Process Book: Exploratory linguistic correlations

Alicia Howell and James Plante

May 2, 2022

1 Overview and Motivation

The languages of the world are divided into linguistic families where languages share similar traits, such as grammar or pronunciation. A commonly known language family is the Indo-European family which includes languages such as Celtic, English, German, and Italian. However, there are language families that contain many speakers but are less well known, such as the Sino-Tibetan family, which has 3,237,999,904 speakers today, and austroasiatic family, which has 116,323,040 speakers today [6].

The Sino-Tibetan family contains 453 languages, including Mandarin and other dialects in China, but many of the languages are not well documented and research often looks at subsets, such as isolating 50 langauges to research [9]. Due to the large size of this language family and the minimal sources for some of the smaller languages, we decided to focus on the austroasiatic family first, despite it having less speakers. The austroasiatic family is the tenth most spoken language family in modern day. It contains 169 languages, all of which have been documented before with some historical context [10]. Therefore, we believe the information to create a dataset containing the origin dates and locations of each language in the austroasiatic family is feasible. From this dataset, we aim to provide an exploratory visualization where users can investigate the correlation between languages in the family based upon when and where they were first spoken.

2 Related Work

In class, Professor Harrison had mentioned currently working on some linguistic projects, which got us thinking about fNIRS research in the lab that uses brain patterns to detect someone's familiarity with a language. We wanted to investigate an adjacent field to this to understand how closely some languages may be related as this could affect how the brain perceives it. This is what inspired us to map out how languages have migrated over time and parent-children relationships. While we had to pivot the scope rather early in the project, we still did a substantial amount of background research on this idea.

Maps are a powerful visualization strategy to display data in relation to location. There are a variety of methods to do so; previous research has visualized linguistic data on maps through use of open-source programs, accessible APIs, and traditional cartographic methods [11]. presents various methods of putting data onto maps. This includes different open-source programs, accessible APIs, and the original cartographic methods.

While a common method to mapping the migration of languages is through documenting the times and regions it was spoken in, researchers have also made a connection between genetics and language families [7]. They showed that in Southeast Asia, genetically similar populations often spoke languages within the same families, including Austroasiatic, Sino-Tibetan, Hmong-Mien, and Austronesian. This contributes to efforts to reconstruct the human genetic history in Southeast Asia but can also assist with determining historical paths that individual languages took and how they evolved.

Worth noting is the potential ethical problems surrounding migration maps of human populations. Migration maps for immigrant pathways have shown that due to time, funding, and other biases, many migration maps often omit valuable information [3]. This can present as specific populations not being included in the maps or certain regions receiving better or worse research to ensure correct representation. Adams (2018) developed guidelines to ethically visualizing mobile populations. While this certainly has greater impact on refugee and asylum populations, it can also affect the research we are conducting. We chose to focus on Eastern languages because of the normalized focus on Western history, but this leads to an inherent lack of details in the data we have access to.

The classification of languages can be difficult to solidify and a researcher's classification decisions are often disputed by others. Unlike fields such as genetics, which has a relatively easy code to compare for relationships (DNA), where one language family ends and another starts is often a blurred line centered around grammatical structure, similar linguistic prefixes or suffixes, and common words. The difficulty here is that there is no way to take a sample of a language and parse it down into all of its components, as one can do with a genome. When comparing sets of words between languages, one has to make a distinction between cognates (words that sound/look similar with similar meanings) and loan words (words from one language adopted into another with the same meaning) [8]. Loan words mean that two or more languages have had significant interaction in the past, but did not necessarily come from the same parent language or are in the same language family, while cognates are generally shared between words that are within the same family. For example, the word for bread in Spanish and Portugese is 'pan' and is considered a cognate between the two languages (which are both in the Romantic family) while 'pan' also means bread in Japanese. Japanese is not truly related to Spanish or Portugese, however, due to Western sailors visiting Japan hundreds of years ago and introducing the grain wheat, along with its main use to make bread, introduced the word 'pan' into the Japanese language.

Part of the reason why data is so scarce on Eastern languages is because as

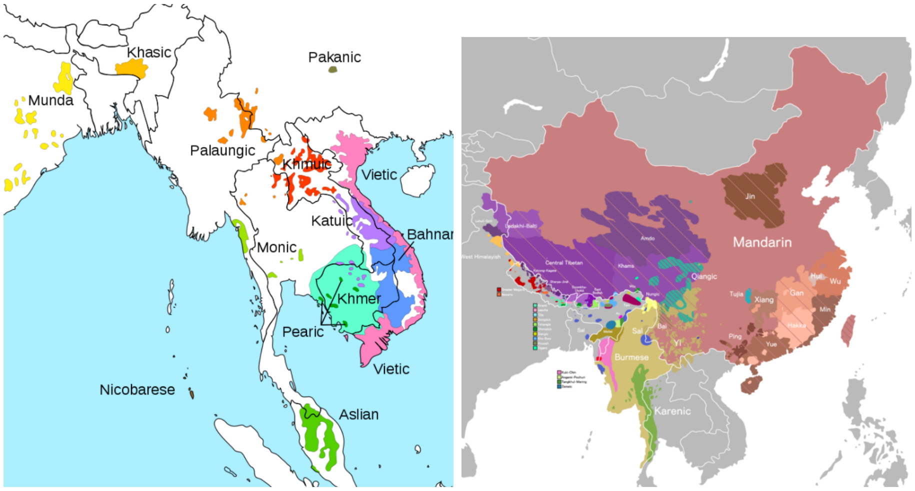


Figure 1: The geographical distributions of subdivisions in the (**Left**) austroasiatic family [1] and the (**Right**) Sino-Tibetan family [2].

recently as 2015, researchers were trekking through the thick forests of Eastern Asia to document locations of endangered languages in the Sino-Tibetan family (Figure 1 Right) [11]. An endangered language is one in which there are fewer native speakers in the younger generations than there are native speakers passing away. Commonly, a language becomes endangered because the younger generations are learning a language that is more widely spoken in order to receive an education or get jobs outside their local community. Many of the languages in the Sino-Tibetan and austroasiatic (Figure 1 Left) families with few native speakers, even if they are not endangered, are in communities that do not have access to technology such as computers where researchers could ask them to fill out a digital survey.

