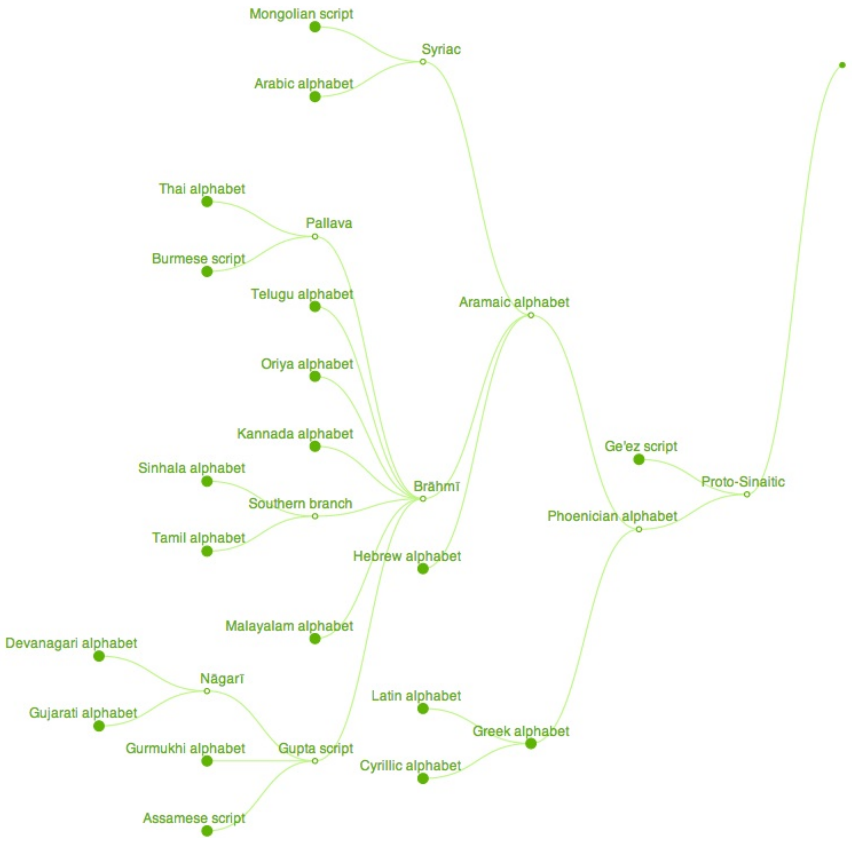


The origin of languages and their scripts

cs-171 Project III Process Book

Tamar Rucham

tamarulrul@gmail.com

Abstract

The origin of languages and their scripts is a project that aims to collect and visualize the evolution tree of different languages and their writing systems. We tend to consider spoken language and written language together, when in fact the two had very different evolution paths and form some unexpected connections between them.

Table of Content:

Visualization:

The landing page for this visualization contains the overview and video. The main visualization page has a column of language origins on the left and a column of script origins on the right. Clicking on a language family brings up the tree for that family (image 1). Languages that have an alphabet connected to it are larger and in full color. Selecting one of those nodes displays the alphabet node at the same level, connected to it's origin. The selected node is highlighted while the rest are greyed out (image 2). At the bottom there are details on demand - a frame displaying the wikipedia article about the language or the alphabet (image 3). Selecting the alphabet will open the writing system tree, with that node already selected. When a script tree is open, selecting one of the selectable scripts will present a list of all languages that use that alphabet (image 4). Initially, when the page loads, the Indo-European tree is ospen and English is selected, this is done to guide the user into the story of the languages, with the language families being the start of the story, the tree, English and Latin alphabet being the middle and the additional information from wikipedia being the end.

