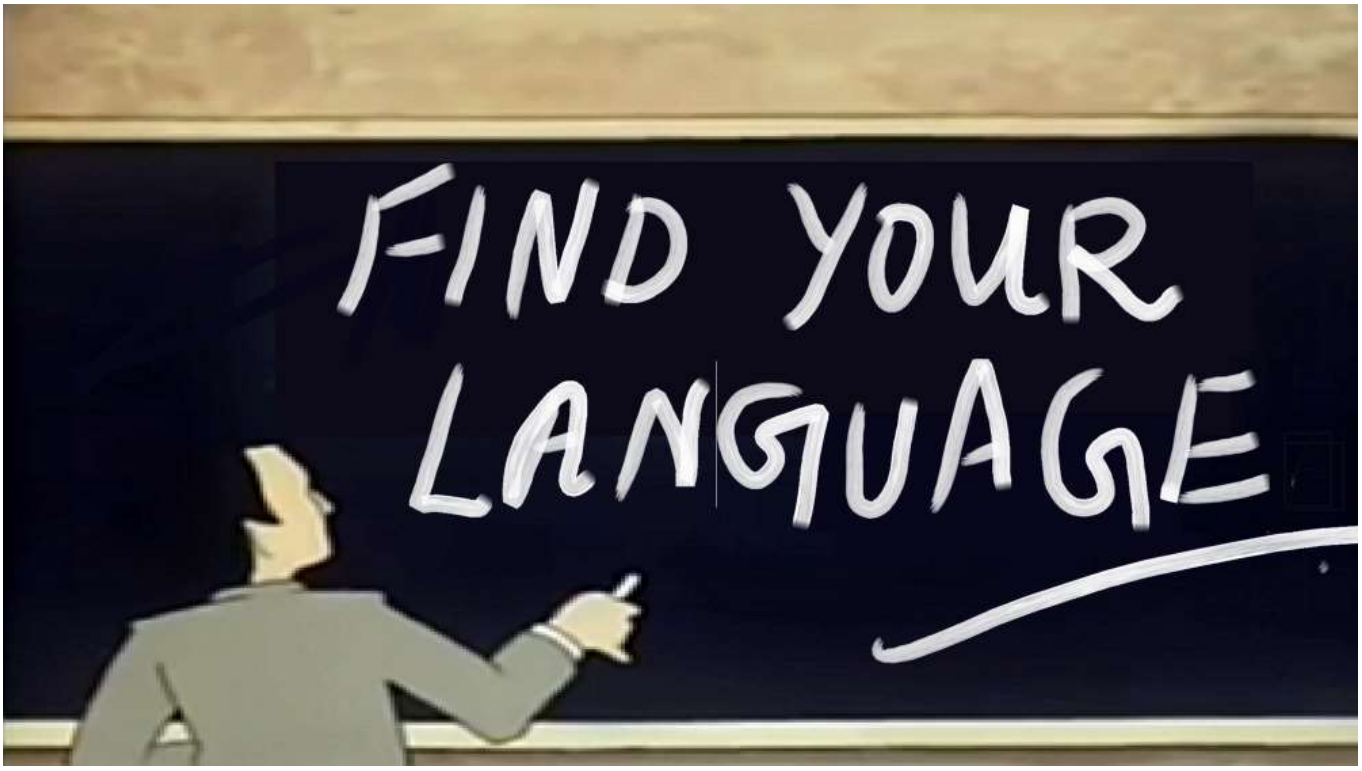
Find your Language!

Vasudha Varadarajan
Apr 25 · 6 min read



April 25: The Languages dashboard is a page for you to visualize the languages all around the world and the relationships between them. The insights we could get from this are exciting, and I would like to see how this turns out too. I will maintain this page to update my progress on the project.
This is a part of the final project for the Visualization course at Stony Brook University.

