Process Book

11.30.2018

About

Team Members:

- Archanasri Subramanian , u1141789@utah.edu
- Krithika Iyer , u1135255@utah.edu

Github Repository: https://github.com/archanasris/dataviscourse-pr-languagesoftheworld

Overview and Motivation

We looked at many datasets for visualization like effect of droughts on crops in USA, women in law dataset from World Bank repository etc. We felt that these datasets didn't provide us much scope for visualization. After some more search we found the WORLD ATLAS OF LANGUAGE STRUCTURES dataset which had a variety of features ranging from geographical prevalence to semantic structures which motivated us to come up with interesting and unique ways to visualise them.

People don't speak one universal language, or even a handful. Instead, today our species collectively speaks over 7,000 distinct languages.And these languages are not spread randomly across the planet. Hence we thought it would interesting to study the different properties and structures of languages spoken in the world.

Related Work

Our inspiration was the visualizations created by the Puff Puff team of the Dipartimento Di Design at Politecnico di Milano. During their Design Density course, they developed interesting visualizations using the same dataset to understand the popular families of languages, the number of people that spoke them and the loanword status of popular languages.

Link: http://www.puffpuffproject.com/languages.html

Questions

Before We are trying to answer the following questions through visualizations:

1. What are the different families of languages that has existed over time?
2. How many countries and people speak a particular family of language?
3. How many countries and people speak each language within a particular family of language?
4. What are the different gender systems used by different languages?
5. How are the word orders and grammar rules different between languages?

As we went about the implementation the questions we were trying to answer have evolved to be :

1. What are the popular families of languages?
2. How many countries speak a particular family of language?
3. How many languages each country speaks?
4. What are the different gender systems used by different languages?
5. What is the distribution of languages in each gender system?
6. What are the locations of the various categories of languages?

Data

The data set used for this project is the **"WORLD ATLAS OF LANGUAGE STRUCTURES".**

Link: https://wals.info/

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 (many of them the leading authorities on the subject)

The data set is rich in features (~ 200 features) and has data for 2680 languages. We do not plan on using all the features for visualization. As we went around working on the visualizations we created our data structures with the necessary features on the fly using javascript.

Implementation and Design Evolution

Landing Page

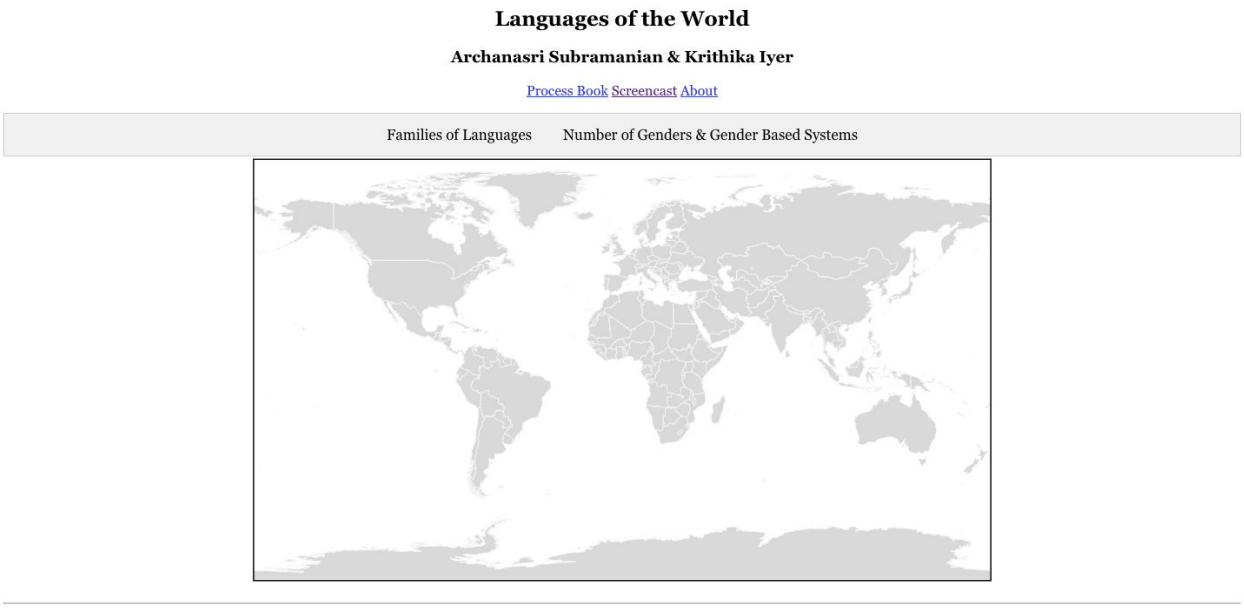

Fig.Landing Page

The landing page has a map which is constant with tabs for viewing the two main visualizations that we have

1. Families of Languages
2. Number of Genders & Gender Based Systems

Clicking on any of them shows the associated visualization below the map. In addition to these two tabs, we have links to view the Process Book and the Screencast.

