Правительство Российской Федерации
Федеральное государственное автономное образовательное учреждение высшего образования

Национальный исследовательский университет
«Высшая школа экономики»

Факультет гуманитарных наук
Образовательная программа
«Фундаментальная и компьютерная лингвистика»

Воронов Михаил Кириллович

Программная библиотека для лингвистической типологии на языке Python

Python Library for Linguistic Typology

Выпускная квалификационная работа студента 4 курса бакалавриата

Академический руководитель образовательной программы	Научный руководитель
канд. филологических наук, доц.	канд. филологических наук, доц.
Ю.А.Ландер	Б.Орехов

4 июня 2019 г.

Москва 2019

Contents

1 Introduction

There are multiple linguistic tools for Python. Most of them concentrate on natural language processing (e.g. NLTK (Bird, Steven, Edward Loper and Ewan Klein 2009)). There are also some libraries that are there to assist linguistic research.

A good example of such package is LingPy (List, Greenhill, and Forkel 2017). It provides multiple calculation and visualisation algorithms for historical linguistics.

Another example of this kind of libraries is LingCorpora (Koshevoy et al. 2018). It allows to perform queries in multiple online text corpora. It is in active development at the moment and already supports more than 25 corpora.

However, I did not find any Python tools that allow to work with online linguistic databases.

Most of such databases are stored in Cross-Linguistic Linked Data format (Haspelmath and Forkel 2013). This specification also provides framework (Forkel et al. 2019) that allows creating CLLD apps. However, it does not provide user-friendly API for the stored data and cannot access databases from remote repositories.

There is a tool for Glottolog (Hammarström, Forkel, and Haspelmath 2019) that provides an API and console application for Glottolog (Forkel 2019). However, this tool requires a local copy of Glottolog data that takes more than 700 megabytes of storage.

Also, there seems to be no Python tools for linguistic interactive mapping and researchers have to use libraries such as Folium (Filipe et al. 2019b) which is a general tool for interactive mapping and is not designed for linguistic maps specifically.

So, the first gap that my package attempts to cover is lack of Python tools that provide an interface for online linguistic databases. The second gap is lack of Python tools designed for linguistic interactive mapping.

There is a package for the R programming language called 'lingtypology' (Moroz 2017). It provedes an API for linguistic databases, a tool to work with Glottolog data and a tool to create interactive linguistic maps. My package was inspired by this R library and I consider it to be its counterpart for Python. Therefore my package is also called 'lingtypology'.

One of the main purposes of my package is to provide a tool for easy reproducibility. Usually researchers manually find the data from multiple linguistic research. With my package it is possible to reproduce such research very quickly.

In the following chapters, I will describe technical aspects of 'lingtypology', provide documentation and several small studies to demonstrate its usability.

2 Project Description

In this section I will provide description for `lingtypology`. To avoid repetition I will not include documentation into this section, it is present in Chapter 3 'Usage'.

2.1 General

`lingtypology` is written in the Python programming language version 3.7 (Python Software Foundation 2019). It also supports versions 3.5 and 3.6. It is planned to keep maintaining all the versions of Python that are supported except for 2.7 branch. It is not supported due to its coming end of life in early 2020.

This project uses `git` distributed version control system. The source code of the project is stored in the remote repository (Voronov 2019b).

2.2 Dependencies

`lingtypology` package requires a number of additional libraries.

- Folium and Branca (Filipe et al. 2019b). Folium is a Python wrapper for `leaflet` library for JavaScript (Agafonkin 2017). Branca is the additional package for Folium that allows editing HTML code of the maps while Folium works with JavaScript only.
- Pandas (Augspurger et al. 2019). Pandas introduces dataframes in Python.
- pyglottolog (Forkel 2019). pyglottolog application is used to extract necessary data from Glottolog.
- matplotlib (Caswell et al. 2019). Matplotlib is a tool for creating plots.
- jinja2. A template engine.
- colour. A library for colors.

2.3 Package

`lingtypology` package consists of different modules and data files.

2.3.1 Modules

- `__init.py__` contains imports and version.
- `maps.py` is the module for linguistic interactive mapping.
- `glottolog.py` contains a number of useful functions for Glottolog.
- `db_apis.py` contains API for multiple online databases.

2.3.2 Data Files

- `legend.html` HTML-template for map legends and title.
- `language_elevation_mapping.json` data on elevation for each language from Glottolog.
- `autotyp_lang_mapping.json` mapping from language IDs from Autotyp to languages from Glottolog. Taken from the R counterpart under the rule of GNU GPL license (Moroz 2017).
- CSV file that starts with `glottolog-languoids` contains some of Glottolog data. It is generated with `pyglottolog` application with `glottolog --repos=glottolog languoids` command.
- JSON file that starts with `glottolog-languoids` contains metadata for the CSV file above.

2.4 Interactive Maps

In this subsection I will describe `lingtypology.maps` module.

2.4.1 General

`lingtypology.maps` contains the `LingMap` object. This object has attributes and methods that allow to render interactive maps. This object stores the data that a user wants to be rendered and when `create_map` method is called all the data is processed and passed into `folium.map` object. Then it can be rendered as HTML.

For example, method `add_overlapping_features` applies different colors from `colors` atrribute to each feature, finds out the proper size of markers based on the amount of features for each language, creates `folium.CircleMarker` objects, adds popups and tooltips to the markers if necessary and then adds the markers to the map.

Also, `lingtypology.maps` contain several supplementary functions.

Full description of the functions and the `LingMap` class can be found in Chapter 3.

2.4.2 Elevation Data

`lingtypology` provides elevation data for each language in Glottolog (as of version 3.4). The source if this data is SRTM dataset (Jarvis A., H.I. Reuter, A. Nelson, E. Guevara 2008). It was processed using locally run Open Elevation API server (Lourenço and Developer66 2019).

2.4.3 Legend

Folium does not provide API to create a legend. Nevertheless, Lingtypology has a legend. It is created automatically based on the data that user passes to lingtypology.

The templated for the legend is based on the HTML code from here (Talbert 2018).

2.4.4 Strokes

The default appearance of standard `folium.CircleMarker` objects does not fit the needs of the package due to the wrong behaviour of strokes. Compare pictures 1 and 2.

