog.org (Accessed: 2026-02-19).

**Haynie, H. J.,** & **Gavin, M. C.** (2019). *Modern Language Range Mapping for the Study of Language Diversity* (Tech. Rep.). SocArXiv. https://doi.org/10.31235/osf.io/9fu7g

**Hochstetler, J. L., Durieux, J. A.,** & **Durieux-Boon, E. I.** (2004). Sociolinguistic survey of the Dogon language area. *SIL International*.

**Mace, R.,** & **Pagel, M.** (1995). A latitudinal gradient in the density of human languages in North America. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *261*(1360), 117–121. https://doi.org/10.1098/rspb.1995.0125

**Mankoff, J.** (2022). Russia's war in Ukraine. *Identity, History and Conflict, Washington, DC: Centre for Strategic and International Studies*. Retrieved from https://www.csis.org/analysis/russias-war-ukraine-identity-history-and-conflict (Accessed: 2026-02-19).

**Matacic, C.** (2020). World's largest linguistics database is getting too expensive for some researchers. *Science*, *15*. https://doi.org/10.1126/science.abb2422

**Matsumae, H., Ranacher, P., Savage, P. E., Blasi, D. E., Currie, T. E., Koganebuchi, K., … Bickel, B.** (2021). Exploring correlations in genetic and cultural variation across language families in northeast Asia. *Science Advances*, *7*(34), eabd9223. https://doi.org/10.1126/sciadv.abd9223

**Messineo, C.** (2011). Aproximación tipológica a las lenguas indígenas del Gran Chaco. Rasgos compartidos entre toba (familia guaycurú) y maká (familia mataco-mataguayo). *Indiana*, *28*(2), 183–225. https://doi.org/10.18441/ind.v28i0.183-225

**Ministerio de Educación de Argentina.** (Ed.) (2009). *República Argentina – Pueblos Indígenas.* Retrieved from https://dia.upenn.edu/en/maps/ARG0005/ (Accessed: 2026-02-19).

**Native Land Digital.** (2025). *Native-Land.ca — Our home on native land.* https://native-land.ca/contact/. (Accessed: 2025-09-05).

**Natural Earth.** (2024). *Natural Earth: 1:10m Physical Vectors - Free vector and raster map data at 1:10m, 1:50m, and 1:110m scales.* Retrieved from https://www.naturalearthdata.com/downloads/10m-physical-vectors/ (Accessed: 2026-02-19).

**Norder, S., Becker, L., Skirgård, H., Arias, L., Witzlack-Makarevich, A.,** & **van Gijn, R.** (2022). glottospace: R package for language mapping and geospatial analysis of linguistic and cultural data. *Journal of Open Source Software*, *7*(77), 4303. https://doi.org/10.21105/joss.04303

**Perlin, R., Kaufman, D., Lampel, J., Daurio, M., Turin, M.,** & **Craig, S.** (2022). *Languages of New York City (digital version), map.* http://languagemap.nyc. New York: Endangered Language Alliance. (Accessed: 2025-09-05).

**QGIS Development Team.** (2025). *QGIS Geographic Information System* (Version 3.34 ed.). Retrieved from https://qgis.org (Accessed: 2026-02-19).

**Queixalos, F.,** & **Renault-Lescure, O.** (2000). *As línguas amazônicas hoje* (F. Queixalós & O. Renault-Lescure, Eds.). São Paulo: IRD/Instituto Socioambiental/MPEG. Retrieved from https://www.documentation.ird.fr/hor/fdi:010022973 (Accessed: 2026-02-19).

**Ranacher, P.** (2025). *guess_glottocode: Guess the Glottocode for a language using either a Wikipedia query or a large language model.* Retrieved from https://github.com/derpetermann/guess_glottocode (Accessed: 2026-02-19).

**Ranacher, P.** (2026a). *Glottography Tutorials.* Zenodo. Retrieved from https://glottography.github.io/tutorials/ (Accessed: 2026-02-19).

**Ranacher, P.** (2026b). *Rglottography.* Zenodo. Retrieved from https://github.com/Glottography/Rglottography (Accessed: 2026-02-19).