April 27: I have successfully set up an interactive map for selecting areas (not longitudes and latitudes yet, but that should be easy) on the world map. I intend to let the user select the region of the world they are interested in knowing about, and then presenting the linguistic profile of the area. Next, I am trying to check how the data should be displayed since it is so vast.

Interactive world map

There are around 240 language families that could be initially presented on the world map with certain coordinates. When the user chooses one of these families, he/she is presented with a dendrogram of the language hierarchy, and information of each family/language/dialect on hover, maybe with some interesting facts if possible. This is a deviation from the initial plan of letting the user choose the countries. I will explore both the options to check which might be better.

April 30: As planned before, I will be setting up a radial dendrogram to show 450 families of languages in a single screen as well as I possibly can. Are there better ways? I should think about that. I want to set up a structure of a radial dendrogram first so that I can later plug in values from my database.

May 1: I have successfully set up a radial dendrogram. I am also looking into the possibility of using other hierarchical data representations, such as sunbursts or trees with zooming functionality. While with sunburst the depths are made more compact and zooming in could be natural, it wouldn't be able to show the families right away, and

it might be a little non-intuitive for my data. Zoomable trees take too much space for me to be able to present a dashboard for analyzing other aspects of language families, other than the hierarchy itself.

While I aim to show the hierarchy in the language families, my primary aim is to check if certain parameters affect the language spread and dominance. So I have chosen a zoomable radial dendrogram, where each node is taken to a dendrogram with that node as the central parent. Clicking on the central element takes you back to the parent dendrogram, and so on. We could delve deep into the trees to check the hierarchies out. But that's not all. We want to show the other parameters, the geographical spread of subfamilies and languages, and interrelations between them. Zoomable radial dendrogram seems to be a good choice for this.

May 5: Today I learned that the Glottolog data uses a certain format for storing trees called the Newick format. While it is supposed to be an efficient representation, the time taken to compute the children of some language families are very high. Sometimes, it takes up to 5 hours. Should I precompute the values? If I did that it might make the tool a little sluggish.

> *How could a zoomable radial dendrogram be accomplished?*

May 6: Precomputed values are not very feasible because parsing the Newick trees for the last few levels of the tree still takes a lot of time. For the sake of the final project's deadlines and the time crunch, I have clustered the language families based on continents they are spoken on, and then show a demo of how a zoomable radial dendrogram could be accomplished. There are two levels, continent names, and families. Clicking on continent names zooms in on the continent. Each word in the dendrogram is clickable.

Radial dendrogram with only 2 levels at a time (to enable zooming)

May 12: My hypothesis for this project: Does the number of children languages give some information about the popularity of the language family? We can see that the Indo-European language family seems to have taken over the world. By doing this exercise, we could take a look at the rarer languages to see how things could be more accessible, and how we could recognize and help the rarer languages of the world.

