The companion paper for the DHG501 final project

Mo Gengte

3140409

The GitHub link for my project (The map) and the datasets:

https://github.com/mokgengte/Final_project_for_DHG501

Background

Due to the special historical background, the language distribution in Taiwan is very diverse, there are four kinds of language used in the island currently: Formosan languages, which is indigenous languages for 16 aboriginal Austronesian in Taiwan island, Hokkien, also known as Minnan language(閩南語) or Taiwanese, Hakka language, a language from southern China, and Mandarin, which is Chinese in a narrow sense. Besides these, the Japanese also play an important part in the construction of modern language in Taiwan, specifically manifested as the Japanese loanwords that exist in all four languages mentioned above. This situation that been influenced by the Japanese language or vocabulary, doesn't show in the same kind of language as the Hokkien, Hakka, and Mandarin in the Chinese mainland, or other Austronesian languages in the Pacific. It is a sign, or a consequence of social change.

There have been several times of social and political change happened during the past hundreds of years in Taiwan, from the large number of immigrants from the coastal southern Han ethic group during the Qing dynasty, to the strong or forced cultural influence from the Japanese colonists after it defeated the central empire, to the arrival of the KMT regime after the surrender of its Japanese owner, to the dictationship under the martial rule after the KMT regime was expelled from the mainland China, to the democratization after the termination of the martial law. Each of these social changes would lead to a reflash of the distribution of languages.

The immigration flood of the coastal southern Han ethnic brought their own local languages to the island, which are the Hokkien and Hakka language, the former is one

of the major languages in Taiwan nowadays.

The occupation of the Japanese colonists brought their culture to the island, but unfortunately, this was a forced process, local languages were forbidden, local cultures were extinguished, and the colonial cultures were forced to be introduced, making all the languages in Taiwan deeply influenced by Japanese.

The arrival of KMT regime from the Chinese mainland brought Mandarin to the island, which was set to be the official language during the mainland period of the KMT regime. Due to the similarities between the local languages and Mandarin, and the popularization efforts made by the authorities, it quickly became the most common language in the island, but the ideology of the KMT regime also led to a strict limitation of the local language under the martial law.

At the end of 20th century, Taiwan started its democratization, the withdrawal of the martial law also brought a retaliatory rebound of local languages, especially the Hokkien language. Therefore, I want to know how diverse it is and how it changes during the past years. What I care most about is what it is now and the reason behind it.

Data

The palace where I can get the most exact data of the distribution of language usage is the official census. However, the Taiwanese census during the martial law period did not take the language into consideration, which makes it impossible to find out the number. While the most recent two times of census undertaken recently got the details of the distribution of language, so I took them as my original dataset, which is

the Taiwanese census in 2010[1] and 2020[2].

The only data I need for the analysis is the distribution of language using, so I search further for detail of the distribution of each county or city in Taiwan, then gather them into one csv file with different administrative division units separately by two censuses, then I got two files from two census containing the language distribution, which is the file “language_census2010.csv” and the file “language_census2020.csv”.

Once I got the data of language distribution, the next step is to select a proper method of presenting the digitalization, my requirement is to show how languages were distributed and changed during the past 10 years between 2010 and 2020, thus making it a good choice to create an interactive map with the detail of administrative division in Taiwan.

To achieve this, I need some extra data, besides the language distribution and usage percentage data I mentioned above from the official census, I also need the exact geographic data of Taiwanese administrative division, which I supposed that it can be download from the official department about geoinformatics of Taiwan, but I was wrong as their website kept returning me the sign of 404 during the several times of my try, so I choose to search the unofficial data from the GitHub as many same kind of projects were been uploaded here, then I got the geoinformation I want from a GitHub organization called “g0v[3]”, which is run by an organization that aims to improve the transparency and disclosure of the social and political information of Taiwan, where the

[1] Taiwanese census of 2010: https://www.stat.gov.tw/News_Content.aspx?n=2755&s=234399
[2] Taiwanese census of 2020: https://www.stat.gov.tw/News_Content.aspx?n=2755&s=230160
[3] GitHub organization “g0v”: https://github.com/g0v/

users can not only found the repository about geoinformation of Taiwan, but also various kinds of social dataset related to Taiwan, and there is a repository called "twgeojson [4] " inside of the organization which share the JSON file of the geoinformation of Taiwanese administrative division from 1982 to 2010, since the administrative division didn't change after 2010 in Taiwan, I chose the file "twCounty2010.geo.json" as my geoinformation dataset.

Digitalization

After all of that, I start my digitalization process, with the help of GitHub Copilot, I can easily ask the AI assistant to achieve what I want to make by just simply dropping the data to the assistant and providing a detailed prompt explaining what my development goal is. Due to the inertia brought by DHG504, I am used to create a digital map in HTML file through a python process, but this time I decided to ask the AI assistant to create HTML file directly, as in this way I can see the content inside of the file and make compares between each version of the HTML file me and the AI assistant changed the content.

So with all the data and methodology were chosen, I dropped the CSV file I gathered and the JSON file I download to the VS Code and ask GitHub Copilot to process them into an interactive heat map not only contain the distribution of language by each county and city, but also with another function that when user click each blocks of the administrative division, there will be a pop-up window that shows the language suing percentage in the exact administrative division.

