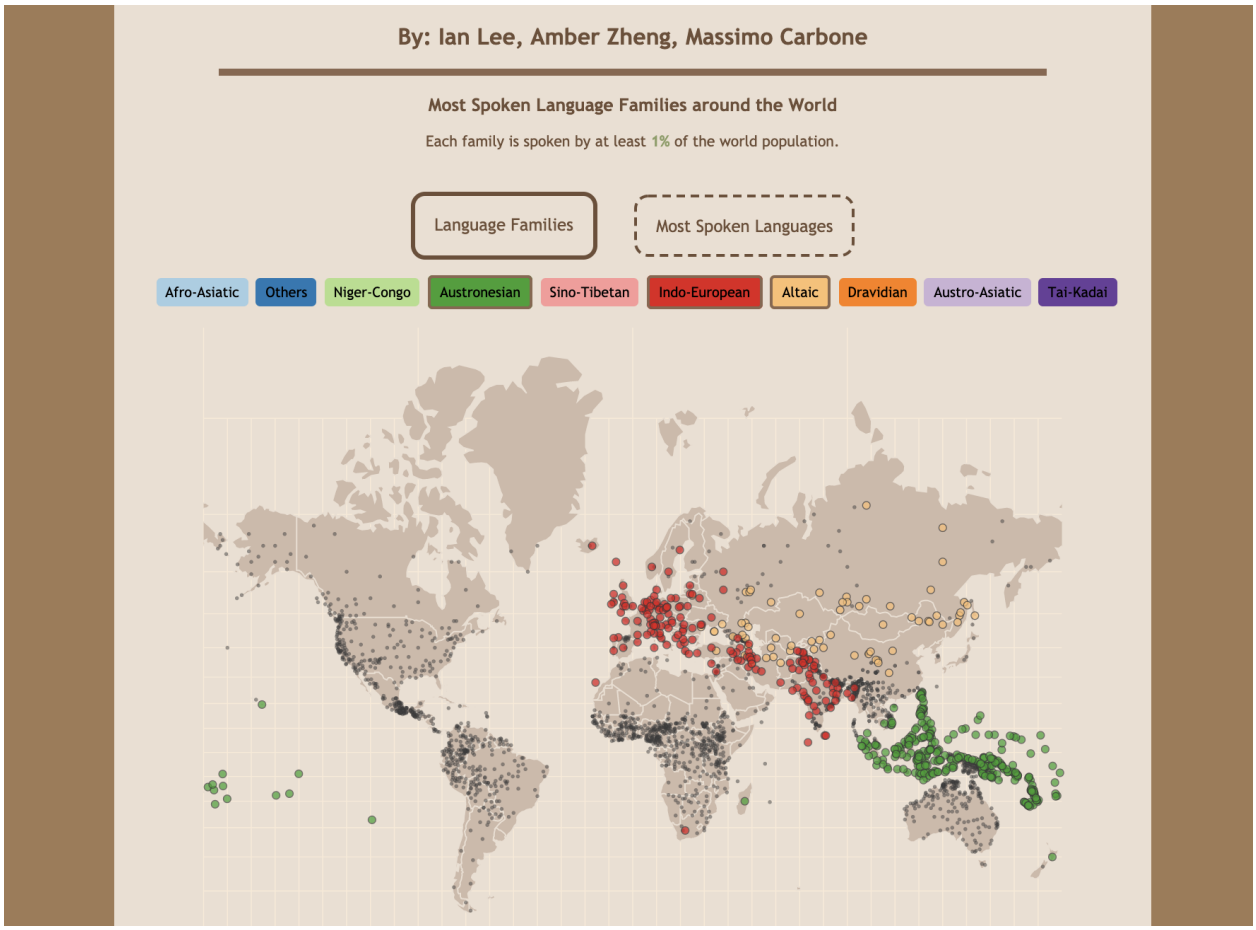