While there are sources that list out the history and geographical distributions of the Sino-Tibetan [9] and Austroasinatic [10] families, they are not completely inclusive works. Typically, this papers or books focus on specific subdivisions of the language families or are too old to be entirely reliable, as new communities of speakers have been found in the seventy years since [10] was published. This stunts the depth of exploratory research that can be performed with these under-represented language families; nevertheless, it is important that work is done with these language families to broaden awareness of the lack of representation and robust datasets available for large percentages of the global population (as the Sino-Tibetan and austroasiatic families are in the top ten most spoken families [6]).

3 Questions

A general question we are trying to solve is "How does the visualization of linguistic data impact the type of interpretations that can be made?" Originally, we had aimed to do this specifically with the migration of languages over time and how they evolved from a parent to child language, such as French, Spanish, and Italian evolving out of Latin. In addition, we wanted to focus on languages that were less studied, such as East/Southeast Asian languages.

Our main question remained intact throughout the project, however, the specifics changed. First, we had wanted to investigate the Sino-Tibetan language family as it has a large population of speakers but little research. This ended up being problematic because we could not finding lists of even the languages within the family let alone more descriptive data like their history and current speakers. We pivoted our question to focus on the austroasiatic family, as it had fewer languages within it but was still within the top ten linguistic families. While the austroasiatic family is better documented, there are still a great number of descriptive details missing. We were able to determine around half of the reported languages within the family, partially because different sources disagreed with which languages are within the austroasiatic family. There was also little historical data about when the different languages, or even just subdivisions of the family, had originated.

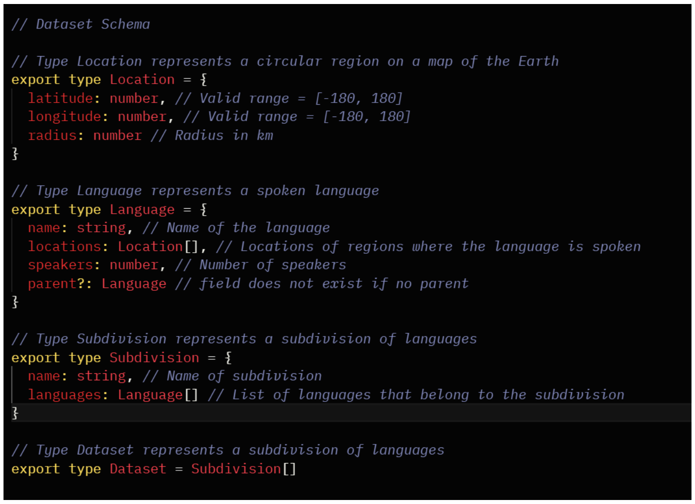
We changed our research question from focusing on the migration and evolution patterns of language to the modern spread and demographic relationships. We were interested in visualizing which languages are in geographically similar locations and how that relates to their subdivisions. Questions that can be investigated from this are "Do languages within the same subdivision remain in the same area?", "What subdivision has languages that are currently the furthest apart?", and "Are there predominant languages based upon region, subdivision, or both?"

4 Data

We collected the data from a variety of sources and parsed together our own dataset. Our primary sources were Britannia [5], linguistic encyclopedias [6], and looking at citations on Wikipedia pages [1]. The features we looked for were the subdivision, spoken locations (longitude, lattitude, and radius), number of native speakers, endangerment status, and known linguistic parents or children. We ended up not using the linguistic parent and child field because the available data was extremely sparse in that regard.

We organized the data first in a table where each column was a feature and each row was an individual language. Then, we transferred it to a .json file (Figure 3). The .json file had the following organization originally:

- Family
  - Name



```
// Dataset Schema

// Type Location represents a circular region on a map of the Earth
export type Location = {
  latitude: number, // Valid range = [-180, 180]
  longitude: number, // Valid range = [-180, 180]
  radius: number // Radius in km
}

// Type Language represents a spoken language
export type Language = {
  name: string, // Name of the language
  locations: Location[], // Locations of regions where the language is spoken
  speakers: number, // Number of speakers
  parent?: Language // field does not exist if no parent
}

// Type Subdivision represents a subdivision of languages
export type Subdivision = {
  name: string, // Name of subdivision
  languages: Language[] // List of languages that belong to the subdivision
}

// Type Dataset represents a subdivision of languages
export type Dataset = Subdivision[]
```

Figure 2: An early example lines of code for transferring data from the table to the .json file for the linguistic data that was used to populate the visualization. This was updated in further versions of data wrangling.