image 1

image 2

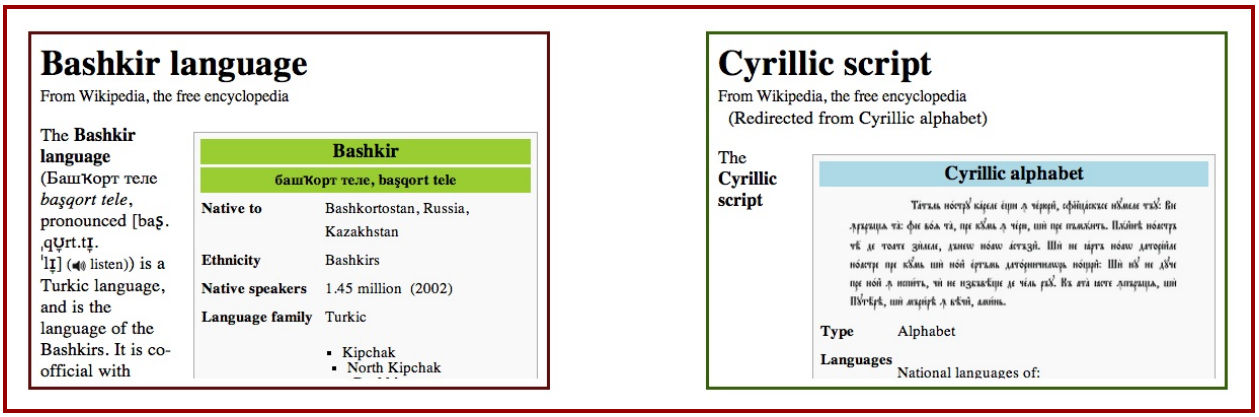


image 3

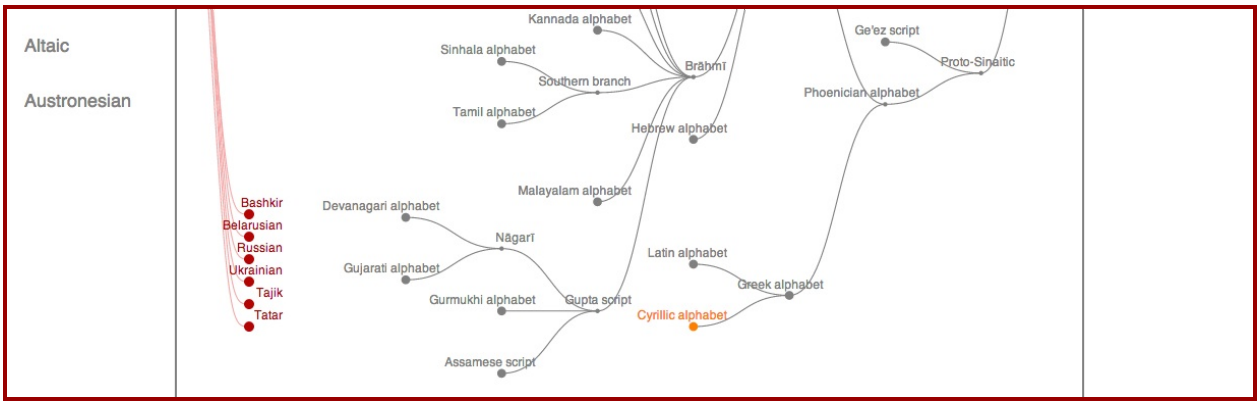


image 4

On first entry to the visualization, there are 3 popups that create a walkthrough in order to highlight some interesting points and familiarize the user with how the visualization works (image 5).

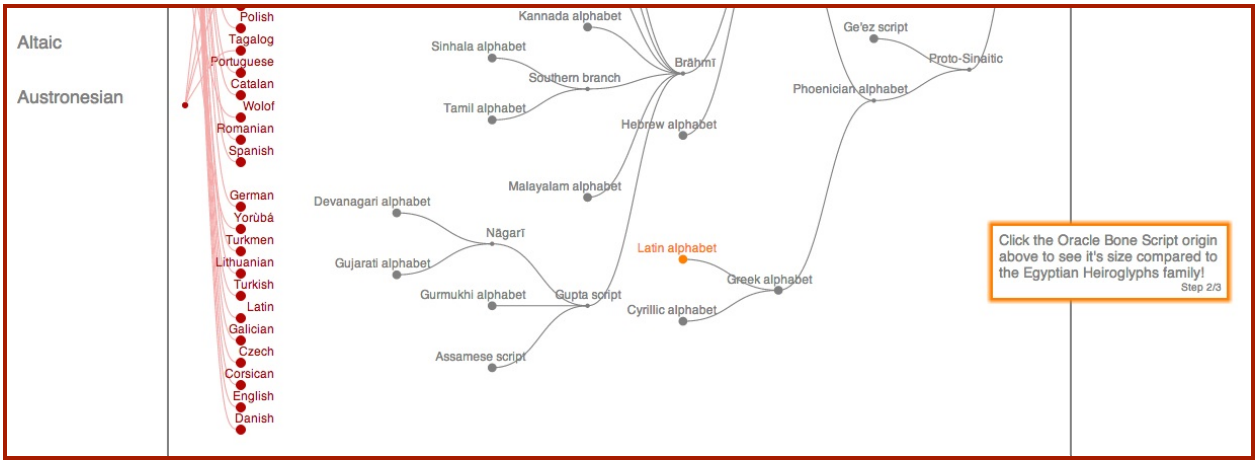
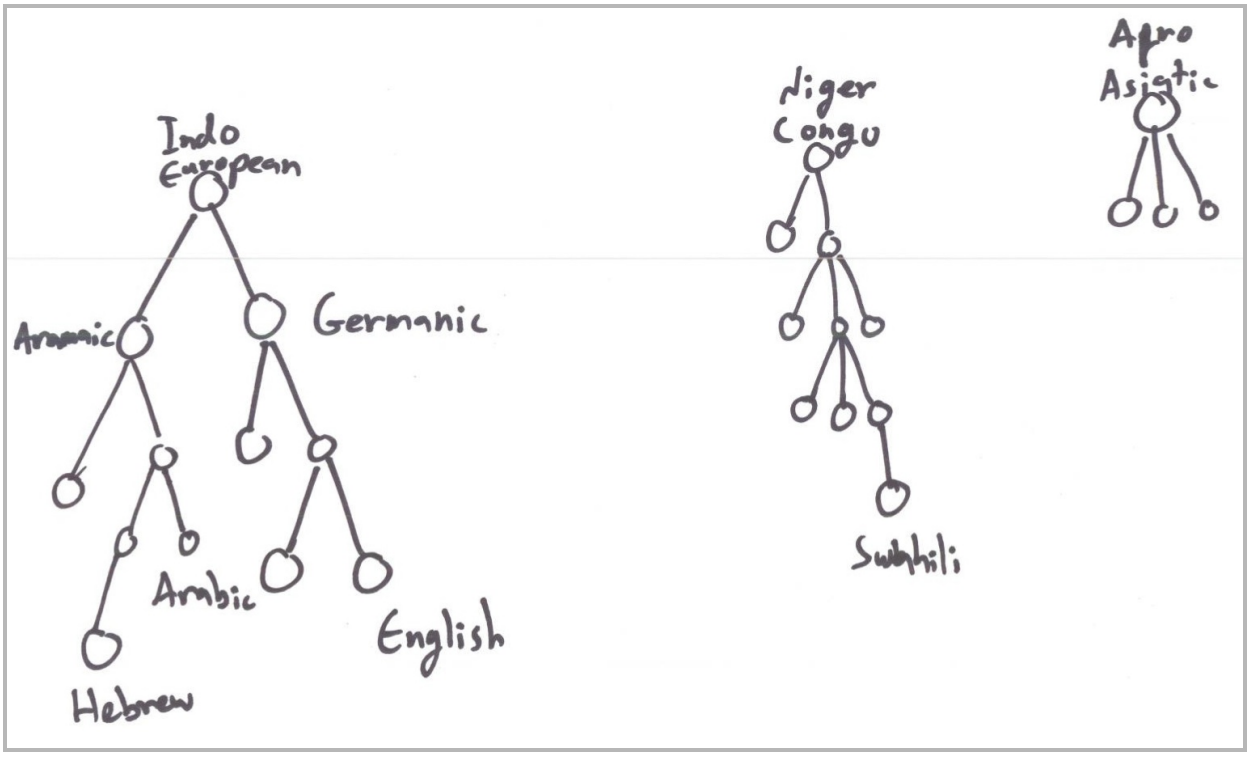