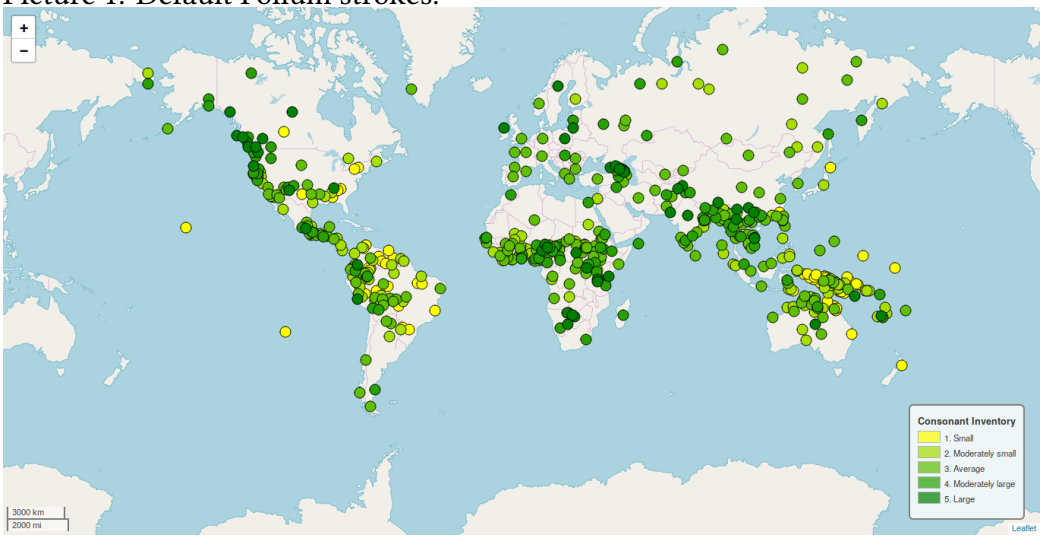
In the first picture strokes appear after each individual marker (the default behaviour of `folium`). In the second picture strokes appear after groups of markers. This bahaviour is impossible to achive using `folium.CircleMarker` objects by default. Therefore, instead of rendering one marker with stroke, `lingtypology` renders two markers: a marker without stroke and a black marker that 115% larger. Markers are not added to the map right away. They are rendered in the proper order:

1. Draw all black background markers.
2. Draw other markers.

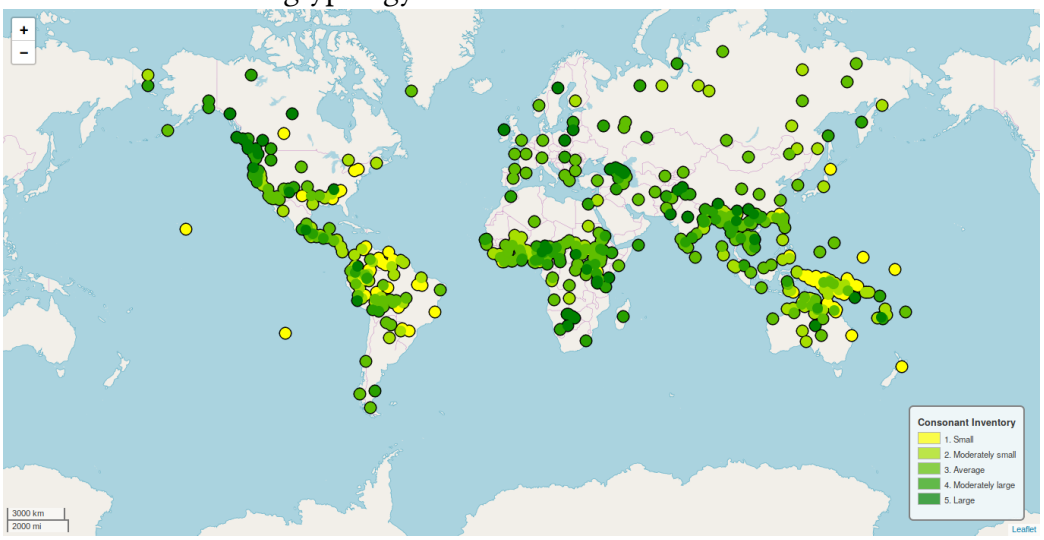
The way markers are rendered in the second picture is the default behaviour of `lingtypology`. However, the option to use the standard behaviour of `foilum` is present as well.

Both pictures were generated using `lingtypology`. Listings are in the attachment ('Listing 2' and 'Listing 1').

Picture 1. Default Folium strokes.



Picture 2. Default Lingtypology strokes.



2.4.5 Minicharts

I had a feature request to add functionality to draw minicharts instead of markers in the map.

Due to the fact that neither `folium` nor `folium.plugins` does not provide such functionality and I did not find a Leaflet plugin with such functionality, minicharts are rendered as SVG with Matplotlib (Caswell et al. 2019).

Matplotilib does provide functionality to render plots as SVG images and save

them as plain text files. Nevertheless, it does not provide documentation for methods that allow to get SVG as Python type `str`. Therefore, the plots rendered as SVG have to be caught with `io.StringIO` object from standard Python library.

Object `folium.Marker` allows passing HTML elements that define their appearance. However, it is not possible to pass the SVG as is: if there are newline symbols, the HTML file of the map will be broken. Therefore, all such symbols have to be removed from the SVG string.

Due to lack of unified API for different charts in Matplotlib, the number of available minicharts is limited. At the moment only pie-charts and bar-charts are supported.

During this research I received a request from one of the Folium developers to add an example of minicharts usage to the Folium gallery. My example is available in the ‘examples‘ directory of the Folium repository (Filipe et al. 2019b).

2.4.6 PNG

Foium does not provide official support for rendering maps as PNG. Nevertheless, I found undocumented API that allows it (`folium.Map._to_png` method).

It was implemented into Lingtypology. However, this method requires additional application (Geckodriver) which is not included into main repositories of popular operating systems (Windows, macOS, Debian GNU/Linux, OpenSUSE etc.) and requires additional efforts to install. Due to this fact, this functionality of Lingtypology is marked as experimental until another way to implement is found. Also, this functionality requires additional Python dependency: selenium. It is used to render HTML as PNG.

2.5 Glottolog

Lingtypology uses Glottolog data at its core. Therefore, accessing it online each time would significantly slow down the package. Glottolog data necessary for the package is included in the package.

However, Glottolog data is updated continually. I update included Glottolog data with each new release of Lingtypology. If user requires newer version of the data before new release, I provide the instruction how to update it manually.

`lingtypology.glottolog` provides multiple functions to work with Glottolog data. Usage information may be found in Chapter 3.

2.6 Databases API

Lingtypology provides API for the following lingustic databases:

- WALS (Dryer and Haspelmath 2013).
- Autotyp (Bickel et al. 2017).
- AfBo (Seifart 2013).
- SAILS (Muysken et al. 2016).
- PHOIBLE (Moran and McCloy 2019)

2.6.1 General

Databases usually can be retrieved in CSV format. They are read into `pandas.DataFrame`, processed and returned to user in easy to read and use format.

`lingtypology.db_apis` contains classes for each database. The module attemts to provide a unified API for all datasets, so there are methods that work the same way for all the databases.

The moment when the database is downloaded depends on the database. In some cases the data is received right after the initialization of the class, in other cases the data is downloaded when `get_df` method is called.

If the data in the database is divided into different pages, tables etc., Lingtypology is able to process and merge several of them.

2.6.2 WALS

WALS: 'The World Atlas of Language Structures (WALS) is a large database of structural (phonological, grammatical, lexical) properties of languages gathered from descriptive materials (such as reference grammars) by a team of 55 authors.' (Dryer and Haspelmath 2013). The data from wals is retrieved from multiple web-pages that contain data for each chapter when `get_df` method is called.

2.6.3 Autotyp

Autotyp is database that contains of multiple modules. Each module represents a grammatical feature (e.g. Agreeement), it contains information on this feature for various languages (Bickel et al. 2017). The data is downloaded when `get_df` method is called.