  - Subdivision

- subdivision
  - Name
  - Languages

- Language
  - Name
  - Family
  - Subdivision
  - Status (for endangerment)

- Area
  - Name
  - Language
  - Locations
    - Latitude
    - Longitude
    - Radius
  - Population

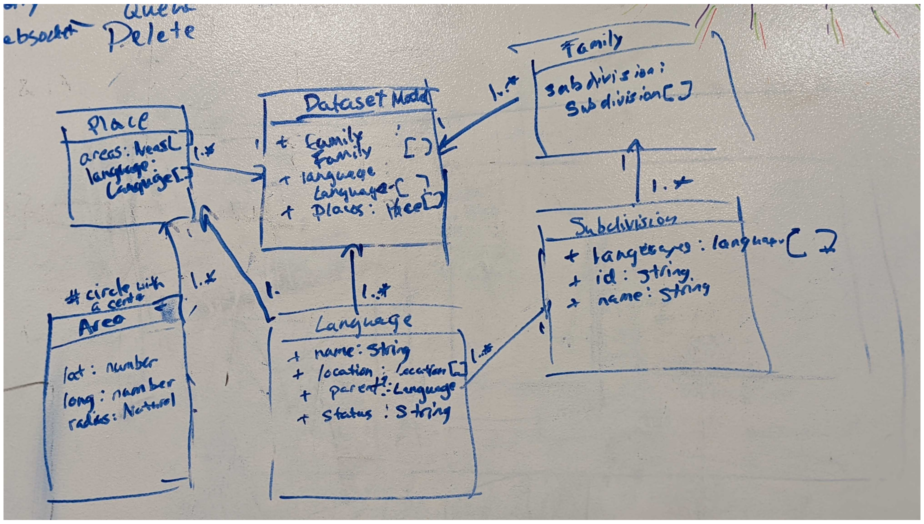
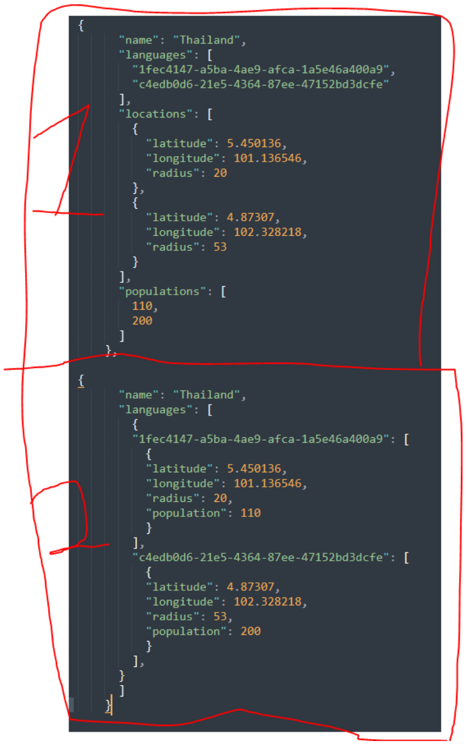


Figure 3: The data schema that shows how each set of data connects together.

However, as we started to code the visualizations, especially the network, we realized that this format was not suitable, primarily due to the "area" category. We had multiple areas with the same name as they were originally categorized by language. In addition, we need to add in a category for the links of the network that would include a [source] and a [target] node. We had two ideas of how to change the area category to work better within the data structure. We decided to go with option two shown in Figure 4. In list format, that looks as follows:

- Area
  - Area Name
  - Languages
    - Language Name
      - Latitude
      - Longitude
      - Radius
      - Population

From this format, we were able to map the nodes and links for the network and add that to the .json file. We drew out a database schema for the dataset to be able to check how the keys and values would interact with each other (Figure 3). This was an important step to ensure the data did not self-reference itself and to map out the logic of how to connect the different visualization components.



```
{
      "name": "Thailand",
      "languages": [
        "1fec4147-a5ba-4ae9-afca-1a5e46a400a9",
        "c4edb0d6-21e5-4364-87ee-47152bd3dcfe"
      ],
      "locations": [
        {
          "latitude": 5.450136,
          "longitude": 101.136546,
          "radius": 20
        },
        {
          "latitude": 4.87307,
          "longitude": 102.328218,
          "radius": 53
        }
      ],
      "populations": [
        110,
        200
      ]
    },

{
      "name": "Thailand",
      "languages": [
        {
        "1fec4147-a5ba-4ae9-afca-1a5e46a400a9": [
          {
          "latitude": 5.450136,
          "longitude": 101.136546,
          "radius": 20,
          "population": 110
          }
        ],
        "c4edb0d6-21e5-4364-87ee-47152bd3dcfe": [
          {
          "latitude": 4.87307,
          "longitude": 102.328218,
          "radius": 53,
          "population": 200
          }
        ],
      }
      ]
    }
```

Figure 4: Two different configurations for linguistic regional information. The top one uses ordering to connect the different variables, so the first language in the ”languages” array would have the latitude, longitude, and radius of the first element in the ”location” array and the first number in the ”population” array. The bottom one gives each language within the region its own array to record the location and population information, which is the one we went with.

5 Exploratory Data Analysis

First, we looked at how regional linguistic data is usually represented, especially when maintaining the familial hierarchy of languages. This fell into two categories for the most part: 1) map layouts with with a colored overlay to represent which regions a language is spoken in (Figure 1) and 2) trees to denote the structure of the family (Figure 5). Visualization 1 typically does not capture details of the language family, such as which languages are within the same subdivision while visualization 2 does not capture regional information. Both visualizations typically lack context for how many native speakers there are of a specific language as well as migration patterns of the language.

We used these visualizations to brainstorm how we would display our data set with the additional information that neither of the previous visualizations typically contain. While we had not managed to procure historical information

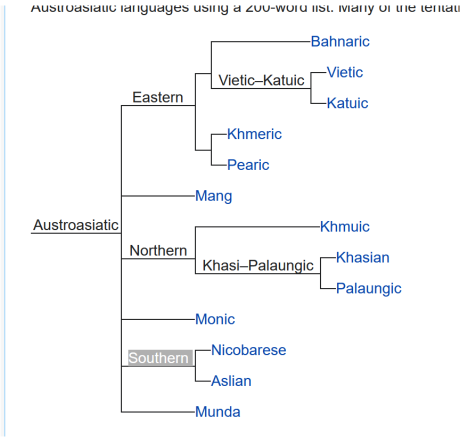

Austroasiatic languages using a 200-word list. Many of the tentati

Figure 5: An example of one theorized configuration for the austroasiatic family tree. This example separates the languages first into a regional distribution, then into the subdivision and finally to a single language.

for each language, we were able to find information regarding how many native speakers there are. We decided the best approach would be to have two separate visualizations that were interconnect, essentially combining visualizations 1 and 2 from above. This way, we could maintain the type of information both convey as well as make simple adjustments to include population data.