Evolution:

After the peer feedback session, we decided to have a landing page which would have two different pages for viewing the two main visualizations but after a TA review session, we decided to have tabs instead of pages. Our old landing page looked like this:

Fig.Landing Page (old)

Visualization: 1

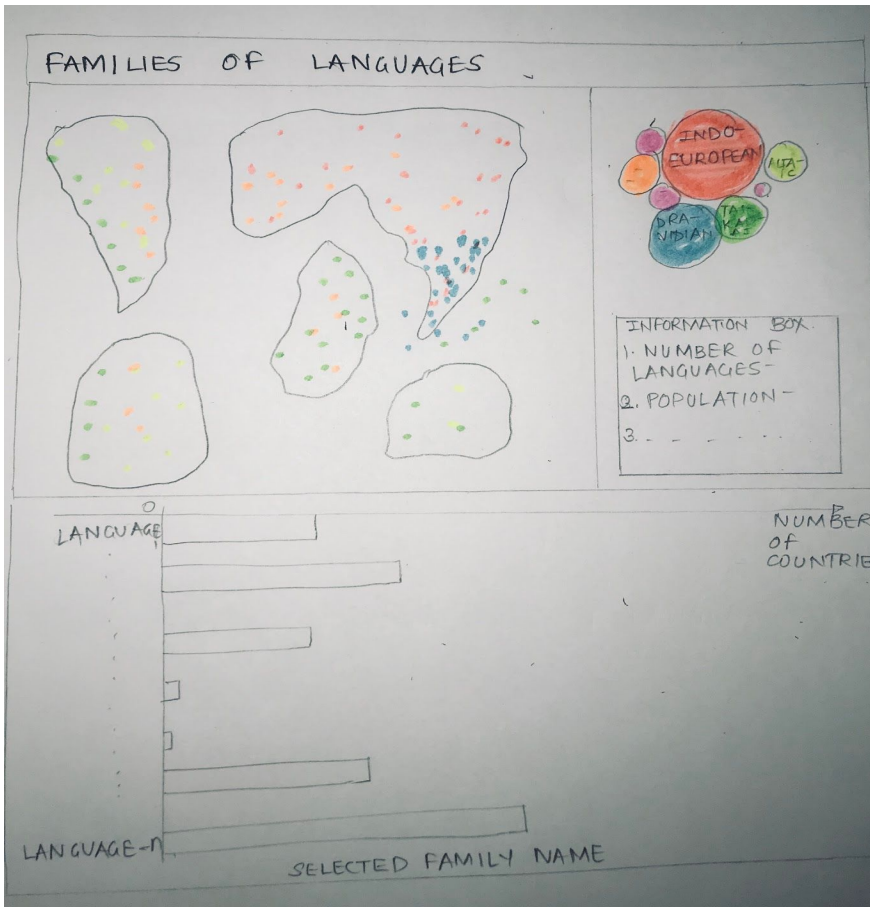

Fig.Design of Visualization 1

Data Processing:

1. In this visualization, all the families of languages with more than 20 countries speaking them were first extracted. They were later sorted and stored as an object.
2. A hierarchy structure (d3.hierarchy(myObject)) is used on the extracted dataset to identify the root nodes and children nodes.
3. The size of each bubble is dependent on the size of the family of language to show prominence of each of them and color coded with the help of d3.scaleOrdinal(d3.schemeCategory10).

Fig.Data used to plot the Bubble Chart

Fig.Data used to plot the Bar Chart

Visualization and Interactivity:

1. On selection of a particular family, the languages associated with that particular family are highlighted on the map based on the region where they are spoken. The data points on the map are color coded.
2. For each family of language, a bar chart will be created indicating all the countries belonging to the family and the number of languages that each country speak. This bar chart will help in visualizing the popularity of each language within a family.
3. The tooltip on the bubble chart and bar chart gives more details such as the number of languages that belong to that family, genus etc. This would help in understanding the spread of macro languages over the world.

Evolution:

1. To answer our questions regarding the families of languages, we decided to use a bubble chart which gives us insight about the popular families of languages and how popular they are.
2. At first, we thought we will use the bar chart to display the number of countries that speak the languages within a family of language. This led to too many bars which didn't help in answering our question so we decided to use the bar to look at the different countries that belong to the family of language and number of languages spoken in that country.