2.6.4 AfBo

AfBo: A world-wide survey of affix borrowing (Seifart 2013). AfBo contains information about borrewed affixes in different languages. It provides data in ZIP archive with CSV files. The data is downloaded with initialization of the class.

2.6.5 SAILS

'The South American Indigenous Language Structures (SAILS) is a large database of grammatical properties of languages gathered from descriptive materials (such as reference grammars)' (Muysken et al. 2016). Like in the case of AfBo, SAILS data is available in ZIP archive. The data is downloaded with initialization of the class.

2.6.6 PHOIBLE

'PHOIBLE is a repository of cross-linguistic phonological inventory data, which have been extracted from source documents and tertiary databases and compiled into a single searchable convenience sample.' (Moran and McCloy 2019). Unlike other databases supported by Lingtypology, PHOIBLE is not a unified dataset. It contains data of the following datasets: UPSID, SPA, AA, PH, GM, RA, SAPHON.

3 Usage

In this part I will provide full guide for 'lingtypology'. It is divided into four parts: Installation, Interactive Maps, Glottolog functions and the API for databases.

3.1 Installation

The package is uploaded to PyPI software repository, therefore it can be installed with the `pip` utility with the following command:

```
pip3 install lingtypology --user
```

3.2 Interactive Maps

Interactive maps can be created with `lingtypology.maps` module.

3.2.1 LingMap Class

class lingtypology.**LingMap** *(languages=[], glottocode=False)*
Bases: *object*
Parameters:

- **languages**: *list* or *pandas.Series* of strings, default *[]*.

  A list of languages. The language names should correspond to their names from Glottolog unless you use *add_custom_coordinates* method.

  Instead of language names you could use Glottocodes (language ID in Glottolog). In this case you need to set *glottocode* parameter to true.

- **glottocode**: *bool*, default *False*.

  Whether to treat *languages* as Glottocodes.

Attributes:

- **tiles**: *str*, default *'OpenStreetMap'*

  Tiles for the map. You can use one of these tiles (list of tiles is borrowed from the Folium Documentation (Filipe et al. 2019a)):

  - "OpenStreetMap"
  - "Mapbox Bright" (Limited levels of zoom for free tiles)
  - "Mapbox Control Room" (Limited levels of zoom for free tiles)
  - "Stamen" (Terrain, Toner, and Watercolor)
  - "Cloudmade" (Must pass API key)

  - "Mapbox" (Must pass API key)
  - "CartoDB" (positron and dark_matter)
  - or pass the custum URL.

- **start_location**: *(float, float)* or *str*, default *(0, 0)*

  Coordinates of the start location for the map *(latitude, longitude)* or a text shortcut. List of available shortcuts: *"Central Europe", "Caucasus", "Australia & Oceania", "Papua New Guinea", "Africa", "Asia", "North America", "Central America", "South America"*.

- **start_zoom**: *int*, default *2*

  Initial zoom level. Bypassed if you are using a shortcut *start_location*.

- **control_scale**: *bool*, default *True*

  Whether to add control scale.

- **prefer_canvas**: *bool*, default *False*

  Use canvas instead of SVG. If set to True, the map may be more responsive in case you have a lot of markers.

- **base_map**: *folium.Map*, default *None*

  In case you want to draw something on particular *folium.Map*.

- **title**: *str*, default *None*

  You can add a title to the map.

- **legend**: *bool*, default *True*

  Whether to add legend for features (*add_features* method).

- **stroke_legend**: *bool*, default *True*.

  Whether to add legend for stroke features (*add_stroke_features* method)

- **legend_title**: *str*, default *'Legend'*

  Legend title.

- **stroke_legend_title**: *str*, default *'Legend'*

  Stroke legend title.

- **legend_position**: *str*, default *'bottomright'*

  Legend position. Available values: 'right', 'left', 'top', 'bottom', 'bottomright', 'bottomleft', 'topright', 'topleft'.

- **stroke_legend_position**: *str*, default *'bottomleft'*

  Stroke legend position. Available values: 'right', 'left', 'top', 'bottom', 'bottomright', 'bottomleft', 'topright', 'topleft'.

- **colors**: *list* of html codes for colors (*str*).

  Colors that represent features. You can either use the 20 default colors or set yours.

- **stroke_colors**: *list* of html codes for colors (*str*)

  Colors that represent additional (stroke) features.

- **shapes**: *list* of characters (*str*)

  If you use shapes instead of colors, you can either use the default shapes or set yours. Shapes are Unicode symbols.

- **stroked**: *bool*, default *True*

  Whether to add stroke to markers.

- **unstroked**: *bool*, default *True*

  If set to True, circle marker will merge if you zoom out without stroke between them. It multiplies the number of markers by 2. For better performance set it to False. More information and examples in Chapter 2, Paragraph 4.4.

- **languages_in_popups**: *bool*, default *True*

  Whether to show links to Glottolog website in popups.

- **control**: *bool*, default *False*

  Whether to add LayerControls and group by features.

- **stroke_control**: *bool*, default *False*

  Whether to add LayerControls and group by stroke features.

- **control_position**: *str*, default *'topright'*

  Position of LayerControls. May be 'topleft', 'topright', 'bottomleft' or 'bottomright'.

- **colormap_colors**: *tuple*, default *('white', 'green')*

  Colors for the colormap.

Methods:

- **add_custom_coordinates** *(custom_coordinates)*

  Set custom coordinates. By default cooordinates for the languages are taken from the Glottolog database. If you have coordinates and want to use them, use this function.

  Parameter custom_coordinates: *list* of custom_coordinates (*tuples*)

  Length of the list should equal to length of languages.

- **add_features***(features, radius=7, opacity=1, numeric=False, control=False, use_shapes=False)*

  Add features to the map.

  Parameters:

  - **features**: *list*

    List of features. Amount of features should be equal to the amount of languages. By default, if you add features, a legend will appear. To shut it down set *legend* attribute to *False*. To change the title of the legend use *legend*$_t$*itle* attribute. To change legend position use use *legend_position* attribute.

  - **radius**: *int*, default *7*

    Marker radius.

  - **numeric**: *bool*, default *False*

    Whether to assign different color to each feature (*False*), or to assign a color from colormap (*True*). You can set it to True only in case your features are numeric and stroke features are not given. To change the default colors of the color scale use *colormap_colors* attribute.

  - **control**: *bool*, default *False*

    Whether to add LayerControls to the map. It allows interactive turning on/off given features.

  - **use_shapes**: *bool*, default *False*

    Whether to use shapes instead of colors. This option allows to represent features as shapes. Shapes are Unicode charaters like ⊠ or ⊠. You can replace or add to default symbols by changing *shapes* attribute. If colors are not a viable option for you, you can set this option to True.

- **add_stroke_features***(features, radius=12, opacity=1, numeric=False, control=False)*:

  Add additional set of features that look like strokes around markers.

- **features**: *list*

  List of additional features. Amount of features should be equal to the amount of languages. By default, if you add stroke features, a legend will appear. To shut it down set *stroke_legend* attribute to *False*. To change the title of the legend use *stroke_legend_title* attribute. To change legend position use use *stroke_legend_position* attribute.

- **radius**: *int*, default *12*

  Marker radius. Note that this radius is absolute as well.

