Language Diversity in the United States

Alexia Adams

# **Project Goals and Motivation**

The goal of this project is to highlight the language diversity in the United States. I was motivated to do this because in March, English was designated as the official language of the US. While this can be viewed as a good or bad thing, I wanted to take the time to acknowledge and showcase the 100s of languages that are spoken here.

# **Intended Audience and Use Case**

My intended audience are people who are curious about the languages spoken in the US. Some intended use cases are:

* Finding how many people speak a particular language in the US (not English)
* Discovering regional differences in the languages spoken
* Determining the most dominant languages in the US and in particular states

# **Inspirations**

A visualization that inspired my second visual section (ranked map) is from the [Visual Capitalist](https://www.visualcapitalist.com/most-common-language-spoken-in-the-u-s-map/). They have a visual (below) that shows the most common languages in each state besides English and Spanish. While the visual looks interesting, it can be difficult to read all the languages/colors being shown. Additionally, I wanted to expand upon this idea, showing the third most common, fourth most common etc. instead of limiting it to the second most common or the most common that is not English or Spanish.

**A map of the united states

Description automatically generated**

Visualization showing the most common languages in the US besides English and Spanish by the Visual Capitalist.

There are many other visualizations you can find online that show the most common languages besides English. Many of them seem convoluted yet provide a lot of detail. I wanted to focus on being simpler and integrating user engagement to show the specific languages rather than show everything up front.

# **Data Source and Preprocessing**

The dataset comes from the United States Census Bureau, specifically the dataset: Detailed Languages Spoken at Home an Ability to Speak English for the Population 5 Years and Over from 2009 to 2013. While the data is a little outdated, this was the most recent dataset I could find with specific languages and counts.

My preprocessing code is in “VizLang.ipynb”. The original 53 datasets were Excel files, so I had to convert them into csv files and then merge them together. I renamed columns and removed unwanted columns like those relating to margin of errors (since the data are estimates). I had to convert the values of come columns into numerical values by removing any commas of larger numbers or replacing the text such as “(D)” which represented an undisclosed number into -1 to make it easier to track during coding. The original dataset had groups (ex: “Other Indo-European”) and subgroups (ex: “Other West Germanic languages”) as rows alongside the languages. I wanted to be able to filter languages by groups/subgroups, so I made columns “group-labels” and “subgroup-labels” for each language storing this information. Specifically for the state data, I added the columns “state” (name of the state) and “abbr” (abbreviation of the state) to be able to keep track of which rows/language data were associated with which state.

# **Design Iterations**

A notebook with writing on it

Description automatically generated

Sketch of original plan for the project.

The original plan for this project was to include the following visualizations: a tree map depicting the distribution of English only speakers vs English and another language vs Another language only, a map showing the most common languages, a lollipop chart of most common languages based a selected state, a map showing the distribution of a selected languages, and a lollipop chart showing the number of speakers of different languages in a group (ex: Indigenous Languages in North America). I wanted to be able to extract as much information as possible form the dataset into visualizations, but I tried to focus on narrowing what I wanted to create.

A paper with writing on it

Description automatically generated

Sketch of possible design for the first visual section.

Something I thought users would want to see were the most common languages overall in the US, so I opted to create a visualization depicting these languages. I initially wanted to include English as part of this visualization but thought that it would be extremely difficult to read the other languages (since about 80% of the population speaks English only). Thus, I focused on the most common languages besides English. I also considered whether I should include the top ten languages, all (before I knew how many languages there were 🡪 would be too cluttered), or show some that pass a particular threshold. I decided to use a threshold of 500k because these languages had similar counts of speakers and half a million is easier of a number to interpret. During the sketching phase, I considered creating this visual as a lollipop or bar chart with the marks representing humans (rectangular body and circular head) that would show a speech bubble when hovering over them. However, I didn’t go through with this idea because of how tiny the humans would be for the languages besides Spanish, some of them you can’t even see the lines of the lollipop. I also didn’t want to get rid of Spanish because I wanted to highlight how it is a very common language in the US.