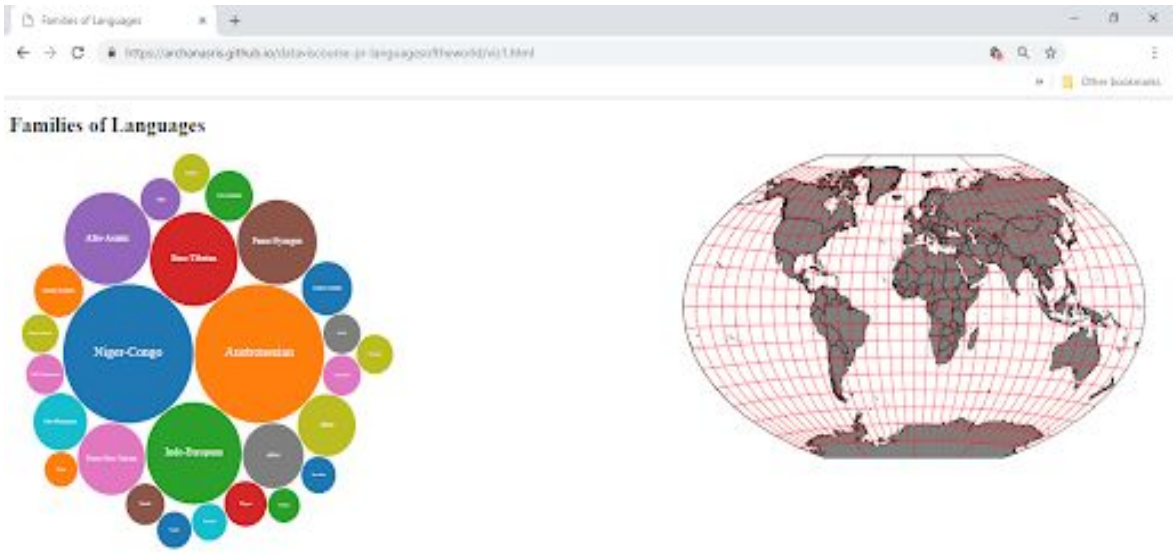

Fig. View of Visualization (old)