- **control**: *bool*, default *False*

  Whether to add LayerControls to the map. It allows interactive turning on/off given stroke features.

- **add_overlapping_features** *(features, radius=7, radius_increment=4, mapping=None)*:

  Add overlapping features. For example, if you want to draw on map whether language 'is ergative', 'is slavic', 'is spoken in Russia'. It will draw several markers of different size for each feature.

  Parameters:

  - **features**: *list* of lists

    List of features. Amount of features should be equal to the amount of languages.

  - **radius**: *int*, default 7

    Radius of the smallest circle.

  - **radius_increment**: *int*, default *4*

    Step by which the size of the marker for each feature will be incremented.

  - **mapping**: *dict*, default *None*

    Mapping for the legend.

- **add_minicharts** *(*minicharts, typ='pie', size=0.6, names=None, textprops=None, labels=False, colors=[], startangle=90)*:

  Create minicharts using Maplotlib.

  Parameters:

  - ***minicharts**: list-like objects

    Data for minicharts. Two list-like objects.

- **typ**: *str*, default *pie*
  Type of the minicharts. Either *pie* or *bar*.
- **size**: *float*
  Size of the minicharts.
- **texprops**: *dict*, default *None*
  Textprops for Matplotlib.
- **labels**: *bool*, default *False*
  Whether to display labels.
- **colors**: *list*, default *[]*
  Minicharts colors.
- **startange**: *int*, default *90*
  Start angle of pie-charts (pie-charts only).

- **add_heatmap**: *(heatmap=[])*

  Add heatmap.

  Parameter heatmap: list-like object with tuples

  Coordinates for the heatmap. To create a heatmap-only map, do not pass any languages.

- **add_popups** *(popups, parse_html=False)*

  Add popups to markers.

  Parameters:

  - **popups**: *list* of strings List of popups. Length of the list should equal to length of languages.
  - **parse_html**: *bool*, default *False* By default (*False*) you can add HTML elements. If you need to add full HTML pages to popups, you need to set the option to *True*.

- **add_tooltips** *(tooltips)*:

  Add tooltips to markers.

  Parameter tooltips: *list* of strings

  List of tooltips. Length of the list should equal to length of languages.

- **add_minimap** *(position='bottomleft', width=150, height=150, collapsed_width=25, collapsed_height=25, zoom_animation=True)*

  Add minimap.

  Parameters:

  - **position**: *str*, default *'bottomleft'*
  - **width** and **height**: *int*, default *150*
  - **collapsed_width** and **collapsed_height**: *int*, default *25*
  - **zoom_animation**: *bool*, default *True*
    You can disable zoom animation for better performance.

- **create_map** *()*

  Create the map.

  Returns folium.Map

- **render***()*

  Returns the HTML code as *str*

- **save** *(path)*

  Save the map as HTML.

  Parameter path, *str*

  Path to the file.

- **save_static** *(path=None)*

  Save the map as PNG. Experimental function. Requires additional Python package Selenium and additional application Geckodriver.

  If path is not given **returns** the PNG as *bytes*.

3.2.2 Functions

function lingtypology.**merge** *(*maps)*
Accepts *LingMap* objects and creates map of them.
Parameter *maps: *LingMap* objects.
Returns foilum.Map

function lingtypology.**get_elevations** *(languages)*
Get data on elevation for languages. More information in Chapter 2, Paragraph 4.2.
Parameter lagnauges: *list* of strings

Returns list

function lingtypology.**gradient** *(iterations, color1='white', color2='green')*
Creates color gradient of given length.
Returns list of HEX-colors

3.3 Glottolog

3.3.1 Functions

Glottolog module includes various functions to work with Glottolog data.

The only funcntion that accepts list-like objects and returns *list* is **get_affiliations**. Its **parameter** is language names, it **returns** the genealogical information for the given languages.

The **parameter** of all the other functions is *str* and they **return** *str*.

The following functions use language name as the **parameter** and **return** coordinates, Glottocode, macro area and ISO code respectively:

- lingtypology.glottolog.**get_coordinates**
- lingtypology.glottolog.**get_glot_id**
- lingtypology.glottolog.**get_macroarea**
- lingtypology.glottolog.**get_iso**

The following functions use Glottocode as the **parameter** and **return** coordinates, language name and ISO code respectively:

- lingtypology.glottolog.**get_coordinates_by_glot_id**
- lingtypology.glottolog.**get_by_glot_id**
- lingtypology.glottolog.**get_iso_by_glot_id**

The following functions use ISO code as the **parameter** and **return** language name and Glottocode respectively.

- lingtypology.glottolog.**get_by_iso**
- lingtypology.glottolog.**get_glot_id_by_iso**

3.3.2 Versions

Processed Glottolog data is stored statically in the package directory. It is updated with each new release of `lingtypology`.

The version of the Glottolog data which is currently used is stored in lingtypology.glottolog.**version** variable.

It is possible to use local Glottolog data. To do so, it is necessary to perform the following steps:

- Download the current version of the Glottolog data (Hammarström, Forkel, and Haspelmath 2019).
- Create directory `.lingtypology_data` in your home directory.
- Move `glottolog` to `.lingtypology_data`.
- Run the following command: `glottolog --repos=glottolog languoids`
- It will generate two small files (csv and json). Now you can delete everything except for these files from the directory.
- Lingtypology will automatically use the local data.

3.4 Databases API

3.4.1 General

One of the objectives of LingTypology is to provide a simple interface for linguistic databases. Therefore, classes used for acccessing them have unified API: most attributes and methods overlap among all of them. In this subsection I will describe this universal interface.

Attributes:

- **citation**: *str*

  Citation for the database.

- **show_citation**: *bool*, default *True*

  Whether to print the citation when *get_df* method is called.

- **features_list** or **subsets_list**: *list* of strings

  List of available features for all the databases except for *Phoible*. In the case of *Phoible* it is list of available subsets (UPSID, SPA etc.).

Methods:

- **get_df**

  In all cases **parameters** are optional. They depend on the particular class.

  In the case of *Wals* it has optional *str* parameter *join_how*: the way multiple WALS pages will be joined (either *'inner'* or *'outer'*). If the value is *'inner'*, the resulting table will only contain data for languages mentioned in all the given pages. Else, the resulting table will contain values mentioned in at least one of the pages. Default: *'inner'*

  In the case of *Autotyp* and *Phoible* it has optional *list* parameter *strip_na*. It is a list of columns. If this parameter is given, the rows where some values in the given columns are not present will be dropped. Default: *[]*.

  Returns the dataset as *pandas.DataFrame*.

- **get_json**

  It works the same way as *get_df* but it **returns** *dict* object where keys are headers of the table.

3.4.2 WALS

class **Wals** *(*features)*

Parameter features, *list*
List of WALS pages that will be present in the resulting table. E.g. *['1A']*

Additional attribite general_citation, *str*
The general citation for **all** the WALS pages.

3.4.3 Autotyp

class **Autotyp** *(*tables)*

Parameter *tables: *list* of strings
List of the Autoptyp tables that will be merged in the resulting table. E.g. *['gender']*

3.4.4 AfBo

class **AfBo** *(*features)*

Parameter *features: *list* of strings
List of Autotyp features that will be present in the resulting table. E.g. *['adjectivizer']*.