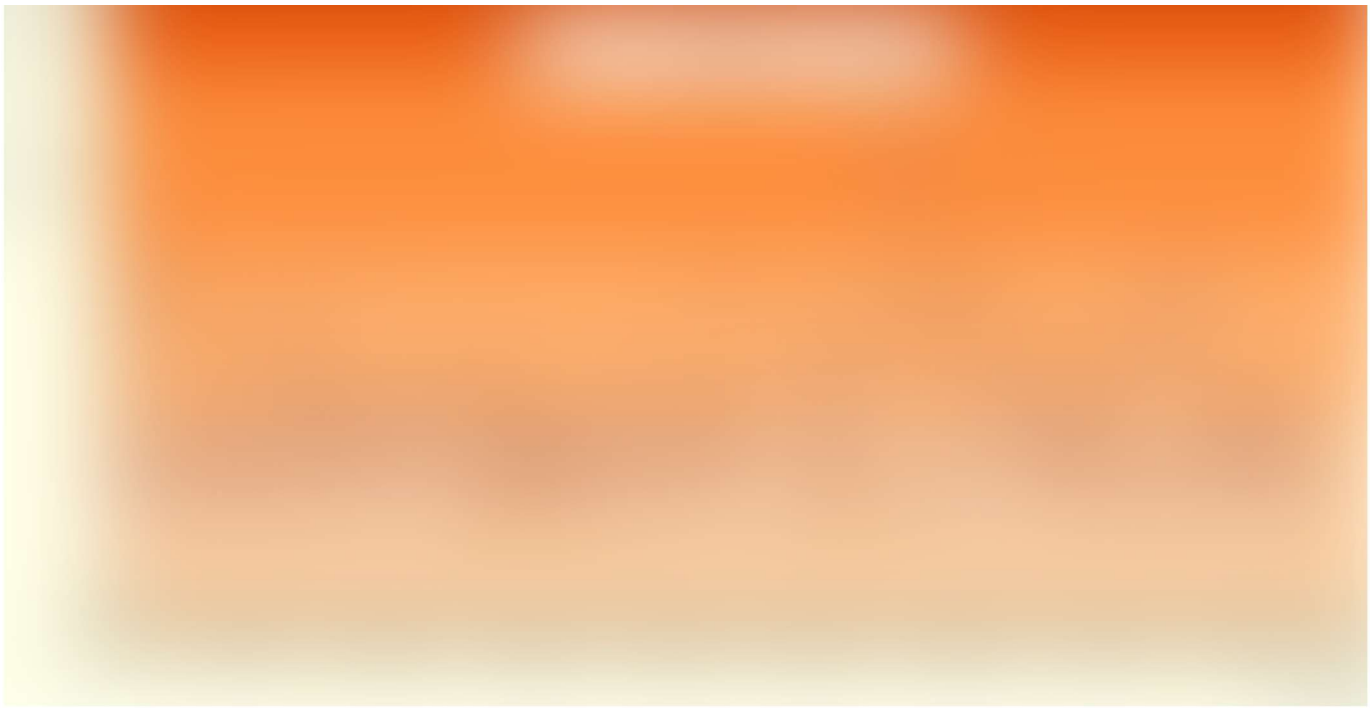
> *Does the number of children languages give some information about the popularity of the language family?*

Zoomed dendrogram

> *Temporal information might come in useful to analyze the popularity of languages over the years but there is very little data out there.*

On the other hand, focusing on the present, we could check for the languages in need of preservation- the critically endangered ones in a much more visual way.

To do this, I have decided to implement a scatter plot of language families against the number of children, and we could help filter out the relevant language to look at by adding a brushing component for the users to pick relevant data points (families) to delve deeper into the radial tree and dig for information. The points could also be mapped to the local areas on the map, although given the time constraint I will only be able to map to continents or countries at the most. Glottolog only maps families to continents and the world coordinates data for languages is very sparse, so it wouldn't make sense to show carrots on a map as I had thought of initially. To get the number of children for each family, I will be iterating through the family trees, and it takes a lot of time to parse each Newick trees because of the sheer size.



Scatterplot showing the families and the number of children

May 14: Because of time constraints, I have finally decided to give up on the multilevel zoomable radial dendrogram, though I really think I should do more during the summer. I have implemented something like proof-of-concept: just for 2 levels. More levels would add so much interaction to the dashboard but I will stick to this for now. I am thinking of color-coding the scatterplot to map it to the areas on the geography, but the barrier is that I do not know the dominant languages in each country to be able to unambiguously tell which countries to map each data point to. I will think about it.

Clicking a language family maps it to the geographical area

> *In the scatter plot diagram, a lot of families are concentrated toward the bottom- these are the very endangered languages.*

May 18: Almost done with the final dashboard. I am calling it "Find your Language!", like a pun on "Mind your language!" because I believe we should all mind- pay attention to- the languages. Lots of small bug fixes left. I tried changing the color schema of the page but orange works out the best for me.

May 19: Finally ready for submission. I have a very small screen, apologies!

Final dashboard (somewhat)

Visualization D3js