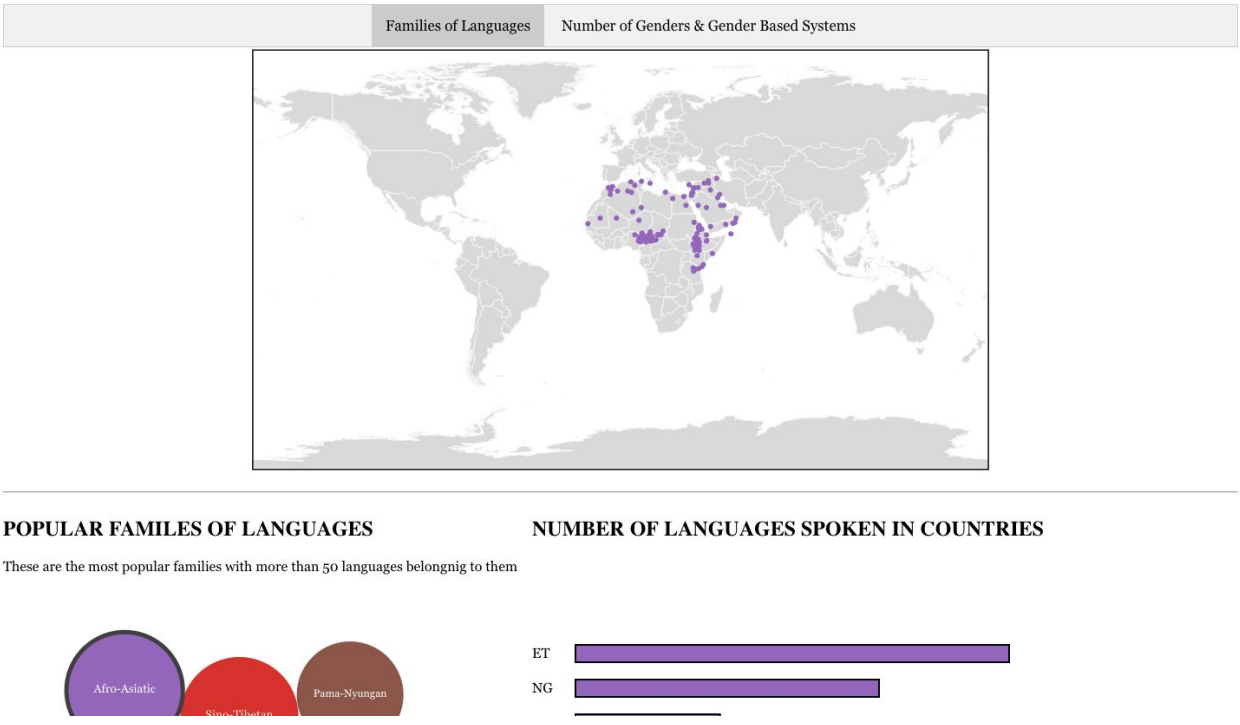

Fig. Visualization 1 - View 1

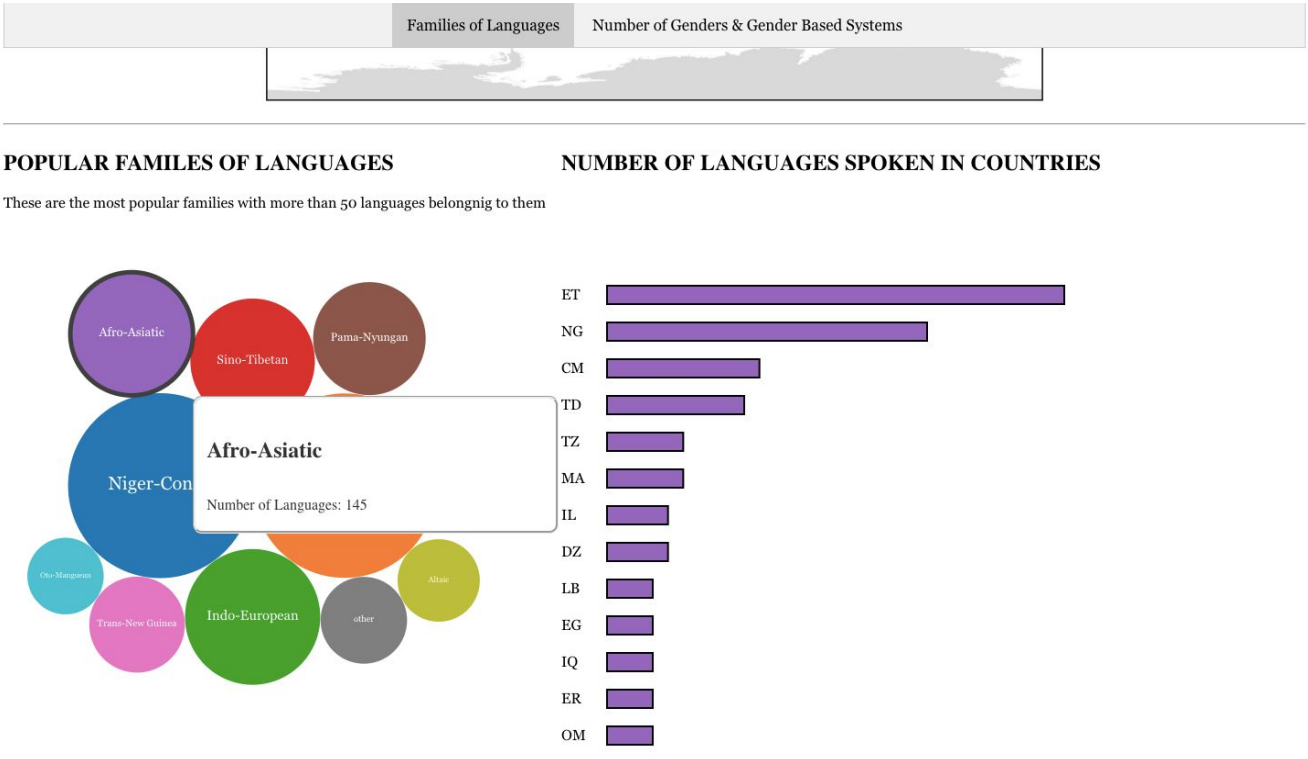

Fig. Visualization 1 - View 2

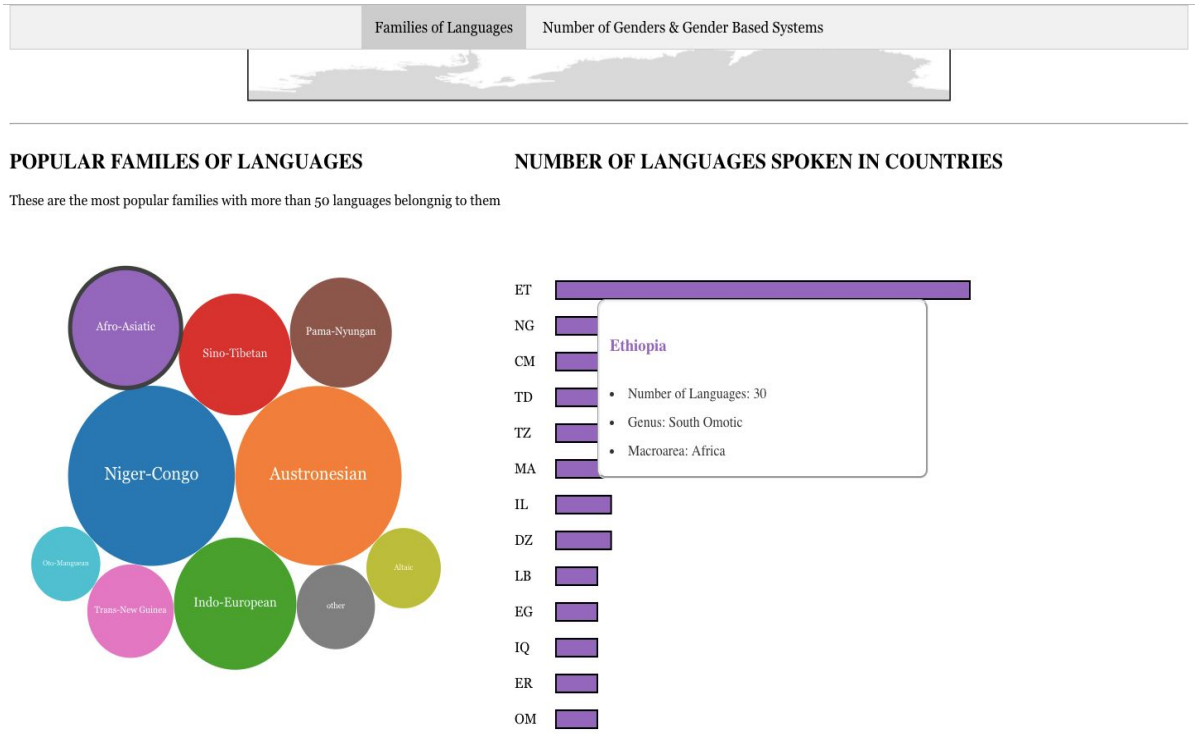

Fig. Visualization 1 - View 3

Visualization 2:

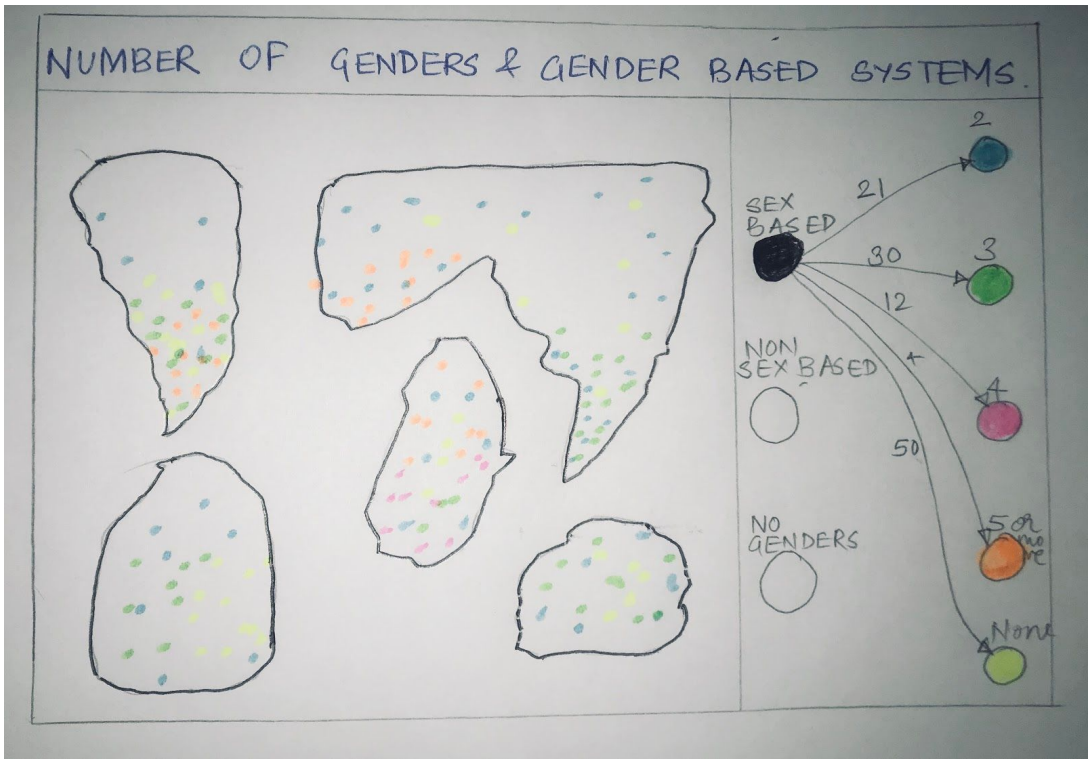

Fig.Design of Visualization 2

In this visualization, we plan to explore the languages and their gender based systems.

Data Processing:

1. Map - We use the **world.json** file for plotting the map. The projection used is **d3.geoWinkel3()**. As we go about with the development of the project we could look into other projections if required. The dataset has the location of the places where the languages are spoken (latitude and longitude) which we will use in order to plot those points on the map.
2. For the tree layout, the data is grouped using d3.nest using the feature gender based system in order to get out first grouping of “Sex-based systems” and “Non-Sex based Systems”.
3. These groups are further separated based on the number of genders used in the languages to get our second level of data. These separated groups are stored in a object specified by us which we could append to the map or other elements to plot them.