3.4.5 SAILS

class **Sails** *(*features)*

Parameter *features: *list* of strings
List of SAILS pages that will be included in the resulting table.

Additional attribute feature**s**_descriptions: *pandas.DataFrame*
Table that contain description for all the SAILS pages.

Additional method feature_descriptions *(*features)*
Returns table with description for each given **feature**.

3.4.6 PHOIBLE

class **Phoible** *(subset='all', aggregated=True)*

Parameters:

- **subset**: *str*, default *'all'*

  One of the PHOIBLE datasets or all of them.

- **aggregated**: *bool*, default *True*

  If is *True* only aggregated data (e.g. amount of consonants) will be included into the resulting table. The table will be small and easy to operate.

  Else, full PHOIBLE data will be included.

3.5 Examples

3.5.1 WALS Heatmap

Listing 1. WALS Heatmap Example.

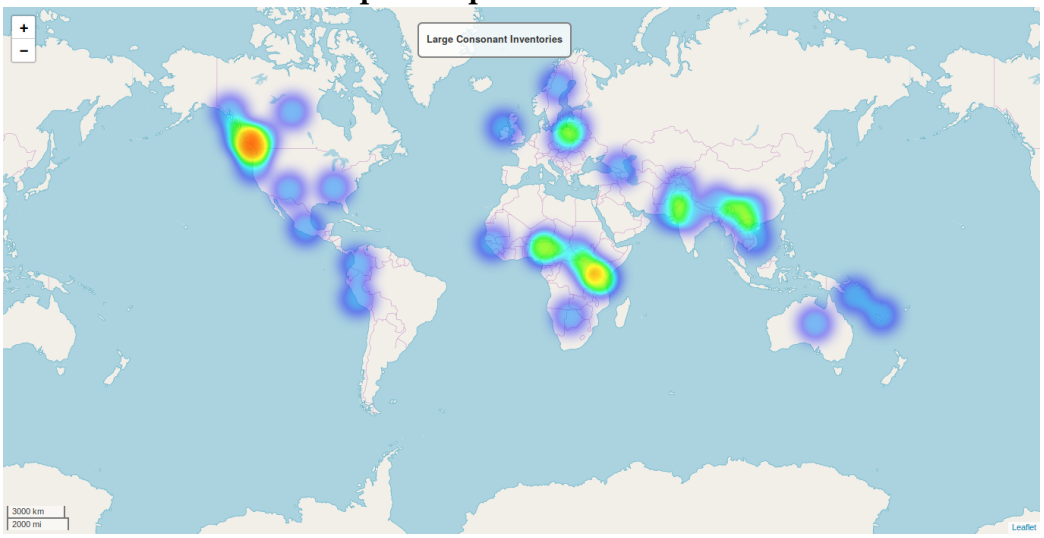
```
import lingtypology

#Get WALS page 1A
wals = lingtypology.db_apis.Wals('1A')
data = wals.get_df()
#Write dataframe to CSV
data.head().to_csv('tables/Wals1A.csv')

#First initialize LingMap without languages
m = lingtypology.LingMap()
#Add heatmap from  the Wals data where the inventory is large
m.add_heatmap(data[data._1A_desc == 'Large'].coordinates)
#Add title
```

```
m.title = 'Large Consonant Inventories'
#Save as PNG
m.save_static('images/WalsHeatmap.png')
```

Table 1. WALS Heatmap Example.

	wals_code	language	...	_1A	_1A_num	_1A_desc
0	kiw	Kiwai (Southern)	...	1. Small	1	Small
1	xoo	!Xóõ	...	5. Large	5	Large
2	ani	//Ani	...	5. Large	5	Large
3	abi	Abipón	...	2. Moderately small	2	Moderately small
4	abk	Abkhaz	...	5. Large	5	Large

Picture 3. WALS Heatmap Example.



3.5.2 Phoible Tones

Listing 2. Phoible Tones Example.

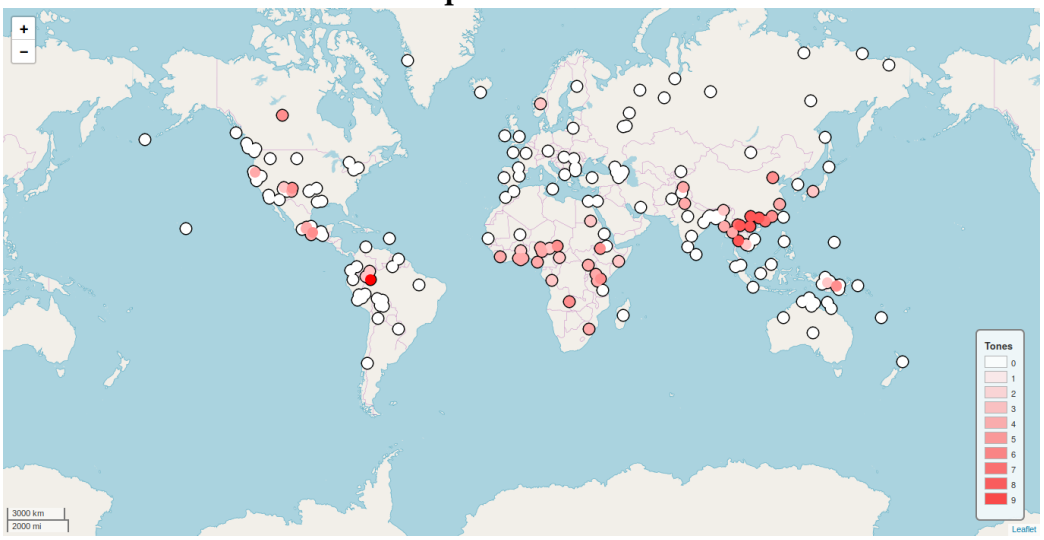
```
import lingtypology

#Get the table for UPSID dataset
p = lingtypology.db_apis.Phoible(subset='SPA')
df = p.get_df(strip_na=['tones'])
df.head().to_csv('tables/phoible.csv')

m = lingtypology.LingMap(df.language)
m.add_custom_coordinates(df.coordinates)
m.colormap_colors = ('white', 'red')
```

```
m.add_features(df.tones, numeric=True)
m.legend_title = 'Tones'
m.save_static('images/phoible.png')
```

Table 2. Phoible Tones Example.

	contribution_name	language	coordinates	...	tones	...
0	Korean (SPA 1)	Korean	(37.5, 128.0)	...	0.0	...
1	Ket (SPA 2)	Ket	(63.7551, 87.5466)	...	0.0	...
2	Lak (SPA 3)	Lak	(42.1328, 47.0809)	...	0.0	...
3	Kabardian (SPA 4)	Kabardian	(43.5082, 43.3918)	...	0.0	...
4	Georgian (SPA 5)	Georgian	(41.850396999999994, 43.78613)	...	0.0	...

Picture 4. Phoible Tones Example.



3.5.3 SAILS Example

Listing 3. SAILS Example.

