Languages of the World
— Process Book —

COM 480 - Data Visualization - Spring 2024

Christodoulos Kechris, Rafael Medina Morillas, Dimitrios Samakovlis

1 Introduction

There are estimated to be over 7,000 languages spoken worldwide. Studying language diversity worldwide is essential for understanding human communication's intricacies and societies' cultural richness. Each language is a unique window into its speakers' collective knowledge, history, and identity. By delving into the vast array of languages spoken globally, researchers can uncover invaluable insights into human cognition, social interaction, and the evolution of languages over time. Furthermore, investigating language diversity fosters appreciation and respect for different cultures, promoting inclusivity and cross-cultural understanding. Ultimately, research in this field enriches our understanding of humanity and provides practical benefits in education, translation, and global cooperation.

Our project explores the origins and characteristics of languages around the world and aim to provide meaningful visualizations that will attract the reader's attention.

1.1 Goals and target audience

The central idea of our project is to show conceptual similarities among widespread languages and their correlation to the geographical position of each language's origin place. It is fascinating to explore in depth how languages may be similar or different in terms of origins, syntax, grammar, phonetics, etc. Similarly, examining the correlation between languages' geographical location and their feature similarity may lead to interesting takeaways and can provide the ground for future research.

Our target audience is the general public. No specific background or academic knowledge is required to understand the presented visualizations. Any person interested in exploring language history and cross-similarities may enjoy visiting our website and playing around with the interactive graphs.

2 Tools and Resources

2.1 Data sources

Our primary dataset is the **World Atlas of Language Structures (WALS)**[1] from Kaggle. WALS is a large database of structural (phonological, grammatical and lexical) properties of languages gathered from descriptive materials (such as reference grammar) by a team of 55 authors. The atlas provides information on the location, linguistic affiliation, and basic typological features of a significant number of the world's languages. There are over 200 features examined for each language, enabling many opportunities for data analysis and visualization. WALS Online is a publication of the Max Planck Institute for Evolutionary Anthropology[2]. It is a separate publication, edited by Dryer, Matthew S. & Haspelmath, Martin (Leipzig: Max Planck Institute for Evolutionary Anthropology, 2013). Preprocessing of this dataset was needed to update or add the ISO 639 Set 1 language codes[3].

Our secondary dataset is the **Countries Info**[4]. This dataset allows us to map the official languages spoken in the world countries, since the WALS dataset focuses on the native languages that originated in the different regions. The Countries Info dataset was processed to add the ISO 3166-1 alpha-2 country codes[5], to allow data matching with the WALS dataset and the GeoJSON.

[1] https://www.kaggle.com/datasets/rtatman/world-atlas-of-language-structures
[2] http://www.eva.mpg.de/
[3] https://www.iso.org/obp/ui/en/#iso:std:iso:639:ed-2:v1:en
[4] https://www.kaggle.com/datasets/pragya1401/countries-info
[5] https://www.iso.org/iso-3166-country-codes.html

For the visualizations including a world map, we employ the countries map data from **Natural Earth**[6], at a scale of 1:50 million, in GeoJSON format.

2.2 Tools

We employed Python to perform the initial dataset exploration and prepare the selected datasets. Python was also necessary to perform data preprocessing and cleaning, and generate complementary CSV and JSON files needed for the Language genealogy and Feature Clusters sections.

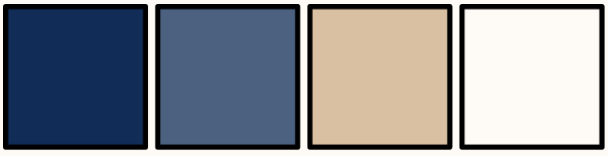
The design of the palette was inspired by the ColorHunt[7] catalogs, and we employed the Chroma.js Color Palette Helper[8] to generate colors coherent within the palette.

To host the website, we employed the Github pages feature.