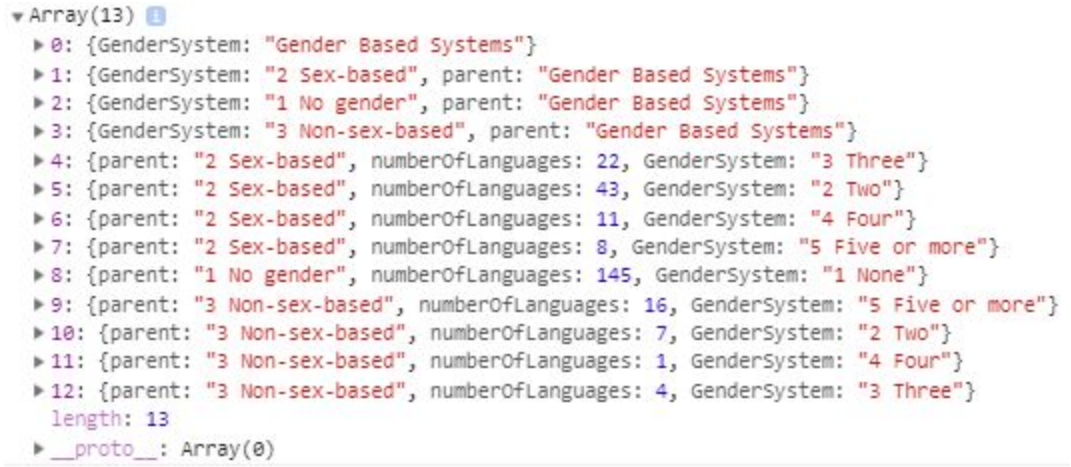

Fig.Data used to plot the tree

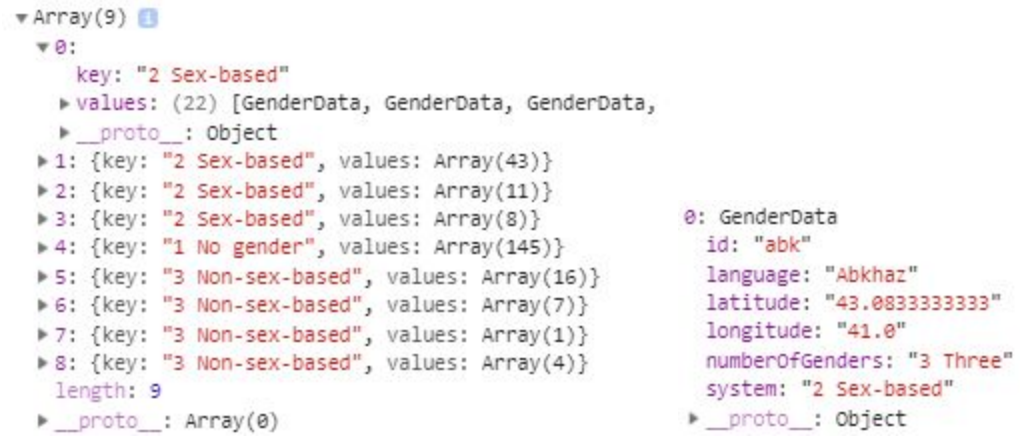

Fig.Dataset grouped by the gender systems and the GenderData object

4. Using the d3.stratify() and d3.hierarchy() a tree structure is rendered for this data.

Visualization and Interactivity:

1. We want to make the tree graph and the map interactive, such that when a node in the tree is clicked all the languages corresponding to that category are displayed on the map.
2. The categories will be color coded on the tree and on the map.

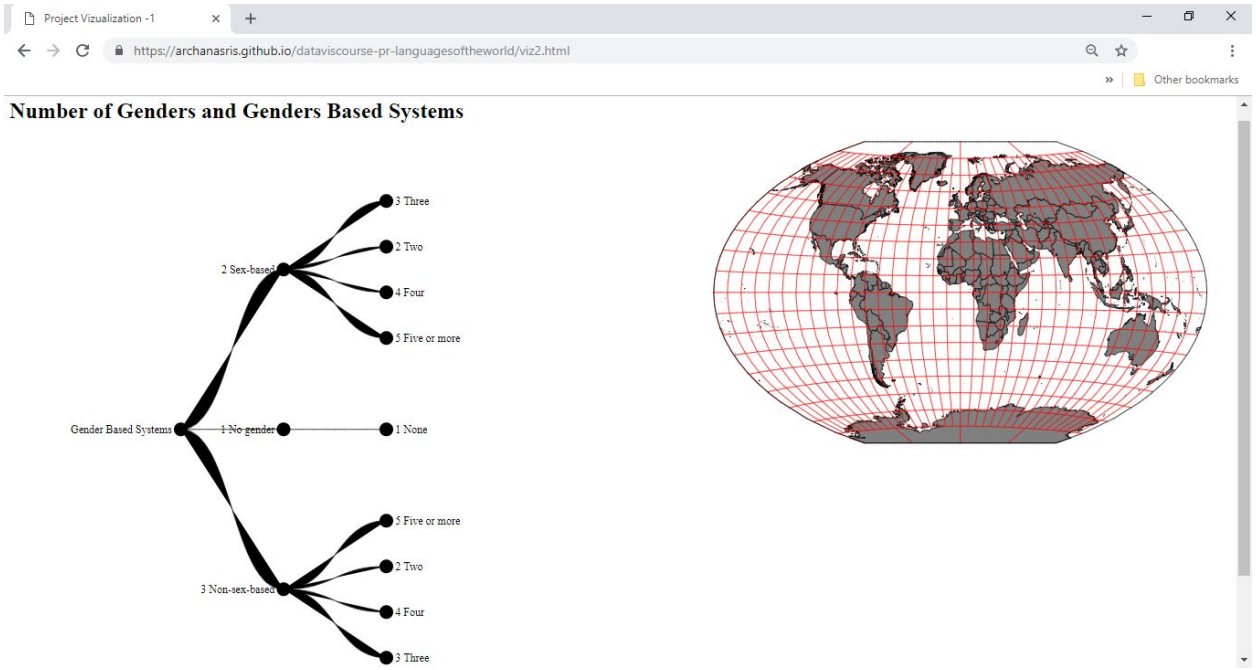

Fig.View of Visualization (old)