6 Design Evolution

Our initial design featured a primary map in the center that contained all of the data paired with a small multiples view underneath that visualized data specific to different regions across periods of time (Figure 6). This design would have included a scroll bar along the bottom of each small multiples pane to change the period of time that was visualized. Each language would have been represented as a colored dot that had a radius proportional to the number of native speakers and would move according to its migration route or split into more dots if some speakers traveled to a new region while others remained in the original one. In addition, a language could spawn from another one for languages that had a parent-child relationship in the defined time period. Due to the lack of research for the austroasiatic language, we were not able to implement a visualization that was designed for showing migration patterns across time, and thus needed to pivot scope and design.

Our second design is relatively the same as our final implementation, though we made modifications to make it more understandable. This design consisted of four components that were all interconnected (Figure 7 Left). The visualizations were a map of the region related to the austroasiatic family as well as a node network that displayed the relationship between members of the family. Then,

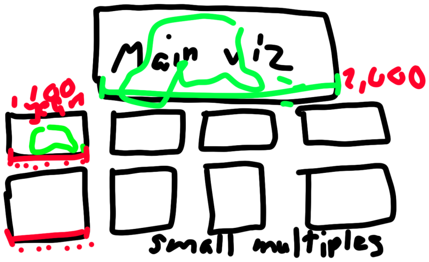


Figure 6: The original design plan for the language visualization when we were still planning to integrate linguistic migration data. The top box was the main visualization component which featured a map of the region of interest, in this case Southeast Asia. There would have been a scroll bar at the bottom that spanned across time. This scroll bar would affect all visualizations synchronously. The bottom component would be a section of small multiples where each box would represent a subregion. Originally, we had planned to make these square regions instead of specific countries since most austroasiatic languages are older than the country borders in Southeast Asia.

there was a static table that listed the subdivisions within the family as well as a dynamic table that lists the languages relevant to the user's exploration.

In this design, we wanted the map and the node network to be interconnected, so when one gets updated, the other does as well. For example, if the user were to select a country within the map, it would update the center table with languages in the country and the node network would populate the nodes and links for the languages in the country. A similar process would happen if the user selects a subdivision from the bottom table as well. We went through a few ideas of how the node network should be constructed, such as what should be set as the parent node for each display. We decided it would be best to have two different node and link lists for the node network (Figure 7 Right), one that is connected to the map and one that is connected to the subdivision, since there are countries that contain languages from multiple subdivisions as well as subdivisions that are spread across multiple countries.

Finally, we went through a variety of design ideas for populating the languages on the map. Our original idea was to just have each language be represented as a small dot, as would have been the case in the migration visualization, with an on-hover event that would display descriptive data about the language, such as endangerment status and number of native speakers (Figure 8). Each language would be a separate color as well. We decided that while this method was possible and would contain the information we had collected, too much of the data has hidden by the on-hover event.

In the end, we decided that adding more features to the dots that represent a language on the map would be the best solution as well as keeping the on-hover event. We changed the color scheme so that each subdivision had a subset of related colors; for example, Vietnamese and Muong both belong to

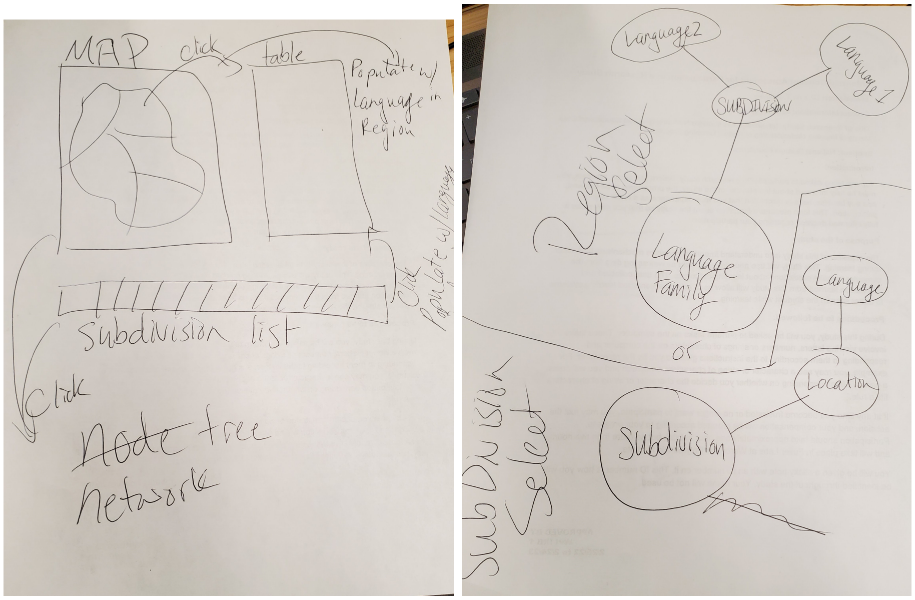


Figure 7: **Left:** A sketched layout of the design for the second plan. It features the main map on the left which can be zoomed in or out and has selectable countries. To it's right is a table listing the languages within the selection and underneath is a selectable table of language subdivisions. When the user selects a country or subdivision, the node network at the bottom would update to reflect the related languages and their connection to each other. **Right:** Two views of the node network, one for if the user selected the region (top) and one for if the user selected a subdivision (bottom). Having two different node network displays would allow for multiple countries to be listed when a subdivision is selected and multiple subdivisions when a country is selected.

the Vietic subdivision and thus one could be a bright yellow while the other is yellow-orange. In addition, we changed the opacity of the dots to represent the number of native speakers for the language in that region (since languages can have multiple regions). The higher the opacity, the more speakers there are. Finally, we changed the on-hover event to change the size of the dot to be proportional with the size of the region the language is spoken in. Originally, we were going to make each dot the size of the area it was spoken in without restricting it to just be an event, however, this would cause a lot of overlap between language regions and make it more difficult for the user to hover over a specific language to get more details.