A paper with writing on it

Description automatically generated

Sketch of possible design for the details of a selected state in the second visual section.

I was considering including detail about the English proficiency based on each language. For example, in the details-breakdown of selected states in the second visual section, I wanted to color the circles of the lollipop based on the proportion of speakers who are proficient in English and then color the stems by the group. I decided against this because I wanted to focus more on just the different types of languages and wanted to keep the colors of the circles consistent across all visuals (the colors of the circles always depict type of language).

# **Final Design (Design Choices and Implementation)**

* **Marks**. The marks for most visualizations revolve around circles. This was to create consistency and be reminiscent of speech “bubbles” (I know they are more elongated but it’s the thought that counts).
* **Consistent** **Colors and Simple Scheme**. For consistency, the color scheme on the groups (ex: Other Indo-European) is kept throughout all visualizations to make it easier for user to interpret the visuals as they scroll through. For the second visual section in particular, having colors by group instead of by language makes the visual easier to interpret; at worst there could be 52 different colors on the map (if each state had a different language) which would make the legend cluttered and potential confuse the user with similar colors. Coloring by group does limit the visual detail but makes it easier to read.
* **Updating Text.** For the second and third visual sections, interacting with the visual can change the text beside the visual. For the second visual section, selecting at state, selecting a language, and changing the rank number will alter the text. For the third visual section, selecting a language will change the text.
* **Emphasize Functionality and Usage.** There are dropdown menus and number inputs that are outlined with dotted lines to draw attention to these interactivities that will change the visual. Information on what happens when the user interacts with these tools can assist the user in discovering new details from the visual.
* **Spanish Separated.** In the second visual section, the details-breakdown separates Spanish from the Everything Else category because most states have Spanish as their second language, so I wanted to emphasize other languages as well. Including Spanish in the lollipop chart would skew the data and make the other languages hard to interpret, like in the first visual section, but worse since the chart is smaller.
* **Tracking Languages.** In the second visual section, it is hard to see which languages are spoken because the coloring is by group. To alleviate this, users can pick a language to track, which will make the circle darker and have the text color be that of the group, to try to make it stand out amongst the other languages.

A screenshot of a computer

Description automatically generated

First visualization section on the languages that have at least half a million speakers besides English.

The first visual section is static (no interactions) but it depicts the most common languages besides English that have at least a half a million speakers. Notice the difference between Spanish and the next most common language. Including this visual at the beginning makes it easier to exclude Spanish in the details-breakdown later on because it is prevalently a common language.

A screenshot of a computer

Description automatically generated

Second visualization section on the most common languages across states.

The second visual section includes a map of the states (+ DC and Puerto Rico) and their most common languages. Users can input a number (1 to 251) to change to a different set of most common languages. Users can hover over the states to see which language is that most common language. Users can also track a language on the map. The dropdown menu for selecting languages is grouped in the same way the colors are (except it excludes English). Selecting a state on the map will produce a details-breakdown of that state.

A screenshot of a computer

Description automatically generated

Second visualization section on the most common languages in the selected state New York.

The details-breakdown includes a proportional bar chart of the proportion of speakers that speak either English, Spanish, or anything else. There is also a lollipop chart that shows the ten most common languages in the Everything Else category. These visuals are static.

A screenshot of a computer screen

Description automatically generated

Third visualization section on the language distribution of Urdu across the states.

The third visual section includes a map that displays the language distribution based on the number of speakers. Users can select which language to look at, which can change the color scale accordingly. Users can also hover over states to see the number of speakers for the selected language. If the selected language does not have any known number of speakers (only has undisclosed amounts), text is added to the visual saying the color scale is not applicable.

# **Contributions**

Since this was a solo project, I did all the sketching, designs, and implementations. I would estimate that this project took me at least 20 hours.