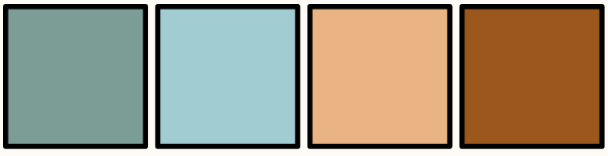
3 Website Design and Implementation

We designed the website as a collection of sections, each one exploring a specific linguistic phenomenon. Overall, our webpage is composed of six parts (Sections 3.2 - 3.7). The user navigates our visual linguistic exploration sequentially, starting from the introduction page and scrolling down until the final section. In each section, the interactive visualization is preceded by an introduction, providing context and inviting the user to immerse into our visualization.

We employed two color palettes for our webpage. The primary one was used throughout the whole page for the background, text and basic map visualizations. The second one was used selectively in visualizations needing high contrast between colored regions. The two color palettes are presented below.



Primary color palette used throughout our website.



Secondary color palette used in visualizations where high contrast was required.

3.1 Map visualization

Visualizing the world map is a core feature of our website, as it appears in all sections. Designing the world map in a modular and flexible way allowed its reusability and seamless integration across different sections.

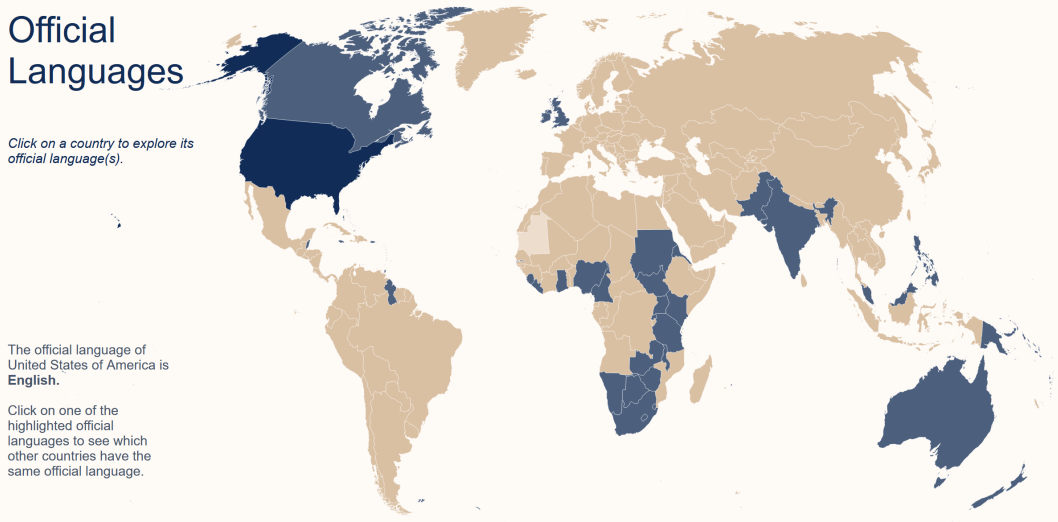
To implement a reusable D3.js map visualization, we developed the JavaScript closure *map()*. The closure is used to hold the configurable map parameters, and returns the function that draws the map accordingly. To leverage D3.js capabilities, the map binds the GeoJSON with the Countries Info dataset and WALS datasets. For the purposes of visualization, we chose the Winkel tripel projection to minimize area, direction and distance distortion. The configurable parameters are the following:

[6] https://www.naturalearthdata.com/
[7] https://colorhunt.co/
[8] https://gka.github.io/palettes/#/9|s||ffffe0,ff005e,93003a|1|1

- Map ID
- Parent SVG
- Position in the SVG
- Size
- Bound GeoJSON and CSVs
- Function to define the color map
- Function to define on-click behavior
- Functions to define on-hover behavior

3.2 Section 1 - Official languages

For the first section, we decided to visualize countries' official languages worldwide. An official language is a language that has certain usage rights in defined situations. Among the world countries, 178 recognize an official language, with 101 recognizing more than one. This visualization allows users to explore the range of languages spoken in different countries.



The users can explore the world map by clicking on a country. This action will result in a detailed description of the country's official languages on the left side. Clicking on the name of an official language will highlight all the countries with the same official language on the map. This visualization aligns with the plans in milestone 2.