While only one version of the node networked ended up (successfully) coded, we had gone through a couple of examples in Observable with our dataset to test what kind of node network we wanted. We found the examples provided by Mike Bostock (https://observablehq.com/@mbostock) to be very valuable and

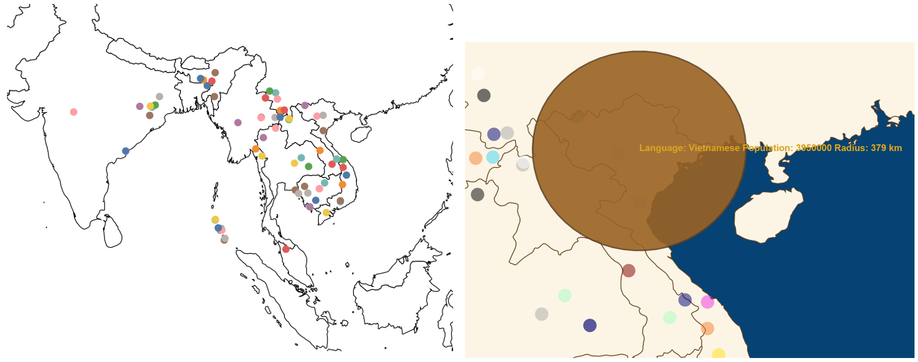


Figure 8: **Left:** The zoomed out view of the map where each language is a separate color. There is not a specific pattern in which the languages are colored e.g. by subdivision. **Right:** Zoomed in view when a user hovers over a specific language. Data about that language is displayed, such as its name, location, and number of speakers.

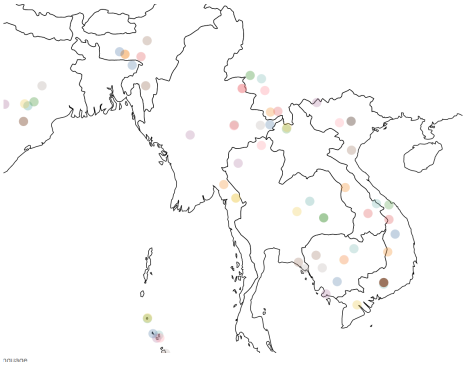


Figure 9: The updated visualization where the opacity of the language's dot reflects the number of native speakers. Most notably is the opaque brown dot in south Vietnam which represents four million Vietnamese speakers.

a good place to start for exploring the different kinds of node networks available. The first one that was of interest to us was a node network that uses Graphviz to structure the shape of the network. Bostock's example [4] he used Graphviz to map the structure of the United States which was fairly accurate, except New England was flipped around (Figure 10).

We used Bostock's Observable code to create a note network that contained the primary language family (austroasiatic), the subdivisions, the individual languages, and the regional locations for each language (Figure 11). We had hoped that Graphviz would have been able to expand the node network out to distinctly show relationships between geographic regions and subdivisions, however, a much larger SVG would have been needed to do so. In Bostock's

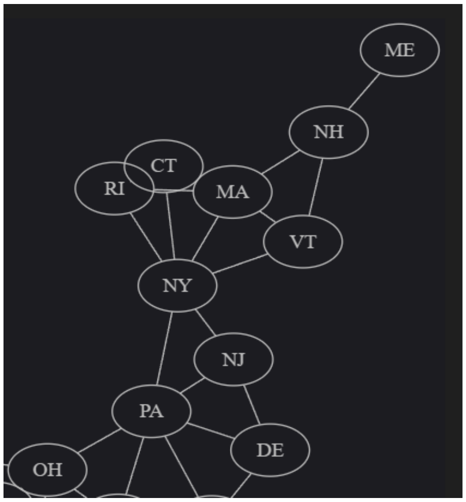


Figure 10: A snippet of Bostock's visualization of how the United States is mapped out using Graphviz [4].

example, there were only fifty nodes (one per state) while we had nintey-three nodes (one per language (70), subdivision (12), region (10), and linguistic family (1).) This resulted in a very cramped node network that was difficult to read (Figure 11 Left) and non-interactive. If Graphviz had included a function to move and reorganized nodes, then the density of the network may have been fixable. However, Graphviz is designed to interpret the locational relationship between all of the nodes, and thus is a static graph.

We trimmed down the dataset to two countries, four subdivisions, and the thirty-five languages within those subdivisions to investigate how a smaller version of our dataset would be represented. In postprocessing, we colored the language family green, the subdivisions blue, and the countries yellow (Figure 11 Right). While with fewer subdivisions we are able to differentiate clusters better, part of what makes this visualization interpretable is the additional of categorized color. Because of the color, we can easily see that one country has only one subdivision, and every language of that subdivision belongs to that country (top left). However, the second country (bottom right) has three different subdivisions that connect to it, with only one language from one of the subdivisions represented. That subdivision in particular (Palaungic, top right) is one of the larger ones in the austroasiatic family with at least sixteen unique languages. Finally, it is difficult to tell how many languages from the bottom left subdivision belong to the bottom country, though it is safe to say at least one. Because of this spacing difficulty with Graphviz, we decided to use other examples of node networks, such as the code provided in class.

