Visualizing Language Diversity in the United States

Calvin Phung
Massachusetts Institute of Technology

Alejandro Camacho
Massachusetts Institute of Technology

Abstract

This project is a visualization of various languages and where they are spoken in the United States. The goal of our visualization is to provide a tool for users to gain a better understanding of the diversity in the US through the unique medium of language, which doesn't get discussed as often compared to race or nationality. We use data from the US census bureau compiled by the American Community survey to showcase the number of speakers by metropolitan areas, counties, and states. In order to assess the effectiveness of our visualization in highlighting diversity, we ask three participants about the most populous ancestries in the US before and after using our visualization to determine whether our visual had any effect in the correctness of their answers. What we found was that our visual did help users in correctly identifying ancestries as well as gaining a better understanding about diversity in the US.

Index Terms: Languages, Diversity, Orthographic Maps, Origins

1 Introduction

The United States continues to diversify every year, and the 2020 US census bureau even estimated that nearly 4 in 10 Americans identify with a race or ethnic group other than white. The data further suggested that the decade from 2010 to 2020 will be the first time the White population proportion actually declined [2]. What's often left out of these discussions about ethnic diversity, however, is the multitude of spoken languages across the United States which can oftentimes paint a richer and more accurate representation of all the different community clusters. The U.S. Census Bureau provides five categorical responses to the race question: White, Black or African American, American Indian or Alaska Native, Asian, and Native Hawaiian or Other Pacific Islander. In contrast, the American Community Survey (ACS), a demographics survey program conducted by the U.S. Census Bureau, found that the number of unique languages and language groups spoken in the United States came out to be 350, based on data collected from 2009 to 2013 [4].

The goal of this paper is to determine whether visualizing the spoken languages in the United States provides a useful tool for understanding diversity. We define diversity in terms of the range of different social, ethnic, and ancestral backgrounds extending from the five main categories provided by the census bureau: for example, German, Irish, Thai, or Japanese. From this work, our contribution will be the creation of a novel interactive visualization of languages spoken in the United States mapped along a geographic dimension to enable the exploration of ethnic and ancestral diversity.

2 Related Work

There have been several works visualizing language diversity in the United States, but none of them provide the level of specificity, interactivity, and functionality that we do. Business Insider created a static visualization of the most common language spoken at home other than English or Spanish for 2017 by state. However, while their visual provides a clean look that showcases language diversity, it is unable to capture the scale of each language because the map displays only categorical values. Ben Blatt at Slate Magazine also published a piece showcasing several visuals highlighting the most commonly spoken languages in each state. His visuals provide greater insight into language diversity because he also includes diversity within different families of languages like the most commonly spoken Native American, Scandinavian, Indo-Aryan, and African language. Both the visuals created by Business Insider and Ben Blatt suit their mediums well because they are easily interpretable at a glance which is ideal for web articles. However, for our purpose, we wanted to create an interactive visualization that not only encouraged exploration, but also serves as an accessible and effective resource for researching language diversity. Our work therefore distinguishes itself from past work by providing greater interactivity, depth, and functionality.

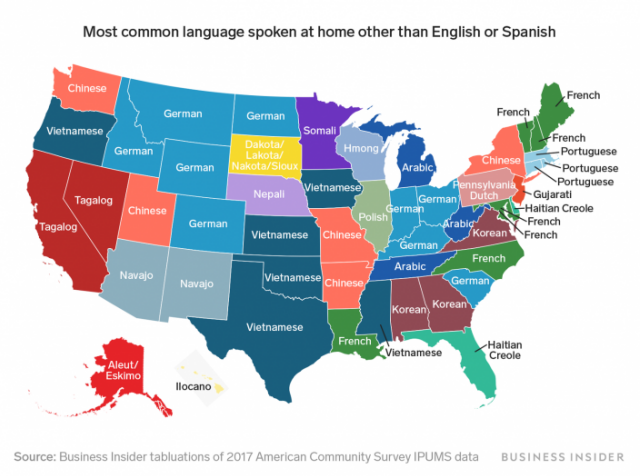

Figure 1: A visualization of spoken languages from Business Insider [3]

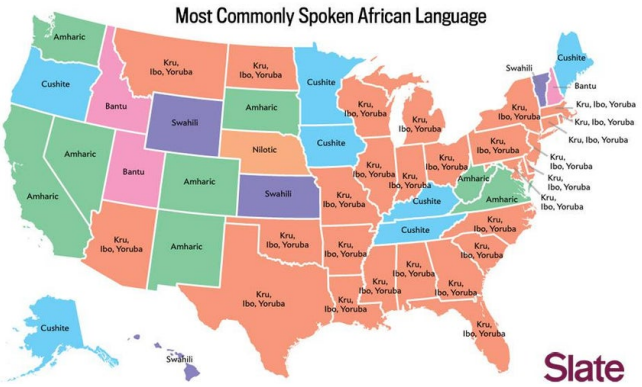

Figure 2: A visualization of spoken languages from Slate [1]