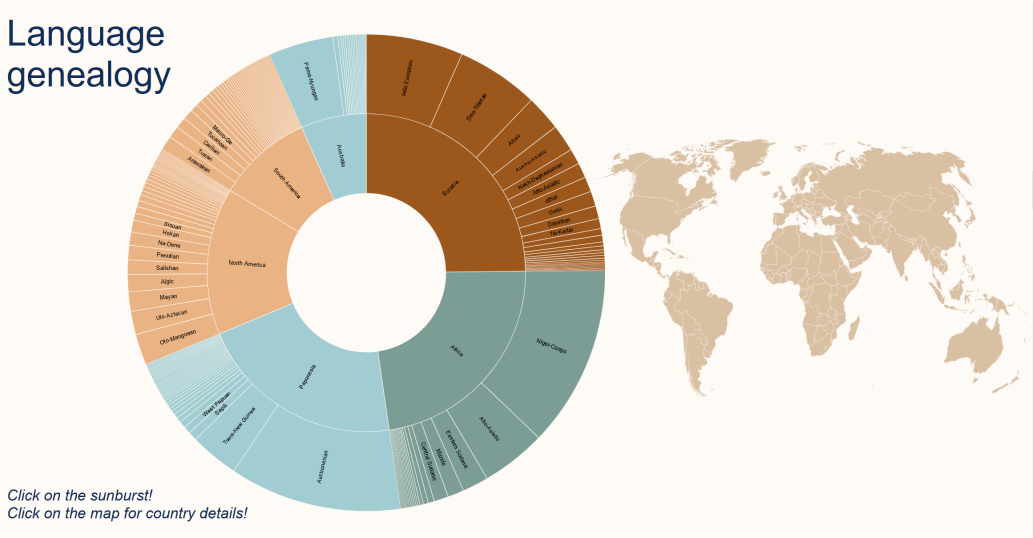
The implementation involved the drawing of the map, including the highlighting of countries on hover to elicit interactivity. When a country is clicked, its color is changed and the name and official lists bound in the JSON are displayed on the left text. This text has an additional D3.js layer of functionality, as we add a further on-click behavior to each string containing the official languages. This allows to again go over the map with to highlight the countries with a matching language ISO code.

Note: This is the only section that uses data from current official languages (Countries Info). For the rest of the sections, we always use all the languages as recorded historically in the WALS dataset.

3.3 Section 2 - Language genealogy

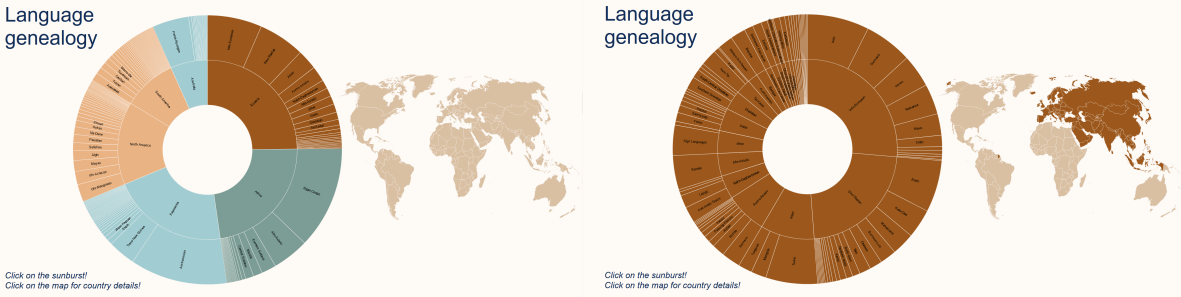
For the second section, we visualized the most general similarity indicators among languages: family and genus. Language genealogy studies the historical relationships between languages, tracing their evolution over time. Like a family tree, languages can be grouped into families based on shared ancestry. For example, the Romance languages (Spanish, French, Italian, etc.) all descend from Latin. This fascinating field highlights the connections between cultures and how human communication has evolved.

The visualization consists of two parts. The first part consists of a zoomable sunburst chart that displays the hierarchical genealogical structure of languages. At the top of the hierarchy is the macro area (i.e., Eurasia, South America). In subsequent levels, we differentiate based on family, genus, and language. The chart displays a maximum of two levels of the hierarchy. Users can click on a label to zoom in further in



the hierarchy or on the empty central circle to zoom out one level. In addition, hovering the mouse over a sunburst section for 1 second will pop up a tooltip illustrating the category name to assist with readability.