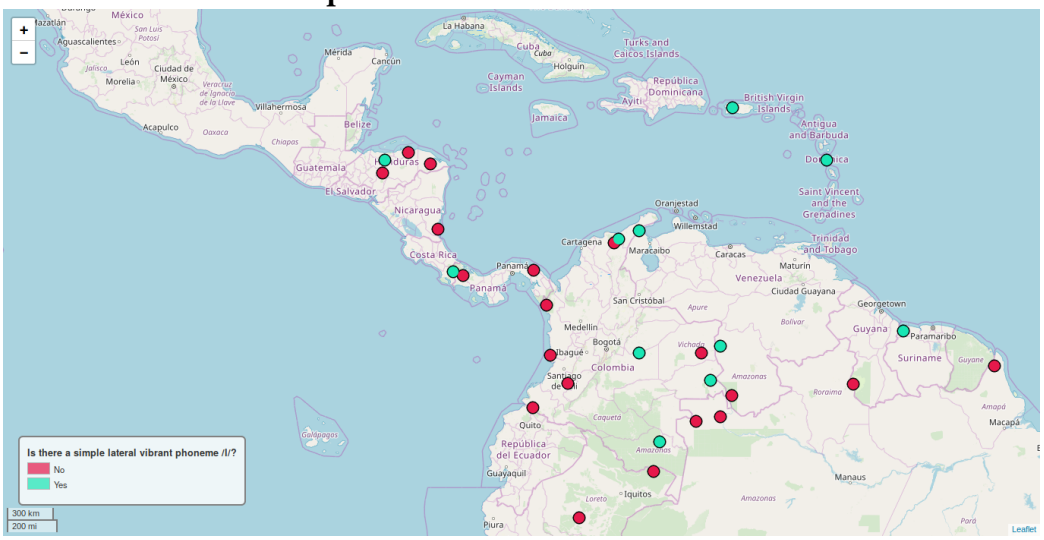
```
import lingtypology

#Get SAILS data for pages 'ICU3' and 'ICU4'
sails = lingtypology.db_apis.Sails('ICU3', 'ICU4')
df = sails.get_df()
df.head().to_csv('tables/sails.csv')

m = lingtypology.LingMap(df.language)
m.add_features(df.ICU3_desc)
#Use page description as legend title
```

```
m.legend_title = sails.feature_descriptions('ICU3').Description.at[0
m.start_location = (9, -79)
m.start_zoom = 5
m.legend_position = 'bottomleft'
m.save_static('images/sails.png')
```

Table 3. SAILS Example.

	language	coordinates	ICU3	ICU3_desc	ICU4	ICU4_desc
0	Baniva	(5.26123, -67.56326999999999)	1	Yes	0	No
1	Apolista	(-14.83, -68.66)”	0	No	?	?
2	Yavitero	(2.800281, -68.08421899999999)	1	Yes	0	No
3	Resígaro	(-2.48139, -71.35778)	0	No	0	No
4	Tol	(14.66859, -87.03719)	0	No	0	No

Picture 5. SAILS Example.



4 Distribution of Languages with Ejective Consonants and Elevation

4.1 Introduction

There are certain studies that suggest that geography may have influence on phonetics. For example there is a study that shows correlation between climatic areas and sonority classes (Munroe, Fought, and Macaulay 2009).

Another example of such studies is article by Caleb Everett that suggests influence of elevation on ejective consonants (Everett 2013).

In this work the hypothesis that ejective consonants are more frequent in high elevation areas (higher than 1500 m) is proven for the data from the respective WALS chapter (Maddieson 2013).

In this section I check this hypothesis on PHOIBLE datasets.

4.2 Analysis

PHOIBLE database contains the following datasets:

- SAPHON: South American Phonological Inventory Database (Lev, Stark, and Chang 2012).
- AA: Alphabets of Africa (Chanard 2006).
- GM: ‘Christopher Green and Steven Moran extracted phonological inventories from secondary sources including grammars and phonological descriptions with the goal of attaining pan-Africa coverage‘ (Moran, McCloy, and Wright 2014).
- PH: ‘Christopher Green and Steven Moran extracted phonological inventories from secondary sources including grammars and phonological descriptions with the goal of attaining pan-Africa coverage‘ (Moran, McCloy, and Wright 2014).
- RA: Common Linguistic Features in Indian Languages: Phoentics (Ramaswami 1999).
- SPA: Stanford Phonology Archive (Crothers et al. 1979).
- UPSID: UCLA Phonological Segment Inventory Database (Maddieson and Precoda 1990).

In this study RA and AA are not use because they either do not contain languages with ejectives or do not provide such information.

For all the other datasets information on amount of ejectives was collected using Lingtypology.

For all the languages I used the Lingtypology function that returns elevation for a given language based on its coordinates from Glottolog.

To check the hypothesis that can be formulated like "is it true that languages higher than 1500m are more probably has ejectives", I calculated chi-square test of the distribution of languages with or without ejectives and higher or lower than 1500m.

To the hypothesis from the Everett's article I added another hypothesis: is it true that if the language is higher, the more ejectives there are. To check this hypothesis I calculated two linear regressions for each dataset: one was calculated on the whole dataset of languages, the other was calculated only for languages that has ejectives.

4.3 Results

Results for the PHOIBLE datasets are stored in Table 4. The firts row is the name of the dataset, the second row is the p-value of the linear regression for languages with ejectives, the third row is the p-value of the linear regression for all languages.

The code necessary to create this table is stored in the remote repository (Voronov 2019a).

Table 4. Ejectives. P-value Table.

	Dataset	Regression (with ejectives only)	Regression (all languages)	Chi2 Test
0	UPSID	0.950559282993466	0.000044964081592	0.000032921681908
1	SPA	0.475539733143422	0.000005592842023	0.000176784757431
2	PH	0.731523538203316	0.392451413030472	0.160190111324293
3	GM	0.038586492300174	0.000000000000000	0.000000000000000
4	SAPHON	0.018874875617294	0.000000005031926	0.000377241915218

There are two possible ways to treat these numbers. The simplest option is to treat the datasets as equal and take median p-values. In this case those would be 0.47554, 0.00001 and 0.00018 for the regressions and chi-square respectively.

The second option is to treat the datasets as diffirent and consider only worldwide datasets. In this case, only UPSID and SPA will be considered.

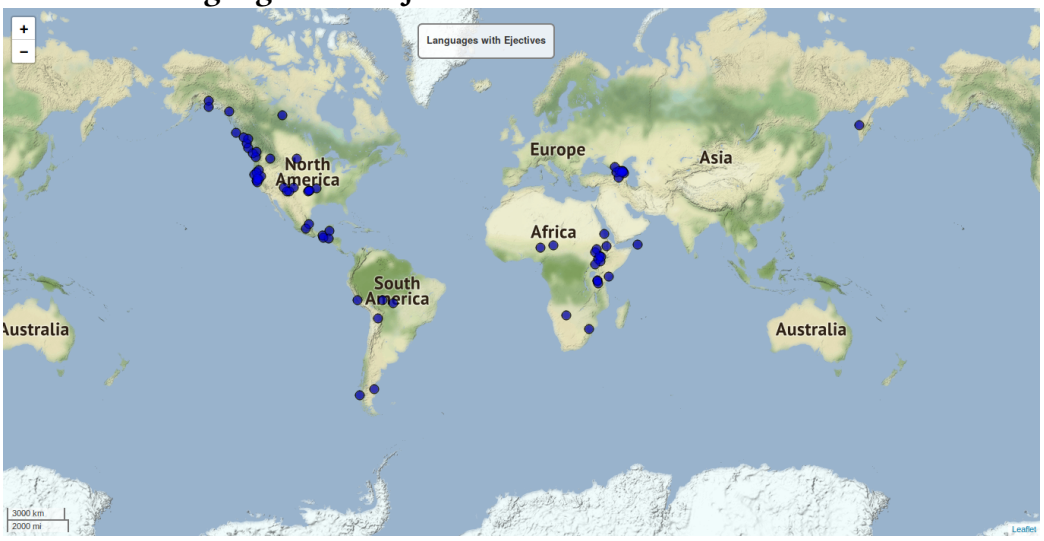
Nevertheless, in both options the result will be the same: the second regression and chi-square test show p-value < 0.05.

