Where did that word come from?

Andrei Dumitrescu
Jason Madeano

1 Introduction

For the A5 visualization, our team was particularly interested in studying the relationship between words in different languages and how they are connected around the world. More specifically, we were interested in working with the World Loanword Database (WoLD). This database includes a plethora of different aspects of language that we found very useful to include in our visualization. For example, indicators of how long ago a word was borrowed, indicators of sources and so on. Our main variable of interest however was the borrowed score of a word. This metric gave us a representation of how likely words across many languages representing a given concept are to be borrowed.

In order to introduce this to the viewers of the visualization, we explain the dataset and key concepts behind it at the top of our webpage. We guide them into a set of interactive bar charts which show the average borrowed score of a category which is then linked to the borrowed scores of concepts/words within that category and then how long ago those words tended to be borrowed across languages. This chart provides the reader with an option to select the category they want to inspect by hovering over the bar, which updates the other barcharts like a drill down.

Our main goal for this overall visualization is to make the general audience interested in the connections between languages, and show a first glimpse of the geographic variation in where languages come from. This is why we decided to use a map of the entire world and plot the individual languages with their respective latitude and longitudes, as well as why we added the lines connecting languages upon a hover from the user. This way, the user can have complete control over the experience that they gather from our visualization and learn about whatever kinds of languages or groups of languages that they desire. With this set up, we also aimed to stress just how complex the languages of the world are to the user.

At the same time, we want the user to have a benchmark to interpret some metrics related to languages that they are viewing. This is why we chose to do bar charts showing comparisons between different ideas, different words and different scores. We felt that these bar graphs were the best suited to convey our message from the data we used since the height differences help the reader quickly distinguish between scores and words. The bar charts are stacked one next to each other in order for the reader to be able to absorb all of the information and supplemental information provided by the charts as quickly as possible.

Alternatively, we know we could have achieved this goal with other visualizations such as a square chart where the age score was a gradient for each word and showed up as the color of the square and the borrowed score could have been represented by the size of the squares for each word. However, we felt that this approach would not have been precise enough and as contrasting as the bar charts are.

2 Related Work

This work was largely inspired by our love of languages and desire to share the diversity of languages with a broad audience. We did not base our work off other language visualizations, and to our knowledge we are the first group to create a set of visualizations based on the WoLD dataset. We hope that our work inspires others to explore the dataset even more thoroughly.

3 Methods

Data Cleaning: A great feature of the WoLD dataset is that it was collaboratively curated by 41 different authors/experts in their respective languages. However, this also presented a challenge: there are many inconsistencies in the dataset. Some authors left certain fields/columns blank while others wrote meticulous details. This limited our analysis since we only wanted to focus on fields that were available for the majority of the words/borrowed sets in the dataset. The dataset was also broken up into many different tables but there was very limited documentation regarding what fields within each table represented and how the tables should be linked (foreign/primary keys). This made data cleaning an arduous process. All of the code used to clean and transform the data is included in the repository to help future data scientists that want to visualize this dataset.

Interactions: With regards to the system and design of our visualization, we tried to avoid frictions in the user experience and make it as easy for the user to interact with the data that we presented. Therefore, we took the following design decisions:

- **Hover:** There are 3 different actions for when a user hovers over something of importance on our webpage. When the mouse hovers over a category in one of the bar charts at the top, this acts as a filter and highlights only the concepts that constitute that category. When the mouse hovers over a specific language, the user can see a line physically connecting it to where the language originated from as well as a tooltip telling the reader what they are viewing. As an alternative, we considered enabling a clicking option instead of hovering but we thought that hovering would be more natural for the user to discover whereas clicking introduces some difficulties for users who are not as enthusiastic about the visualization. The last hover functionality is when a user hovers over a specific region, the tooltip for this action will simply show the reader the name of that country/region. We thought that visualizing these geographic locations and connecting them to their origin as well as the information from the bar charts above would be enough to quickly leave the reader of an impression on the language landscape they are choosing to look at.
- **Tooltip:** We decided to include a tooltip with the name of the language (as explained above).
- **Zoom:** We enabled a feature to zoom in the map once the users scroll with their scroll wheel on the map. As well as an option to click a button and reset the level of zoom on the map. This allows the user to see better the languages, especially in regions that represent clusters of language hotspots. The hover feature keeps working as when seeing the whole world map and the points/lines dynamically resize upon zooming.
- **Slider Bar:** Our final feature that we have added to this visualization is a slider bar. For some context, there is a section of our database that is labeled age score, which provides rough estimates of how long ago the borrowing of a particular word occured. Based on this, we also wanted to add a time aspect to our

visualization and show how exactly languages have changed over time. We put a slider bar on the top map that allows users to drag the bar and filter out the horizontal lines and the languages that exist on the map. This may also help to make the map more engaging since it reduces the number of points for the user to focus on. With that said, we also included a checkbox which disables the slider and allows the user to see all of the points at once.

4 Results

Our chosen problem for this final project was not a direct problem that we could showcase with a certain amount of visualizations that answer a question at the end of it, instead our goal of tackling the complexity of language was much more difficult to do in a manner that came up with a solution. Our overall goal for this project, as we have said throughout our paper and throughout our project, was indeed to enable the reader and the user to explore the languages of the world at their own interest. As a result, the problem that we chose to tackle is simply the problem of illuminating the history of language and the history of how languages across the world have interacted over time. In order to tackle this problem we have produced two major visualizations with key components that help us achieve this goal. Our main visualization that is produced from this overall project is that of the map. This tells a different story for the reader every time they hover over a point. The visualizations that are produced from the map are our main way of communicating the history aspect of language. Our second key visualization that is produced from our project is that of the bar charts. These interconnected and filtered bar charts can again produce many different visualizations and each one tells a story based off of the selected filter. These bar chart visualizations help us tell the story of how connected each language within the region and its ideas are to its surrounding counterparts.

5 Discussion

We hope that the audience leaves our visualization with a better appreciation for the many threads that connect the languages around the world. To our knowledge, this idea of linking languages and focusing on borrowed words in a visualization is novel, but we believe useful for better understanding where words come from. Additionally, while the visualization only contains roughly 400 among the more than 7000 languages spoken around the world, we believe that this exposure is still important. We don’t expect the audience to be aware of the many (most) of the languages in the visualization, so even just displaying them in this way can provide a platform and spark interest in languages that otherwise might be dying out.

Future Work: As we have been developing our visualization, we have been talking about certain other features that we could potentially add to the system if we had more time, and one of the main ways that we could have extended our system that we came up with was mainly for the purpose of adding context to the visualization for the user. We believe we could have done more with the tooltip on the map, but we wanted to avoid over-cluttering the screen; it is likely possible to store a lot more information there such as things like the aggregate borrowed score, so we can see how much on average a language borrows or is borrowed. The aggregate age score as well could be a good indicator of how old a language may be, but in summary, for our tool tip, there are definitely pieces of our data and information for the reader that we could have made show up on the visualization and help the user understand our map more. Another area for expansion is the table below the map. We worry that users may not notice it below the map or understand that you can scroll to see more entries (when applicable). It would be interesting to make the table more interactive, exposing the user to more fields from the dataset, adding search functionality, adding the ability to sort by columns, etc. Lastly, we were thinking of the idea of introducing a play button to the slide bar. This would allow the user to see an automatically updating view of our map that would have the languages showing up over time. We think this could be a great way to extend the story telling of our visualization overall.

References

Haspelmath, Martin & Tadmor, Uri (eds.) 2009. World Loanword Database. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://wold.clld.org/