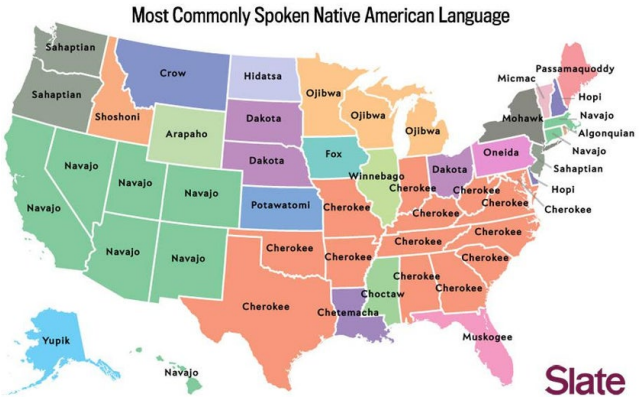

Figure 3: A visualization of spoken languages from Slate [1]

3 Methods

The first step of our inquiry is to identify where languages are spoken, how many speakers are in those areas, and what country that language is primarily spoken in. Then, we will build our site and conduct a quick user study.

3.1 Datasets

In order to visualize language speaking populations geographically, we combined census bureau data compiled by the American Community Survey for metro areas, counties, and states. We looked specifically at the field of "Languages Spoken at Home" for each geographic designation. Although we were able to acquire the most recent data (2019) for the counties and states, the most recent data for metro areas was from a 5-year estimate for the years 2009-2013. It is significant here to acknowledge a limitation of the ACS dataset. Due to concerns of possibly infringing on citizens' privacy in some of the smaller counties with few speakers of a language, the ACS dataset does not publicly provide the same level of language granularity for counties as there is for metropolitan areas or states. Where the states dataset might have data for the number of Cantonese speakers, for example, the counties dataset had to combine Cantonese along with all other Asian languages into a category called "Asian/Pacific Islander" languages. [5]

We then manually augmented the languages datasets with each language's ISO 639-2 Code in order to have a consistent ID to associate each language across different datasets. This was important because we also used Dryer and Haspelmath's "World Atlas of Language Structures Online" to geographically locate language origins and primary country of speakers [**?**].

In order to create the D3 projections of the U.S. map and 3D globe, we used two geoJSON files: the Natural Earth public domain map dataset and a pre-built TopoJSON from the U.S. Census Bureau.

Finally, we acquired sound clips of people speaking out phrases of various languages from Simon Ager's Omniglot website in order to supplement the user's exploration of languages with an audio dimension not usually provided by these types of visualizations. By allowing the user to hear how languages sounded, we hoped to increase engagement by encouraging users to interact with our visuals and learn about less well-known languages.

3.2 Implementation

We integrated D3.js with React and Express.js in order to create a fully dynamic site with a customized web server. Our web server processes and cleans all of our datasets sent to the front-end. We also use Google Drive storage to host the sound bites of language phrases which requires server integration with Google's Drive API. We include further discussion of the website in the Results section.

3.3 Evaluation

In the experimental phase, we will ask participants to guess ten of the twenty most populous ethnic/ancestral classifications that are more specific than the racial categories provided by the census. After interacting with our visualization, we will again ask participants the same question and measure the difference in the number of correct guesses. The answers are evaluated against the 2015 American Community Survey's list of the twenty most populous ancestries in the United States where German is the first most common and Filipino is the twentieth most common. We want to determine whether interacting with the visualization of language diversity will provide a stronger grasp of racial and ethnic diversity in the United States.

4 Results

With the goal of improving users' understanding of diversity in the United States, we created a website that visualizes the ACS spoken language dataset using an interactive map, an animated globe, and a left drawer that provides additional information. After selecting a language from the dropdown at the top of the page, each of these components update to reflect the selected language.

4.1 Interactive Map

Our map has 3 modes: "Metro Areas", "Counties", "States". Each mode visualizes different spatial information for a selected language. "Metro Areas" visualizes a circle at each of 100 different metropolitan areas, where the radius of the circle encodes the number of speakers for the selected language in the area. "Counties" visualizes the territorial lines for each of the 3,143 counties and county-equivalents in the 50 states and District of Columbia, where color encodes the number of speakers for the selected language in the area. "States" visualizes the territorial lines for each of the 50 states, where color encodes the number of speakers for the selected language in each state. On each map, we allow the user to pan and zoom to select more precisely some of the smaller counties or metro areas. Hovering over a metro area, county, or state reveals a tooltip with the name of the feature and the number of speakers.

We decided to have multiple modes to increase the level of granularity, as opposed to just seeing information at the state level, and to allow for more types of encoding, as different people may better understand changes in different encoding channels. Also, by displaying this information on a pannable and zoomable map, the visualized datasets are easier to navigate.

4.2 Animated Globe

When a user selects a language, the globe rotates to and highlights the nations in which that language is spoken by a significant proportion of the population. The user can then hover over each of these nations to reveal their names in a tooltip. Below this globe, we textually provide the names of the highlighted nations, as well as the genus and family of the selected language.