This means that it is indeed true that the share of languages with ejectives is higher if the elevation is more than 1500m. This fact causes the linear regression for all languages (including the ones with no ejectives) to be statically significant. Nevertheless, it is not true that that the higher the language, the more ejectives there are due to the fact that the regression for the languages that has ejectives does not show statistically significant result.

To demonstrate the results on the map, I provide map with languages that has ejectives (Picture 6) for UPSID dataset. It is noticable that these languages tend

to be in high elevation areas, now it is proven statistically for Phoible datasets The code used to generate it is in Appendix: Listing 3.

Picture 6. Languages with Ejectives.



4.4 Discussion

Of course, ”correlation does not imply causation”, and at the moment it is impossible to claim cause-and-effect relation for high elevation and presence of ejective consonants. Certain studies speculate why there may be cause-and-effect relations for geographical properties of regions and phonetics (e.g. (Everett 2013) and (Munroe, Fought, and Macaulay 2009). Nevertheless, it seems that further research is needed.

5 High Elevation: Quantitative Research

5.1 Introduction

Lingtypology allows accessing multiple features from linguistic databases. Therefore, I decided to find out whether high elevation may define distribution of other features from PHOIBLE database. Also, I decided to add morphosyntactic features from Aytotyp database.

5.2 PHOIBLE

To test this idea on PHOIBLE data I take all the phoneme properties amd test them as binary. So, in this part I ask questions like: "Is it true that if elevation is higher than 1500m, then the more likely it is to meet a language that has phonemes with the property?" I use chi-square test, then I find median p-value for all the datasets. The result is in Table 5. *NaN* means that there are no languages where at least on phoneme has the given property. *nan* means that there is not enough data to calculate chi-square.

Table 5. PHOIBLE All.

Dataset	short	long	delayedRelease	tap	trill	nasal
UPSID	0.7304	0.6205	0.6106	0.9272	0.5174	0.7388
SPA	0.4974	0.8311	0.4335	0.9873	0.9605	nan
GM	0.6587	0.0070	0.8435	0.8367	0.9499	0.1603
RA	0.0826	0.1125	nan	0.1125	0.0622	nan
AA	NaN	0.7559	nan	0.9076	0.4865	nan
PH	NaN	0.2549	0.9051	0.7908	0.1327	0.7573
SAPHON	NaN	0.0287	0.4856	0.3496	0.8520	0.7113
Median	0.578074	0.254949	0.610642	0.836724	0.517375	0.725022

Dataset	lateral	labial	round	labiodental	distributed	strident
UPSID	0.1174	nan	0.2667	0.8925	0.8872	0.5576
SPA	0.5463	0.3787	0.3787	0.1592	0.2771	0.7159
GM	0.6415	nan	0.1603	0.5869	0.4575	0.3861
RA	0.9301	nan	nan	0.9249	nan	0.3215
AA	0.0491	nan	nan	0.1428	0.8365	nan
PH	0.3205	nan	nan	0.8006	0.0753	0.4896
SAPHON	0.0000	nan	nan	0.8457	0.0139	0.3705
Median	0.320519	0.378695	0.266709	0.800579	0.367317	0.43784

Dataset	low	front	back	tense	retractedTongueRoot	advancedTongueRoot
UPSID	0.2667	nan	nan	0.2667	0.1243	NaN
SPA	0.3787	nan	nan	nan	0.8936	0.3787
GM	0.4430	0.1603	0.1603	0.1603	0.8242	NaN
RA	0.3215	nan	nan	nan	0.9301	NaN
AA	nan	nan	nan	nan	0.2252	NaN
PH	0.5906	nan	nan	0.2552	0.8665	0.2552
SAPHON	nan	nan	nan	nan	NaN	0.1864
Median	0.378695	0.160319	0.160319	0.255246	0.845344	0.255246

toprule Dataset	epilaryngealSource	spreadGlottis	constrictedGlottis	fortis	loweredLarynxImplosive	click
UPSID	NaN	0.3624	0.1280	NaN	0.5654	NaN
SPA	NaN	0.8858	0.1328	0.8083	0.8776	NaN
GM	0.1603	0.0480	0.0057	NaN	0.2245	0.1603
RA	NaN	0.8941	0.1244	NaN	0.3215	NaN
AA	NaN	0.1302	0.6491	NaN	0.5679	NaN
PH	0.2552	0.8090	0.1432	NaN	0.9455	NaN
SAPHON	NaN	0.0090	0.3423	NaN	0.6432	NaN
Median	0.207783	0.362376	0.132809	0.808315	0.567919	0.160319

5.3 Autotyp

Also, I checked this idea for numeric features from Autotyp. In this case linear regression was used.

Among the 30 numeric features there were 4 features that showed p-value < 0.05. These features are represented in Table 6. In this table the features are represented as abbreviations that are not always clear. Therefore I provide the list of descriptions for features from the Autotyp repository (Bickel et al. 2017):

- ‘Exponence: number of categories that are expressed in the same marker‘.
- ‘Rough approximation of the size of the possessum category in terms of the number of semantic classes covered‘.
- ‘Number of separately marked inflectional categories (including agreement) in position ”post” of the verb‘.
- ‘Number of morpheme types included in a phonologically or grammatically coherent suffix domain‘.

Table 6. Autotyp Features.

Feature	Subfeature	P-value
Grammatical_markers	Exponence.n	0.00000000
NP_structure	NPHeadSemClassSize.n	0.01766784
VInfl_counts_per_position	VInflCatAndAgrPost.n	0.02895302
Word_domains	MphmTypesInCohSuffixDomain.n	0.00196901

5.4 Discussion

So, in the case of PHOIBLE features the result is negative: no statistically significant differences were found in distribution of languages having phonemes with certain characteristics and high elevation.

In the case of Autotyp there were found several linear regressions with p-value < 0.05. However, I did not find any papers mentioning influence of geography on morphosynthax. Therefore, it is secure to say that futher research is needed.

6 WALS: Quantitative Research

7 References

Programming Tools

Bird, Steven, Edward Loper and Ewan Klein (2009). *Natural Language Processing with Python.*

Caswell, Thomas A et al. (May 2019). *matplotlib/matplotlib v3.1.0.* DOI: 10.5281/zenodo.2893252. URL: https://doi.org/10.5281/zenodo.2893252.