Evolution:

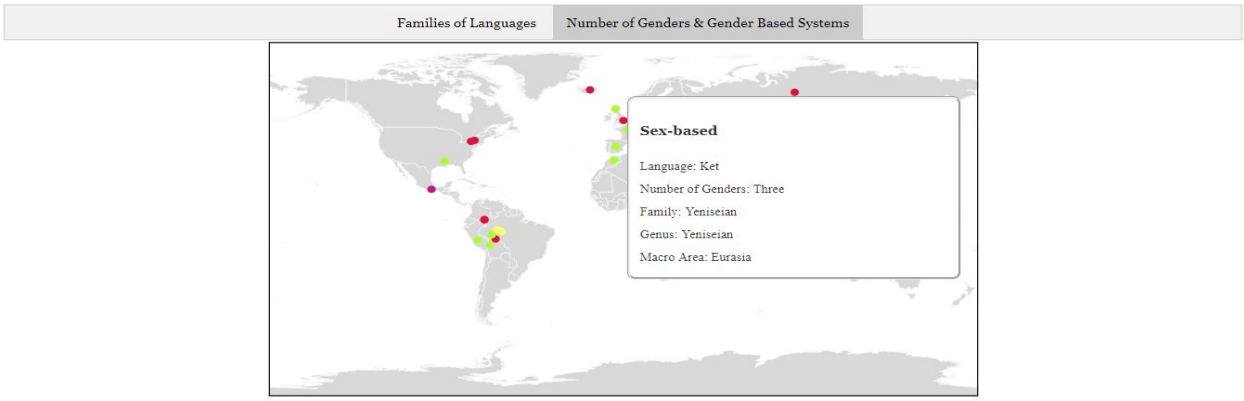

Fig.Visualization 2 - View 1

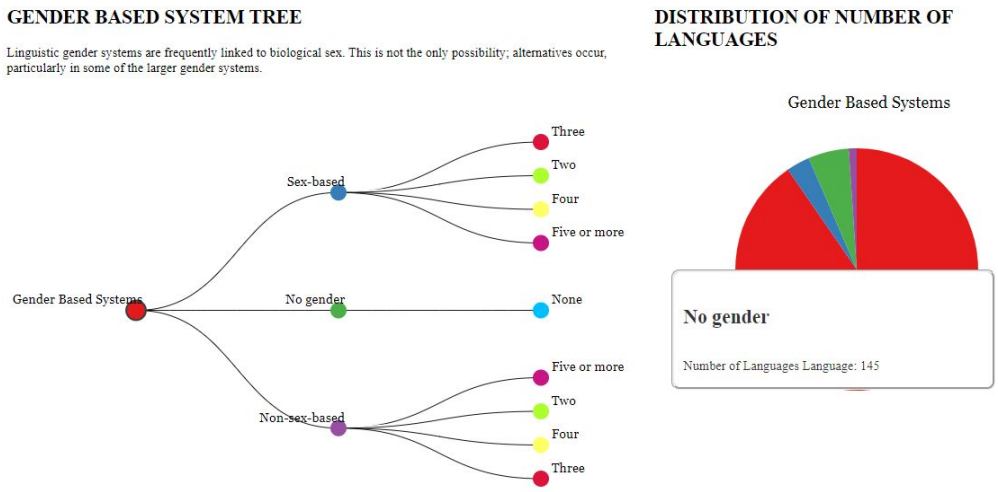

Fig. View of Visualization 2

The following changes were made:

1. We decided to change the projection used for the map from **d3.geoWinkel3()** to **d3.geoPatterson().** The reason we chose other projection was,the Patterson projection is a very good cylindrical for a rectangular map. It still exaggerates the polar areas but to a lesser degree than comparable projections, and preserves shapes of the land masses very well. We observed that while showing the points on the geoWinkel projection because of the curved and compressed projection the points which lie close to each other was not distinguishable which is solved in geoPatterson to a certain extent.
2. The tree is color coded, hence we can easily spot the sector corresponding to the system in the pie chart and we can link the points on the map to the pie chart and the tree.
3. As we you see in the map we display the information regarding the language on which the user hovers. Also we show a tooltip for the pie chart which tells us the number of languages belonging to that node of the tree.
4. Pie chart is displayed only for the first two levels of the tree as the leaf nodes have no further classification to qualify for a pie chart