---

[4] GitHub repository "twgeojson": https://github.com/g0v/twgeojson

Then it was when the first problem came up, I realized that the geoinformation data might have some mistakes because the division blocks shown in the first result are totally empty. After requesting the assistant, it informs me that due to the browser security restriction, the HTML file accessed by the browser can not load the JSON file, to fix that, the AI assistant converted the JSON file to JS file through a python program thus making it a loadable file to the HTML opened in the browser as a script label. Meanwhile, at the previous assignment, I met a problem that the data would get conflict caused by the difference of traditional and simplify Chinese, so an extra conversion code was added to the HTML file to ensure the consistency.

Consequently, I got the interactive heat map with the exact functions I want, which can show the languages distribution and usage percentage by each administrative division of counties and cities in Taiwan. Other function including showing the detail of the percentage for each language in the specific administrative units. Moreover, user can adjust the year of the map to compare the raise and fall of the language distribution between 2010 and 2020.

Then comes to the improvement of readability and other detail feature, for example, for each language, I will assign different shades of color to each administrative division based on the percentage of the language usage list in the CSV file. At first, the default grade is not suitable since it only has 5 classes and the gap between each grade is unequal, so I changed the HTML file directly to 20 grades by 5% of each class.

What's more, I also make another map included in the HTML file to show the dominant language in the island in 2010 and 2020, and due to the unique statue of

Mandarin, I also created another version of the dominant language map excluding Mandarin for fairness and objectivity.

Another small improvement is that when it comes to showing the distribution of each language singly, the window on the top right would display the exact number and percentage of speakers in the whole of Taiwan for that language.

Finally, after make sure that detail functions and the readability are as good as I wish, I ask the AI assistant to integrate all the information and functions into the HTML singly thus making it an offline-available and Independently operable HTML map. However, though I was satisfied, there are some limitations in this part that I will talk about later in the passage.

Limitation

First, I need to talk about the limitation of the data, as shown by two censuses, the total population of Taiwan is around 23 million people, but in the data, I used for the interactive map, the total number of populations in both language census is around 21 million, it is a sum of the language census for each 22 administrative divisions in Taiwan. This difference in the number of populations shows that there are some people who refused to answer or remained unanswered to the question or some there are some answers whose sample size was too small and were categorized as "other".

Next limitation is about a specific administrative, the “Matsu”(馬祖) county, before the KMT regime was expelled from the Chinese mainland, this small archipelago was part of the Lianjiang county of Fujian province of China, which uses another language called the Mindong language(閩東語). Narrowing down to the archipelago of

Matsu, people lives here speak the Matsu dialect(馬祖話), which is not one of any four languages mentioned above, but this language is only uses in the Matsu county, thus some special design were made in the window of this county, more specific, a new label of “Matsu dialect” with its percentage of speakers was added into it.

Another limitation is that for the map of the Austronesian languages, due to lots of reasons, the number of its speakers were in a very small amount, even in the county of Taitung, where got the highest percentage of Austronesian speakers only got 22.2% when it comes to 2020. Thus, making it might not that suitable to simply use the standards for other 3 language to present the language distribution map, especially when it comes to the factors like depth of color or the geographical division since most of the Austronesian aboriginal Taiwanese lives in the central mountainous area or the southeastern part of the island while there is no an independent administrative division for the mountainous area. For this consideration, it might be better to use a more detail administrative map with specific into the third class of the administrative division of Taiwan like township(鄉、鎮) or district(區), but there is no specific language distribution data matching them, so currently I was not able to make it.

Discussion

There are a lot of information can be observed through the map and it will take too many spaces to go through all of them. Consequently, I will only talk about some interesting finding.

First, there are only two dominant languages in Taiwan: Mandarin and Hokkien, but from 2010 to 2020, the population speaking Mandarin increased more than Hokkien,

thus making the percentage gap wider between two dominant languages.

Second, the population of the Hakka language is mainly distribute in the northwestern part of the island, more specific, the Taoyuan, Hsinchu and Miaoli areas (桃竹苗地區), that is the result of historical phenomena since for the early immigrants from coastal southern China, is more easy for them to get there and would also be a wise choice to live with fellow comes from the same place.

Finally, I want to talk about a special phenomenon I found from the map, that is the Matthew effect in the field of languages in Taiwan. From 2010 to 2020, we can see in the map that both the percentage and the population for Mandarin and Hokkien, which is the two dominant languages, are increasing while the Hakka and Austronesian languages are decreasing. What's more, the Matthew effect even happened between the two dominant languages, which I mentioned above that the percentage and the population gap is getting wider between the two dominant languages.

It is understandable since the most common languages would always be attractive for people as I always related to better social, economic or political opportunities, especially for the Mandarin, which not only is the official language of Taiwan but also get a strong influence all around the world.

But what about the difference in the situation between the two Chinese dialects of Hokkien and Hakka? Since after the end of martial law, each languages get the same protective measures, which makes the comparison that the Hokkien is increasing while the Hakka is decreasing is strange. I believe this is due to some political interference, since the Taiwanese government needs to strengthen the status and the use of Taiwanese

to emphasize the subjectivity of Taiwan as to reach some political goal, especially for the DDP government.

Conclusion

I truly realized the applicability that the digital method can bring to the humanities research. Through this project, I truly tried and learnt a lot of methodologies about digitalization in humanities field. At the beginning of this final project, I tried many forms just for test, besides the interactive map, I tried to train LLM with dialect translation function and create an online exhibition. I will save it as the ideas for later.