The second part of the visualization consists of a world map. The map highlights the countries in the group category previously clicked on the sunburst. Users can click on a country for detailed information on its languages, families, and genera. As in the previous section, the visualization aligns with the milestone 2 plans, apart from the tooltips to provide more detailed information.



View before (left) and after (right) clicking on "Eurasia" on the sunburst chart.

The implementation of the sunburst chart was based on the D3 Gallery zoomable-sunburst[9], adapting it to the employed palette and datasets. We also added functionality for changing the opacity on hover to elicit interactivity. In addition, we added a world map object configured to pop up a tooltip box when clicking on a country.

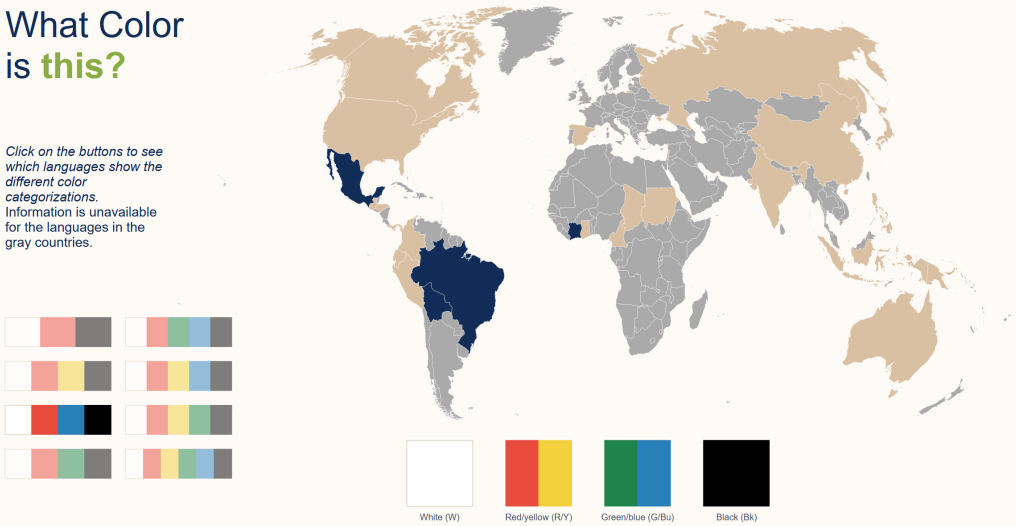
The most challenging part of this section was providing the data in a hierarchical JSON format. We used Python to extract the relevant data from the WALS CSV (i.e., macro area, family, genus, language) and construct the hierarchical JSON as required by the sunburst chart module.

3.4 Section 3 - Color categories

Despite the general disparity between the color terms employed by different languages, all of them contain a series of basic words for color encompassing white, red, yellow, green, blue and black. This section helps the user visualize the number and classification of basic color terms among languages.

In the visualization, the user can click on eight different buttons representing the basic color categories of the associated language; opacity change on hovering and clicking allows the user to learn they are clickable and which button is selected. When one button is clicked, the map highlights the countries where a language