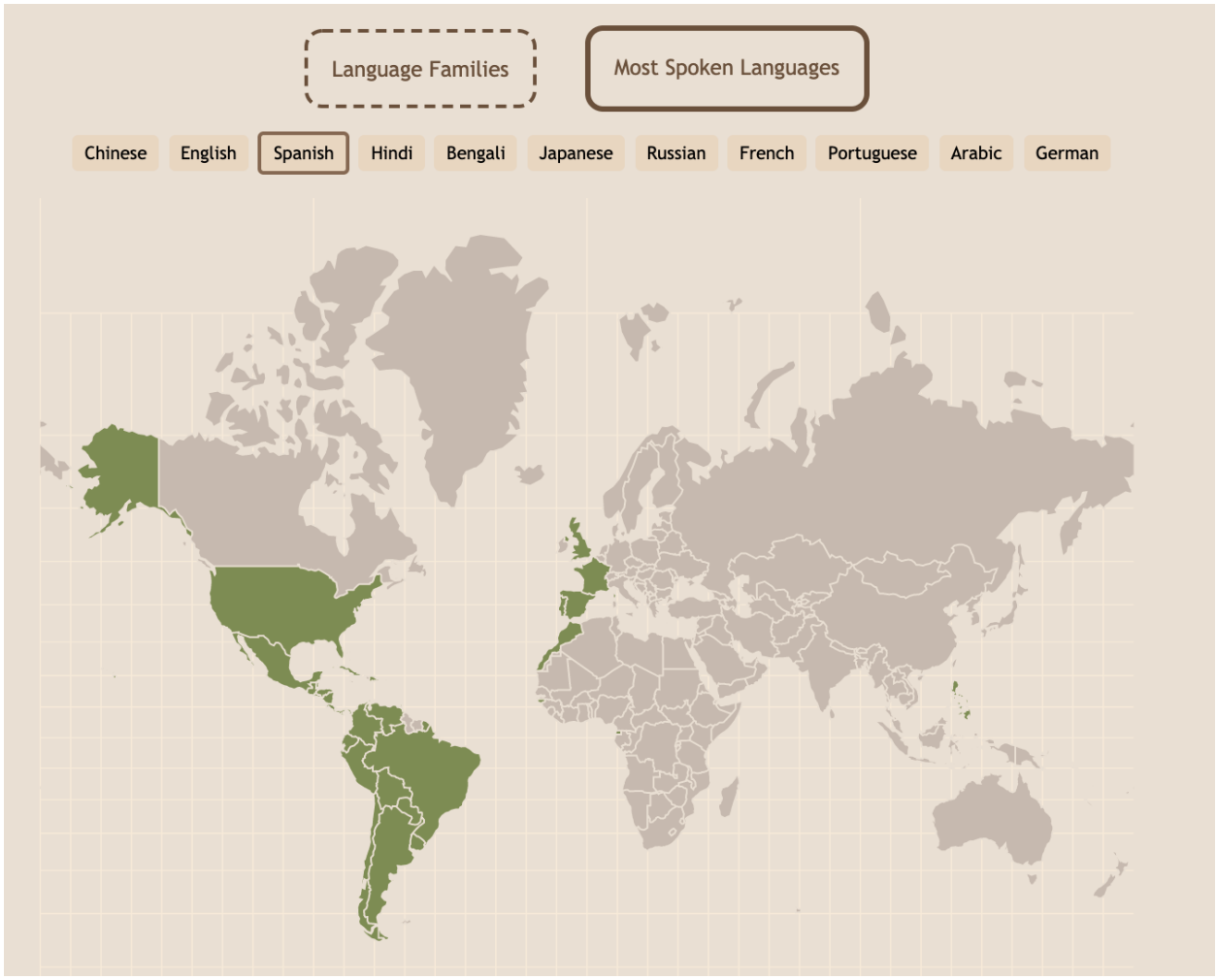
INFO 3300