We decided to include an animated globe as opposed to just listing the nations because not all users will have the same level of understanding of where a nation may be in the world. So, by directly showing them on the globe, they have that geographical context that reduces mental effort by the user.

4.3 Drawer

When a user clicks a feature (metro area, county, or state), a left drawer pops out to provide additional information for the selected language in the context of the clicked feature. We tell the user the rank of the selected language and the total number of speakers. We provide a histogram that compares the selected language to the nearest neighboring languages by number of speakers. We also include an audio clip, audio clip transliteration, and audio clip English translation for the selected language.

We decided on using a left drawer to house the additional information because displaying all of this directly on the map or with multiple channels would have been overwhelming on the user. Also, the left drawer is flexible and allows for future additions (more in Future Work).

4.4 User Study

4.4.1 Phase 1

We had three participants for our user study who were each initially asked to name, without consulting sources, any ten of the twenty most populous ancestries in the United States. The bolded answers are correct. Their responses were the following:

- Participant 1: **China**, South Korea, **India**, **United Kingdom**, Canada, **Mexico**, Vietnam, Brazil, Taiwan, **Russia**
- Participant 2: **Chinese**, **Mexican**, **Indian**, **Irish**, **English**, **German**, **Italian**, **African American**, Israeli, Brazilian
- Participant 3: **African American**, **Irish**, **Mexican**, **German**, **Chinese**, Japanese, **Asian Indian**, Native American, **English**, Vietnamese

4.4.2 Phase 2

We then had the participants interact with our languages visualization for five minutes and then asked them the same question again and got the following results:

- Participant 1: **Mexico**, Central America, **China**, **Filipino**, Vietnam, Korea, Middle East, **India**, Brazil, Haiti
- Participant 2: **Mexican**, **Chinese**, **German**, Vietnamese, **Indian**, **Russian**, **Italian**, **French**, **Dutch**, **Irish**
- Participant 3: **Mexican**, **German**, **Irish**, **English**, **Chinese**, **Asian Indian**, Native American, **Filipino**, Korean, **Russian**

5 Discussion

Based on the results from our three participants, we found that for participants 2 and 3, their answers were much more closely aligned with the real distribution of ancestries after using our visual: participant 2 only had one wrong answer and participant 3 only had 2 wrong answers. For participant 1, although their answers strayed further from the ACS ancestry data, the answers saw more diverse ancestry with the inclusion of Central America and the Middle East. The participants in their initial answers already had some idea about the level of diversity in the United States, but after using the visual, they were able to better answer the question of "how" is America diverse?

Therefore, from our quick user study, we can see that our visualization is in fact beneficial for gaining a better understanding of the diversity in the United States. We provide various cues to help guide their exploration and understanding. By sorting the languages in the selection menu in descending order (and informing the user about this), users can navigate down this list and see how the colors get lighter and the circles get smaller. They can then learn, for example, that the US has a sizable French speaking population. Also, by providing sound clips, users can hear the differences in these languages if they are in more "only English" parts of the US. For example, someone in North Dakota can hear and compare Korean and Japanese, which they may have just previously grouped together as Asian languages without understanding any of their similarities or differences.

As they interacted with the visual, participants were most surprised about the ordering of the most populous languages and the proportion of speakers of a language in key areas. For example, one participant was curious about why there were so many Tagalog and Vietnamese speakers even though the number of people claiming Filipino and Vietnamese ancestry in the US only make up a small proportion of the population. They also learned that even though there were many people of German ancestry, the number of German speakers was drastically lower. What we can conclude is that even from our small sample of participants, users can derive key insights from interacting with our visualization, which encourages them to do more research to figure out why the visual appeared a certain way. This supports our claim that the visualization of language diversity serves as an effective resource for exploring diversity and learning about languages.

6 Future Work

We believe our site maximizes use of the globe and map; however, the left drawer is adaptable and has room for improvement. With more time and resources, we can expand on the left drawer to include more information about the language and its associated culture. In this way, the selected language is a gateway to learning more about the culture and nation it hails from. One way we could visualize this is by including more media from the selected language, such as video clips of people speaking the language, audio clips of popular music in the language, or lists of popular books in the language. Given how significant media is to convey the values of a culture, this could greatly improve our site's capabilities as an education tool.

References

[1] B. Blatt. Tagalog in California, Cherokee in Arkansas. *Slate*, 2014.

[2] W. H. Frey. The nation is diversifying even faster than predicted, according to new census data. *Brookings*, 2020.

[3] A. Kiersz. This map shows the most commonly spoken language in every US state, excluding English and Spanish. *Business Insider*, 2019.

[4] U.S. Census Bureau. *Census Bureau Reports at Least 350 Languages Spoken in U.S. Homes*, 2015.

[5] U.S. Census Bureau; American Community Survey. *Public Use Microdata Sample (PUMS)*, 2019.