[9]https://observablehq.com/@d3/zoomable-sunburst

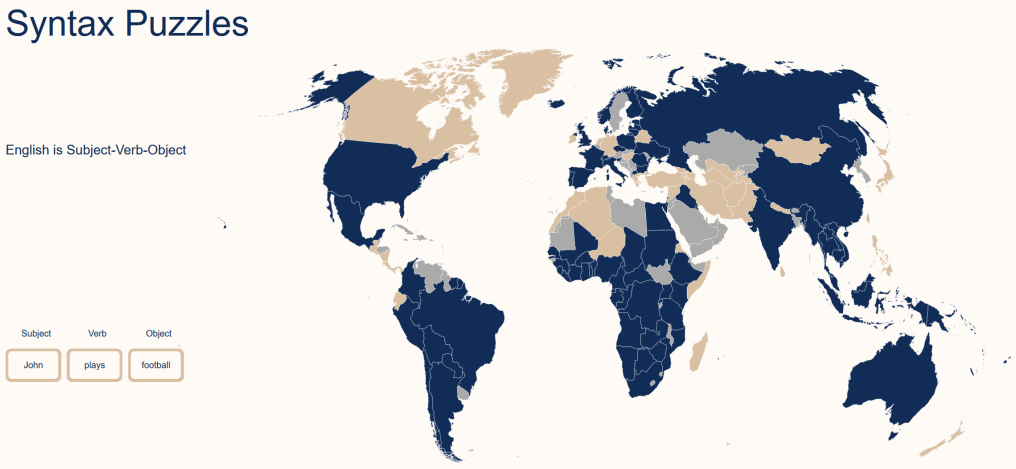


using the selected basic color categorization is spoken. A palette showing how the six basic color categories are grouped in that selection is also shown; the specific color tonalities employed were selected to be coherent with the overall website palette.

Up to this point, the design aligns with the plans in milestone 2. As we realized few countries had the features needed to be categorized, we added the possibility to color countries in gray if none of their languages included the features. Additionally, we added the possibility to click the countries to display a tooltip listing the relevant languages and which color categorization scheme they used.

3.5 Section 4 - Sentence order

In this section we explore different syntaxes used by languages across the world. For this, we chose to present the ways languages order the Object, the Subject and the Verb in a sentence. For example, English uses the Subject - Verb - Object syntax: e.g. "John (Subject) plays (Verb) football (Object)".



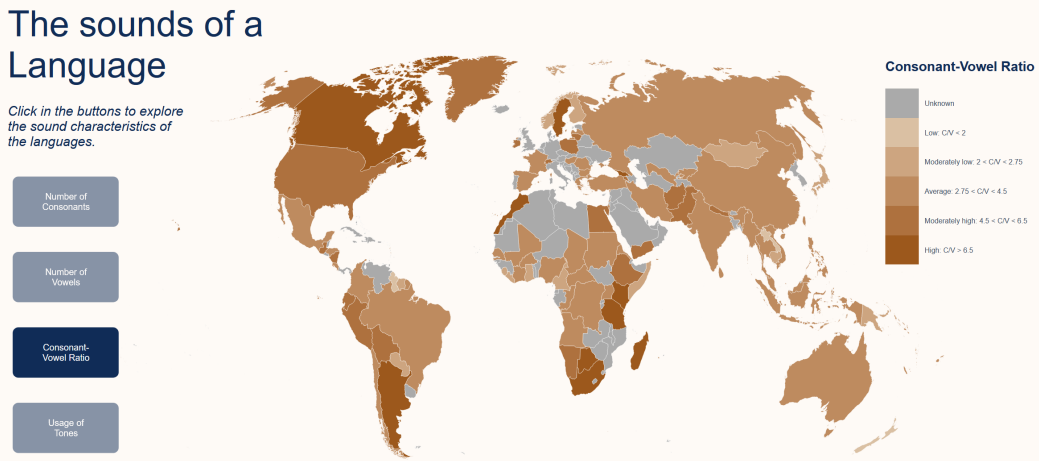
In the visualization, the user can play a mini puzzle by ordering an example sentence in various ways. The user drags and drops the Subject, Verb and Object of the example sentence in an order of their choosing. Once the sentence is complete, a map visualization highlights the countries where the specified syntax is used. By double-clicking on each word, the user can reset it so that they can try different syntagma orderings. This implementation aligns with the plan proposed in the second milestone.

We decided to further enhance the syntax exploration experience, adding two additional features in the visualization. The first one is a set of *fun facts* about syntax across the world. These are displayed in above the syntax-puzzle area once the user has completed a sentence. The second extension is the ability to hover over the world map and select a country to view its languages' syntax. We present information about the syntax of the native languages, not the official ones.

The challenged in this section's implementation was the drag-and-drop tool. We developed a custom word drag-and-drop widget using HTML and javascript elements. The widget allows each word to be dropped in a word-reception area. We also added the reset functionality, where a double-click event signals the corresponding word element to return back to its original place. Once all word-reception places are filled and the sentence is complete, the widget signals to the D3.js map element to highlight the corresponding geographical areas.

3.6 Section 5 - Consonant and vowels

The next visualization illustrates the disparities among the sounds used by different languages. By observing world choropleths, the user can learn about the amount of vowels and consonant in a language, the ratio between them, and if a language uses a tone system and how complex.