**Ranacher, P., Forkel, R., Efrat-Kowalsky, N., Urban, M., Hehli, A., Franz, M., … Norder, S.** (2025). A global and interoperable dataset of linguistic distributions derived from the Atlas of the World's Languages. *Scientific Data*, *12*(1), 1466. https://doi.org/10.1038/s41597-025-05828-6

**Rantanen, T., Tolvanen, H., Roose, M., Ylikoski, J.,** & **Vesakoski, O.** (2022). Best practices for spatial language data harmonization, sharing and map creation—A case study of Uralic. *Plos one*, *17*(6), e0269648. https://doi.org/10.1371/journal.pone.0269648

**Rantanen, T., Vesakoski, O., Ylikoski, J.,** & **Tolvanen, H.** (2021). *Geographical database of the Uralic languages (v1.0) [Data set].* Zenodo. https://doi.org/10.5281/zenodo.4784188

**Šavrič, B., Patterson, T.,** & **Jenny, B.** (2019). The Equal Earth map projection. *International Journal of Geographical Information Science*, *33*(3), 454–465. https://doi.org/10.1080/13658816.2018.1504949

**Schapper, A.** (2020). Introduction to The Papuan languages of Timor, Alor and Pantar. In A. Schapper (Ed.), *The papuan languages of timor, alor and pantar: Volume 3* (pp. 1–52). Berlin, Boston: De Gruyter Mouton. https://doi.org/10.1515/9781501511158-001

**Steever, S. B.** (2019). *The Dravidian Languages* (2nd ed.). London, England: Taylor & Francis. https://doi.org/10.4324/9781315722580

**Suttles, W.,** & **Suttles, C.** (1985). *Native Languages of the Northwest Coast.* Portland: The Press of the Oregon Historical Society.

**Takahashi, T., Hannes, G., Neureiter, N.,** & **Ranacher, P.** (2023). Inferring the History of Spatial Diffusion Processes. In R. Beecham, J. A. Long, D. Smith, Q. Zhao, & S. Wise (Eds.), *12th International Conference on Geographic Information Science (GIScience 2023)* (Vol. 277, pp. 71:1–71:6). Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik. https://doi.org/10.4230/LIPIcs.GIScience.2023.71

**Tarble de Scaramelli, K.,** & **Zucchi, A.** (1984). Nuevos Datos sobre la Arqueología Tardia del Orinoco: La Serie Valloide. *Acta Científica Venezolana*, *35*, 434–445.

**Toan, L.** (2024). Construction of personal identity through linguistic device: An anthropologial linguistics analysis. *Revista De Gestão Social E Ambiental*, *18*(7), e07139. https://doi.org/10.24857/rgsa.v18n7-122

**Vuillermet, M.** (2012). *A grammar of Ese Ejja, a Takanan language of the Bolivian Amazon* (Unpublished doctoral dissertation). Université Lumière Lyon 2.

**Walker, R. S.,** & **Ribeiro, L. A.** (2011). Bayesian phylogeography of the Arawak expansion in lowland South America. *Proceedings of the Royal Society B: Biological Sciences*, *278*(1718), 2562–2567. https://doi.org/10.1098/rspb.2010.2579

**Wikipedia contributors.** (2024). *List of official languages by country and territory — Wikipedia, the free encyclopedia.* Retrieved from https://en.wikipedia.org/w/index.php?title=List_of_official_languages_by_country_and_territory&oldid=1257169394 (Accessed: 2025-11-18).

**Wurm, S.,** & **Hattori, S.** (1981). *Language Atlas of the Pacific Area: New Guinea area, Oceania, Australia* (Vol. 66). Canberra: Research School of Pacific and Asian Studies, Australian National University.

**Zucchi, A.** (2017). *Arqueología de los llanos occidentales y el Orinoco*. Caracas: Centro Nacional de Estudios Históricos.

**TO CITE THIS ARTICLE:**
Ranacher, P., Forkel, R., Efrat-Kowalsky, N., Urban, M., Hehli, A., Franz, M., Biland, G., Kreienbühl, A., Rodríguez, A. H., Azevedo, M. C. B. C., Giebler, J., Takahashi, T., Neureiter, N., van Gijn, R., Roose, M., Vesakoski, O., Weibel, R., Kaiping, G., & Norder, S. (2026). Glottography: An Open-Source Geolinguistic Data Platform for Mapping the World's Languages. *Journal of Open Humanities Data*, 12: 47, pp. 1–16. DOI: https://doi.org/10.5334/johd.459

**Submitted:** 06 November 2025
**Accepted:** 04 February 2026
**Published:** 19 March 2026



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