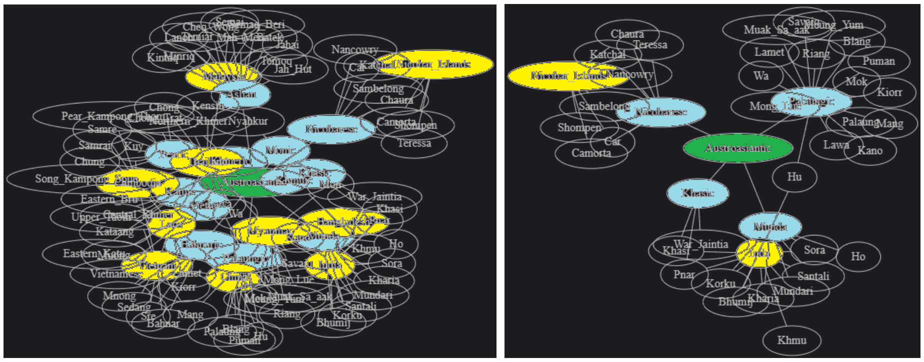


Figure 11: **Left:** The node network with 93 nodes. The center is the austroasiatic family node (green) which is linked to each subdivision (blue). Each subdivision node is then linked to the languages the belong within that subdivision which then link to the regions (yellow) they are spoken in. Multiple languages from different subdivisions can be spoken in the same region, cause the node network to contain many overlapping lines and be difficult to read. **Right:** A trimmed down version of the node network that contains only two countries and four subdivisions. We selected these to countries to specifically showcase the two ways they are linked. The Nicobar Islands node (top left) is only linked with languages that are within the Niccobarese subdivision, creating a distinct cluster. India (bottom right) has languages from three different subdivisions, though in the Palaungic subdivision only one language is spoken in India. This causes the node network to form two different clusters with one language (Hu) connecting the two clusters.

7 Implementation

Figure 12 shows the final implementation of the language exploration tool on the server. On the left side of the UI is an interactive map that the user can zoom in or out on and click-and-drag to move around. The map only features countries in the region that the Austroasiatic is spoken, which is primarily Southeast Asia and parts of China and India. The right features the node network display which by default shows every language family, subdivision, and individual language until the user makes selections. Underneath the map component is a drop-down box to select a specific subdivision to explore in more detail.

Figure 13 is a zoomed in view of the map component on the left of Figure 12. In this example, the user has selected the country of Thailand, which turns green to give feedback to the user that it is selected. The drop-down list of subdivisions becomes limited to only the subdivisions within Thailand and the node network on the right of Figure 12 is updated as well. If the user selects a separate country, Thailand will no longer be highlighted in green and the new country will. Not shown is a function similar to Figure 8 Right where an event deploys on-hover when a user puts their cursor over a language. The dot expands or

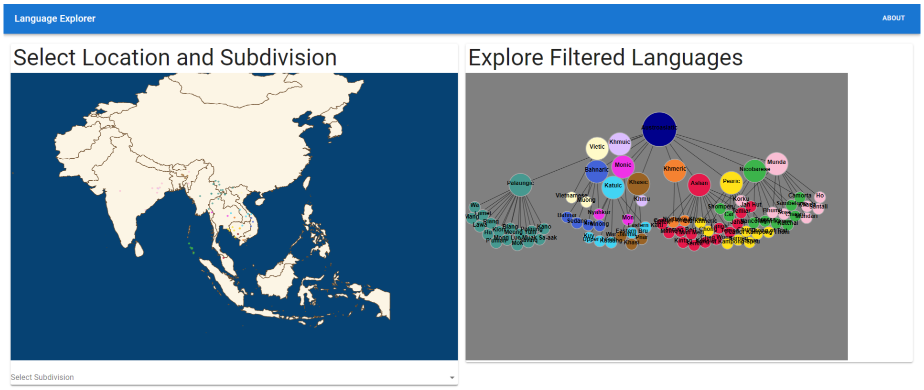


Figure 12: A full view of the user interface when a user first opens up the server. **Left:** A regional map that displays selectable countries and is populated with dots that represent individual languages. The dots are color-coded according to subdivision. **Right:** The node network visualization in the state it would populate in when the user first connects to the server.

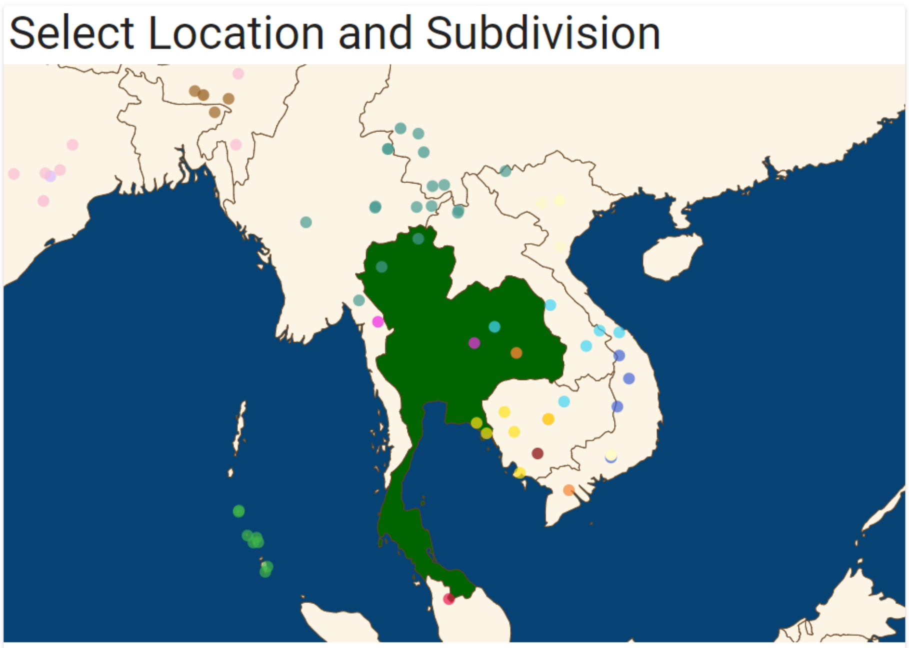
contracts to represent the general area the language is spoken in and provides more descriptive data about the language.