Filipe et al. (May 2019b). *python-visualization/folium: v0.9.1.* DOI: 10.5281/zenodo.3229045. URL: https://doi.org/10.5281/zenodo.3229045.

Forkel, Robert (Apr. 2019). *clld/pyglottolog: Glottolog API.* DOI: 10.5281/zenodo.2620250. URL: https://doi.org/10.5281/zenodo.2620250.

Forkel, Robert et al. (Apr. 2019). *clld/clld: clld - a toolkit for cross-linguistic databases.* DOI: 10.5281/zenodo.2635592. URL: https://doi.org/10.5281/zenodo.2635592.

Hammarström, Harald, Robert Forkel, and Martin Haspelmath (Apr. 2019). *clld/glottolog: Glottolog database 3.4.* DOI: 10.5281/zenodo.2620814. URL: https://doi.org/10.5281/zenodo.2620814.

Koshevoy, Alexey et al. (Feb. 2018). *lingcorpora/lingcorpora.py: initial release.* DOI: 10.5281/zenodo.1172457. URL: https://doi.org/10.5281/zenodo.1172457.

List, Johann-Mattis, Simon Greenhill, and Robert Forkel (2017). *LingPy. A Python library for quantitative tasks in historical linguistics.* Jena. DOI: https://doi.org/10.5281/zenodo.1065403. URL: http://lingpy.org.

Lourenço, João Ricardo and Developer66 (2019). *Open-Elevation - Remake.* URL: https://github.com/Developer66/open-elevation.

Moroz, George (2017). *lingtypology: easy mapping for Linguistic Typology.* DOI: 10.5281/zenodo.1289471. URL: https://CRAN.R-project.org/package=lingtypology.

Voronov, Michael (May 2019b). *lingtypology: a Python tool for linguistic interactive mapping.* DOI: 10.5281/zenodo.2669068. URL: https://doi.org/10.5281/zenodo.2669068.

Documentation

Agafonkin, Vladimir (2017). *Leaflet API reference.* URL: https://leafletjs.com/reference-1.5.0.html.

Augspurger, Tom et al. (2019). *pandas: powerful Python data analysis toolkit*. URL: http://pandas.pydata.org/pandas-docs/stable.
Filipe et al. (2019a). *Folium*. URL: https://python-visualization.github.io/folium/index.html.
Python Software Foundation (2019). *The Python Language Reference*. URL: https://docs.python.org/3.7/reference/.
Talbert, Colin (2018). *How does one add a legend (categorical) to a folium map*. URL: https://nbviewer.jupyter.org/gist/talbertc-usgs/18f8901fc98f109f2b71156cf3ac81cd.

Online Recources

Bickel, Balthasar et al. (2017). *The AUTOTYP typological databases. Version 0.1.0*. URL: https://github.com/autotyp/autotyp-data/tree/0.1.0.
Chanard, C. (2006). *Systèmes Alphabétiques Des Langues Africaines*. URL: http://sumale.vjf.cnrs.fr/phono/.
Dryer, Matthew S. and Martin Haspelmath, eds. (2013). *WALS Online*. URL: https://wals.info/.
Haspelmath, Martin and Robert Forkel (2013). *CLLD – Cross-Linguistic Linked Data*. URL: https://clld.org/.
Jarvis A., H.I. Reuter, A. Nelson, E. Guevara (2008). *Hole-filled seamless SRTM data V4*. URL: http://srtm.csi.cgiar.org.
Lev, Michael, Tammy Stark, and Will Chang (2012). *South American Phonological Inventory Database*. URL: http://linguistics.berkeley.edu/%20saphon/en/.
Moran, Steven and Daniel McCloy, eds. (2019). *PHOIBLE 2.0*. URL: https://phoible.org/.
Moran, Steven, Daniel McCloy, and Richard Wright, eds. (2014). *PHOIBLE Online*. URL: http://phoible.org/.
Muysken, Pieter et al. (2016). *South American Indigenous Language Structures (SAILS) Online*. URL: http://sails.clld.org.
Seifart, Frank, ed. (2013). *AfBo: A world-wide survey of affix borrowing*. URL: https://afbo.info/.
Voronov, Michael (2019a). *Data for Lingtypology Demonstrative Studies*. URL: https://github.com/OneAdder/lingtypology_research.

Articles

Everett, Caleb (2013). "Evidence for Direct Geographic Influences on Linguistic Sounds: The Case of Ejectives". In: DOI: 10.1371/journal.pone.

0065275. URL: https://doi.org/10.1371/journal.pone.0065275.
Maddieson, Ian (2013). "Glottalized Consonants". In: ed. by Matthew S. Dryer and Martin Haspelmath. URL: https://wals.info/chapter/7.
Munroe, Robert L., John G. Fought, and Ronald K. S. Macaulay (2009). "Warm Climates and Sonority Classes: Not Simply More Vowels and Fewer Consonants". In: *Cross-Cultural Research* 43.2, pp. 123–133. DOI: 10.1177/1069397109331485. eprint: https://doi.org/10.1177/1069397109331485. URL: https://doi.org/10.1177/1069397109331485.

Other

Maddieson, Ian and Kristin Precoda (1990). "Updating UPSID". In: *UCLA Working Papers in Phonetics*. Vol. 74. Department of Linguistics, UCLA, pp. 104–111.
Ramaswami, N. (1999). *Common Linguistic Features in Indian Languages: Phonetics*. Central Institute of Indian Languages.

8 Appendix

Listing 1. Wals 1A map with default Lingtypology strokes.

```
import os
os.chdir('images')
import lingtypology

wals_page = lingtypology.db_apis.Wals('1a', '2a').get_df()
m = lingtypology.LingMap(wals_page.language)
m.add_custom_coordinates(wals_page.coordinates)
m.add_features(wals_page._1A)
m.legend_title = 'Consonant Inventory'
m.colors = lingtypology.gradient(5, 'yellow', 'green')
m.save_static('LingtypologyStrokeAppearance.png')
```

Listing 2. Wals 1A map with default Folium strokes.

```
import os
os.chdir('images')
import lingtypology

wals_page = lingtypology.db_apis.Wals('1a', '2a').get_df()
m = lingtypology.LingMap(wals_page.language)
m.add_custom_coordinates(wals_page.coordinates)
m.add_features(wals_page._1A)
m.legend_title = 'Consonant Inventory'
m.colors = lingtypology.gradient(5, 'yellow', 'green')
m.unstroked = False
m.save_static('FoliumStrokeAppearance.png')
```

Listing 3. Languages with Ejectives.

```
import lingtypology

df = lingtypology.db_apis.Phoible(subset='UPSID', aggregated=False).g
#Get all languages with ejectives
df = df[df.raisedLarynxEjective == '+']
#Remove duplicates
df = df.drop_duplicates(subset='Glottocode')

m = lingtypology.LingMap(df.Glottocode, glottocode=True)
#Tiles with terrain
```

```
m.tiles = 'Stamen Terrain'
m.title = 'Languages with Ejectives'
m.radius = 5
m.opacity = 0.5
m.colors = ('blue',)
m.save_static('images/Picture6.png')
```