Project 2

Amber Zheng, Ian Lee , Massimo Carbone





Project and Data Description

The majority of work was done finding usable and relevant datasets to use to create our visualizations. Usable, clean, and *accurate* language data is surprisingly difficult to find, especially when encompassing such a large dataset such as all existing languages across the globe. Our primary datasets were sourced from Ethnologue.com and Wals.info, as these had the most exhaustive and detailed sets. Ethnologue provided a trial dataset which included 3 .tab files describing different languages, their respective ISO 639-3 language codes (an internationally agreed upon set of codes), and country of origin specified by alpha-2 country code. LanguageCodes.tab specified each unique language and its code, as well as the country of origin. LanguageIndex.tab added a field which listed the status of the language, such as whether or not it is living or extinct, and CountryCodes.tab specified each country's country code. Combining this with population data from the United Nations and 3 letter alpha-3 country codes and numeric codes from this GitHub page gave us a good dataset to begin with. While helpful, we were unfortunately unable to acquire access to the entire Ethnologue dataset, which boasts this information and more detailed data such as exact coordinates, language families, precursor languages, and evolution history. We emailed asking for student access however Ethnologue responded with a quote for $600 and

the dataset could not be found in Cornell's library database. This forced us to combine the data from Ethnologue with another GitHub library linking specific language codes and countries to a more precise longitude and latitude, then linked these together with the tables matching language codes to respective countries. In addition, the Wals dataset proved immensely useful as it contained an open source dataset similar in scope to Ethnologue's with more detailed locations and descriptions for languages. Massimo compiled the Ethnologue data into one complete JSON file containing coordinates and population statistics using SQLite and Tableau, and then compiled Wal's dataset using a combination of Python and other scripts to produce a geoJSON file containing pertinent information. Ian then translated the geoJSON file to a topoJSON file so it could be used more readily/efficiently when building the site. We also found this handy world map library for D3 and topoJSON which allowed us to plot the points on our map visualization effectively.

Our overall intended design/look for this project, since we are dealing with data of a historical context, was rustic in nature, so we opted for natural colors such as light browns and beiges for the actual map itself. We found these colors to be differentiable and clear enough to serve as an adequate background while also contributing to our overall intended aesthetic. In addition, we added a background of longitude and latitude lines in white to add readability and most importantly an old cartographer-esque feel. Amber and Ian removed Antarctica from our world projection to remove unnecessary clutter (no unique languages have their root in Antarctica), and decided on a simple dot paradigm to display each language. If we had access to the complete Ethnologue dataset, we would have loved to include some system of clicking or hovering over a particular node to view a language's particular details, however this was not possible given our data. Additionally, doing something similar, such as listing each language name on a mouseover, would detract from readability and overall polish of our design, thus we opted for a category selector interaction where hovering over a language family would highlight corresponding languages on the map in that color, and clicking on said family would trigger a more permanent highlight. We chose an ordinal scale for the language families for best contrast and color distinctiveness. Our "most spoken languages" selector featured a similar UI, and made use of a single olive color to highlight countries and there should be no overlap when choosing a *single* most spoken language. Overall our most striking marks were the circular nodes and country/country borders, and we categorized them using channels of color and geographic location. As touched on previously, our interactive elements were designed with simplicity and "cleanliness" in mind, so a top navigation style bar with clickable visualization filters worked best for us. We also tried our best to keep all visualizations completely visible on one page so a user would not have to scroll through the page to find different features. This way, we hoped to achieve an intuitive user experience where everything is clearly laid out and displayed in a manner that is not confusing. We tried to make these interactive elements as interesting as possible by taking different approaches in the types of data we displayed, whether it would be specific languages or more general language statistics. If we had access to Ethnologue and the corresponding time information on each data, we would have implemented a slider mechanism to watch languages spread geographically through time, but we did the best we could with what we had.

We touched on this a bit earlier, but the initial vision we had for the project was a collection of visualizations that would allow us to see how, where, and when language evolved. Just learning about

the sheer number of languages was surprising in and of itself, and seeing the different densities of languages revealed a lot about linguistic diversity and also showed how humans slowly spread out over the globe, spreading language and culture. It would have been interesting to see the evolution of these languages over time, especially seeing different proto-languages evolve and branch into the multitude of languages we recognize today, but given the constraints we think our visualization does a good job of conveying the diversity and evolution of languages, which was our main goal. Overall a very interesting project to work on!

Group Contribution

Massimo: Found and worked extensively with the Ethnologue and Wals datasets, as well as other reference sets to create a starting point for project. Reached out to Ethnologue to use dataset and found alternative data when this was not viable. Cleaned, filtered, and merged data. Constructed geoJSON file for map visualization and merged .tab data.

Ian: Cleaned, filtered, and merged data. Create the two world map visualization as well as plot the languages and their associated families. The data cleaning process takes the most time as there are around four data sources to work with. Finding the suitable data for visualization also took quite some time.

Amber: Initiated 2 potential visualization ideas, found 2 sites used as data sources towards the final idea. Cleaned, filtered, and merged data. Created the two world map visualizations and implemented the interactive features. Implemented all the interface features including the overall color scheme, color scales, button appearances, etc. Wrote out the texts presented to explain the final visualizations.