Figure 14 is the view of the node network when Thailand is selected in the map. Six different subdivisions are spoken within Thailand with one or two unique languages being spoken per subdivision in the country. In the node network, the subdivisions and their child languages are the same color to provide easier identification when the nodes are more densely compacted. In addition, a user can click and drag any of the nodes to rearrange them which is useful in instances if the individual languages are crossing links between subdivisions and one needs to visually separate them.

The final interactive component of the visualization is the subdivision drop-down box in Figure 15. The subdivisions are organized in alphabetical order and the drop-down repopulates when a country is selected in the map view to only allow selecting subdivisions within that region. If a user selects a subdivision from the drop-down box, the node network will update to display just that subdivision and its children. In addition, if a user selects a country and then selects a subdivision within that country, the node network will populate only with the languages that subdivision contains within that specific country. This provides users with a way to isolate individual subdivisions in a region-oriented approach.

8 Evaluation

Through this visualization we are able to answer the questions posited in section 3.



Select Location and Subdivision

Figure 13: A zoomed in view of the map in the visualization. Thailand is highlighted in green because it was selected by the user. We can see languages with different opacity spread across the region as well as languages from the same subdivision (denoted by the same color) crossing borders. The ocean is colored blue to help separate it from land masses. There are countries, such as Nepal and Indonesia, that are featured on the map but have no languages connected to the Austroasiatic family. These countries can still be selected and the other components (node network and subdivision list) will refresh to be blank.

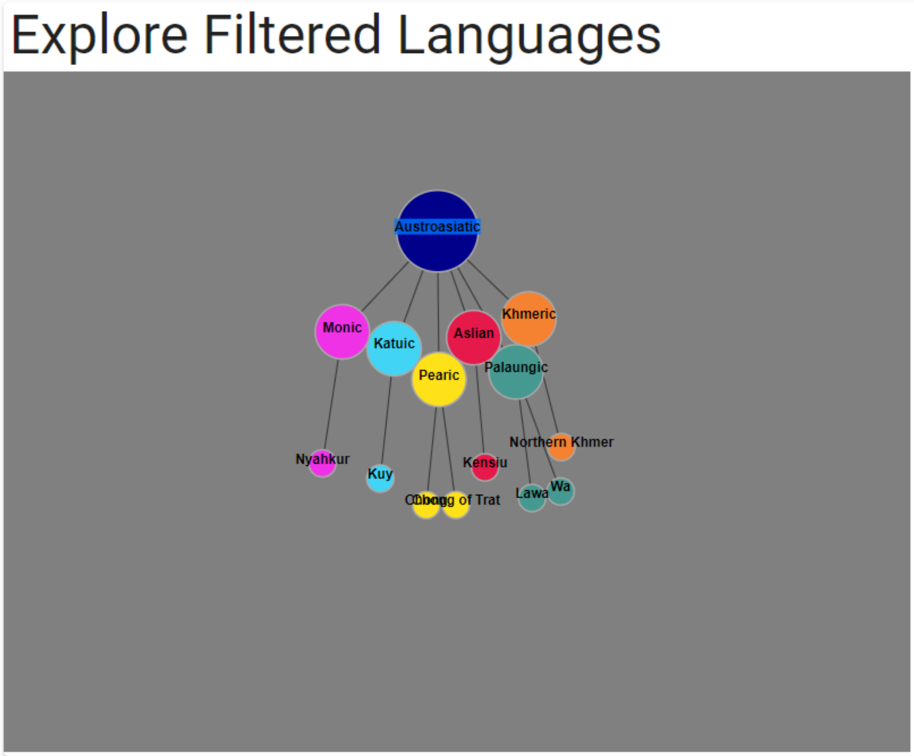


Figure 14: A view of the node network. The language family, Austroasiatic, is at the top in dark blue. The color and position of this node remain the same no matter what else is populated inside of the network. The nodes that link directly to the language family are the subdivisions. In this figure, Thailand was selected in the map view and thus only the subdivisions within Thailand are exposed here. This colors remain consistent across networks as well; if the user selected Myanmar, which also contains a language from the Monic subdivision, those nodes would remain pink. The individual languages within a subdivision are the same color as their subdivision to make it easier to group related languages visually.

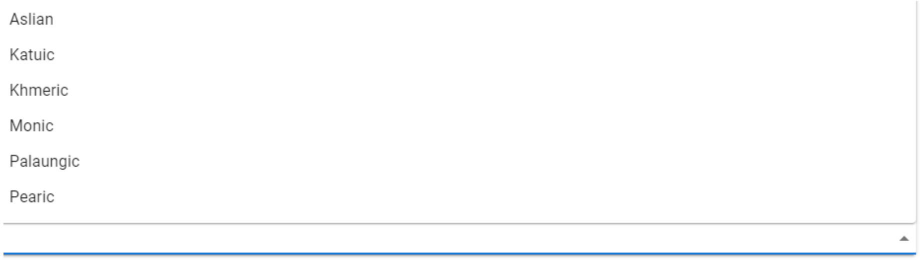


Figure 15: The drop-down list of subdivisions. This figure shows the subdivisions within Thailand because it is selected in the map view. By default, all subdivisions are listed in alphabetical order upon loading the server. A user can select a subdivision here to adjust the display in the node network.