Heatmap Visualization:

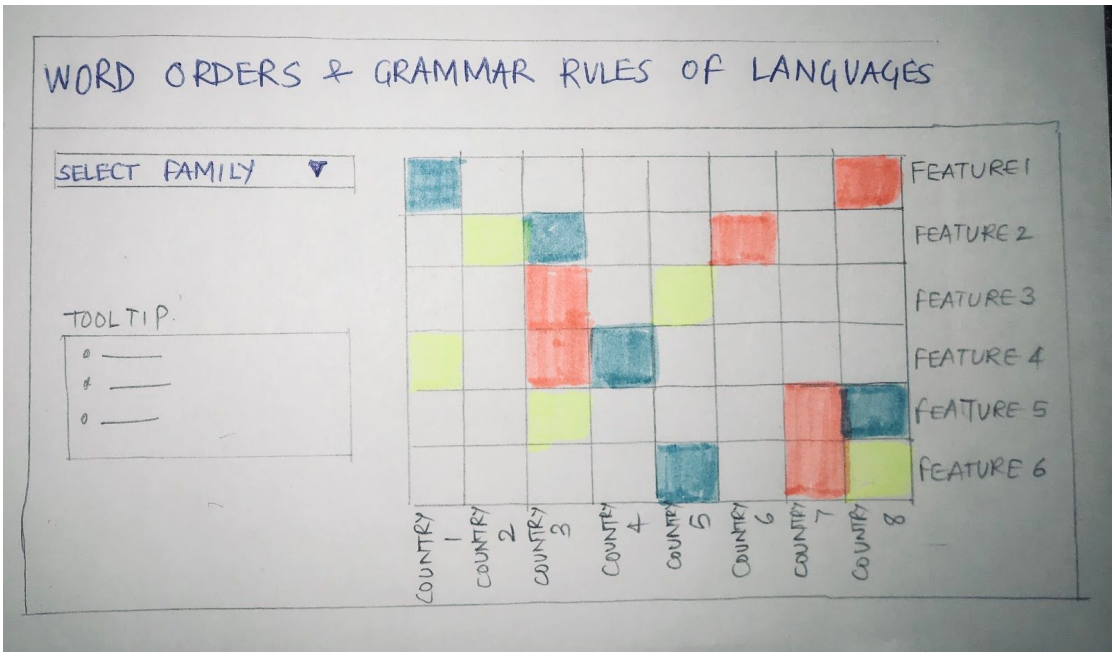

Fig.Design of Heatmap Visualization

Data Processing:

As per peer feedback review for our heatmap implementation, we explored the grammar rules of two categories - countries and languages and decided to go ahead with languages which presented more intuitive patterns.

Visualization and Implementation:

We integrated the heat map with the bubble chart in the first visualisation of family of languages.

Evolution:

Since the heatmap did not show us the information we expected to see we decided to remove it from our visualization.

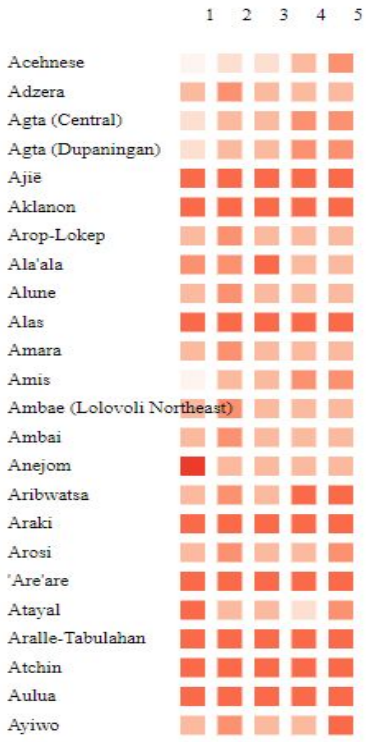

Fig.Heatmap Visualization

Evaluation

What we learnt regarding the data ?

Visualization 1:

Indonesia,Australia and India speak the most number of languages. The families which have the most languages are Austronesian ,Niger-Congo and Indo-European. We learnt that the population of a region does not necessary relate to the number of languages spoken.

Visualization 2:

Sex-based systems are found in almost all areas where there is gender. Given a gender system, the most common number of genders is two. Such languages are found in most gender areas. Of the 112 languages with gender in the sample, three quarters (84) have sex-based systems. The main non-sex-based area is covered by the extensive Niger-Congo family in western, central and southern Africa, which contributes 17 of the 28 non-sex-based systems. Most of these have five or more genders, but Grebo and Koromfe have three. The other substantial non-sex-based area is that of the Algonquian family of

North America.The wide scatter of these languages shows that animacy is a viable basis for gender systems. Nevertheless, it is overshadowed by sex-based-systems.

Possible Improvements

While exploring this dataset, we realised that this database is not very extensive. We also saw that we could not provide enough user interaction in terms of drill down for this dataset. Hence, the visualization can be improved by combining this database with other language database like the loan word system database, language spoken by population database etc.