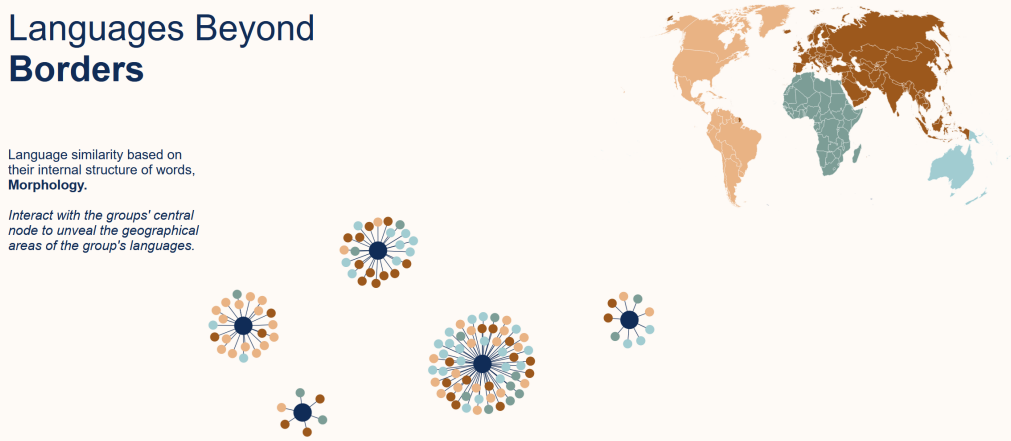
The section shows a world map where the countries are colored according to the described categories. The user can select one feature to display by clicking on the buttons on the left of the page, which will update the colors of the map and display the corresponding legend on the top right corner. Additionally, clicking on one country displays a tooltip detailing the specified category for all the languages spoken there. Finally, changes in opacity and grey countries are used for eliciting interactivity and signaling lack of data, as in Section 3. The visualization mostly aligns with the second milestone's plan, but has been expanded to show several features.

Since the languages spoken in a country can have different values in the analyzed features, our visualization colors the country according to the more frequent value among all of the languages.

3.7 Section 6 - Feature clusters

Each of the previous sections explores an singular aspect of the world languages, e.g. color description. In this section we attempt a more high-level exploration, investigating similarity of languages in the concept-space of morphology. We visualize clusters of native languages based on their *Morphology*[10]. In this section, languages across the globe, separated by borders and long distances, are brought close together in the user's screen space, based on their closeness in the morphological features.

[10]https://en.wikipedia.org/wiki/Morphology_(linguistics)



This section is comprised of two components. The first one, occupying the main space of the screen, is a cluster visualization, where morphologically similar languages are grouped in the same cluster. The language nodes are colored according to their geographical area. The second option is a world map, acting as a color-legend as each geographical area is color-matched with the cluster graph. By hovering over the central node of a cluster, the geographical areas corresponding to the cluster's languages are highlighted.

For the implementation, the first step was to cluster the languages based on their morphology. We used the following WALS morphology features[11]:

1. 23A Locus of Marking in the Clause
2. 24A Locus of Marking in Possessive Noun Phrases
3. 25A Locus of Marking: Whole-language Typology
4. 26A Prefixing vs. Suffixing in Inflectional
5. 27A Reduplication

We used python to group the languages into five clusters. Since the features are discrete and not continuous we chose the *kmodes*[12] python library. Our python script outputs the resulting clusters into a JSON file. For visualizing the groups we used the D3.js disjoint force-directed graph tool. The JSON file containing the cluster information was formatted to facilitate easy and fast parsing by the disjoint graph.

4 Peer assesment

Dimitrios Samakovlis

- Dataset search
- Visualization ideas
- Webpage design
- Worldmap implementation
- Language Genealogy
- Screencast
- Process book

[11] https://wals.info/feature
[12] https://pypi.org/project/kmodes/

Rafael Medina Morillas

- Dataset search
- Visualization ideas
- Webpage design
- Worldmap implementation
- Official languages
- Color categories
- Consonant and vowels
- Screencast
- Process book

Christodoulos Kechris

- Dataset search
- Visualization ideas
- Webpage design
- Worldmap implementation
- Official languages
- Sentence order
- Feature clusters
- Screencast
- Process book