- Do languages within the same subdivision remain in the same area?
  - Yes and no. In general, most languages within a subdivision are clustered together, though on the rare occasion, such as the Monic subdivision, the languages can be spread far apart. Often, however, are languages that cross country borders but are still within close proximity to other languages in their subdivision. This can be seen in the map view when zooming in on border regions, such as the border between Thailand, Myanmar, Laos, and China all containing Palaungic languages (Figure 13. This would have been difficult to answer if we had only used the node network/family tree visualization per country.
- What subdivision has languages that are currently the furthest apart?
  - The Munda subdivision likely has the languages that are spread the furthest apart due to Korku being located in central India while Ho and Santali have speakers in Bangladesh. However, we were not able to find locations and populations for all languages in the Austroasiatic family, only approximately half of them, so this answer may change as more languages are added.
- Are there predominant languages based upon region, subdivision, or both?
  - Yes, as noted before, subdivisions are often clustered in the same area. Notably, Malaysia contains almost all of the languages in the Aslian subdivision. Additionally, if the user hovers over the dot for Ho in India or the dot for Vietnamese in northern Vietnam, they can see the wide geographical distribution of these languages as well as the large number of native speakers.

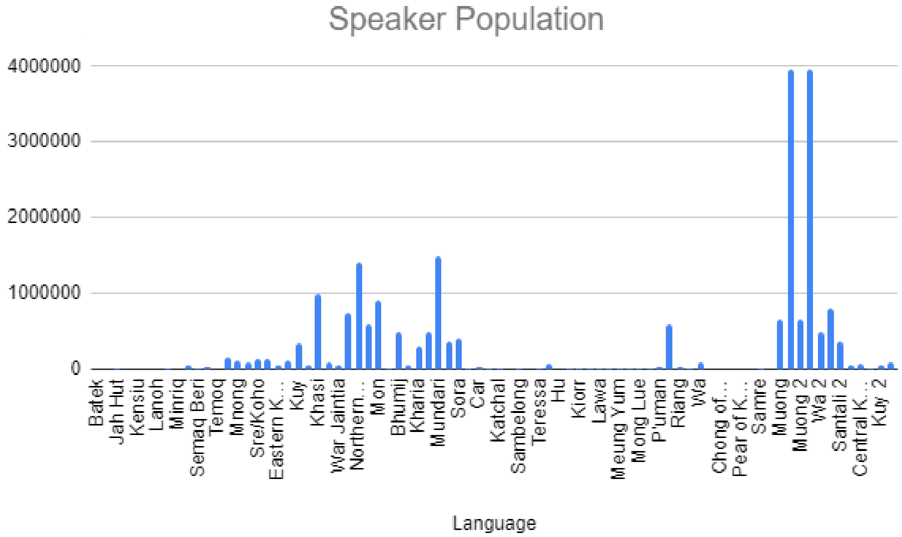


Figure 16: A bar chart of the number of speakers per language. The languages are split by region, so if a language, such as Vietnamese, is spoken in multiple regions it will have two separate bars in the chart. Most languages have less than 1,000,000 speakers per region with only four having greater than 1,000,000.

8.1 Improvements

One improvement is to have the opacity be measured in a logarithmic scale. Currently, it is a direct correlation between the opacity and the population of native speakers. However, we have up to four million speakers in a single are and as few as 50 speakers for another. This caused most of the dots to be relatively the same opacity as the majority of languages have fewer than a million speakers (Figure 16).

Future improvements should add in the ability to deselect countries in the map component and select multiple countries. We would also like to include the bottom node network in Figure 7 Left so that when selecting a subdivision, the network populates which countries have speakers of the subdivision. Additionally, the user should get some form of feedback about the name of the region they have selected, either appearing in the title, on-hover of the country, or in the node network.

References

[1] Austroasiatic languages (Apr 2022), `https://en.wikipedia.org/wiki/Austroasiatic-languages`

[2] Sino-tibetan languages (Apr 2022), `https://en.wikipedia.org/wiki/Sino-Tibetan-languages`

[3] Adams, P.C.: Migration maps with the news: Guidelines for ethical visualization of mobile populations. Journalism Studies **19**(4), 527–547 (2018)

[4] Bostock, M.: How graphviz thinks the usa is laid out (Jan 2022), `https://observablehq.com/@mbostock/how-graphviz-thinks-the-usa-is-laid-out`

[5] Diffloth, G.: Austroasiatic languages (May 2018), `https://www.britannica.com/topic/Austroasiatic-languages`

[6] Eberhard, D.M., Simons, G.F., Fennig, C.D.: Ethnologue: Languages of the World. SIL International, Dallas, Texas, 25th edn. (2022)

[7] Kutanan, W., Liu, D., Kampuansai, J., Srikummool, M., Srithawong, S., Shoocongdej, R., Sangkhano, S., Ruangchai, S., Pittayaporn, P., Arias, L., et al.: Reconstructing the human genetic history of mainland southeast asia: insights from genome-wide data from thailand and laos. Molecular biology and evolution **38**(8), 3459–3477 (2021)

[8] Rogers, J., Webb, S., Nakata, T.: Do the cognacy characteristics of loanwords make them more easily learned than noncognates? Language Teaching Research **19**(1), 9–27 (2015)

[9] Sagart, L., Jacques, G., Lai, Y., Ryder, R.J., Thouzeau, V., Greenhill, S.J., List, J.M.: Dated language phylogenies shed light on the ancestry of sino-tibetan. Proceedings of the National Academy of Sciences **116**(21), 10317–10322 (2019)

[10] Sebeok, T.A.: An examination of the austroasiatic language family. Language pp. 206–217 (1942)

[11] Zastrow, M.: Data visualization: Science on the map. Nature **519**(7541), 119–120 (2015)