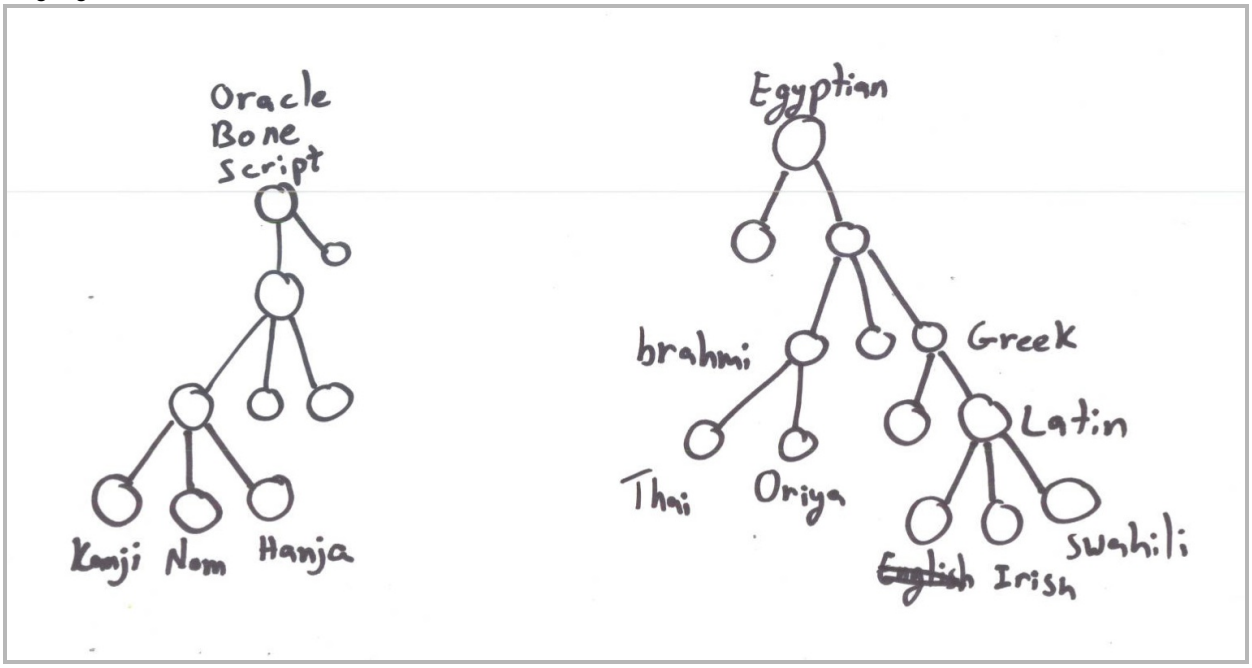
image 5

Design Evolution:

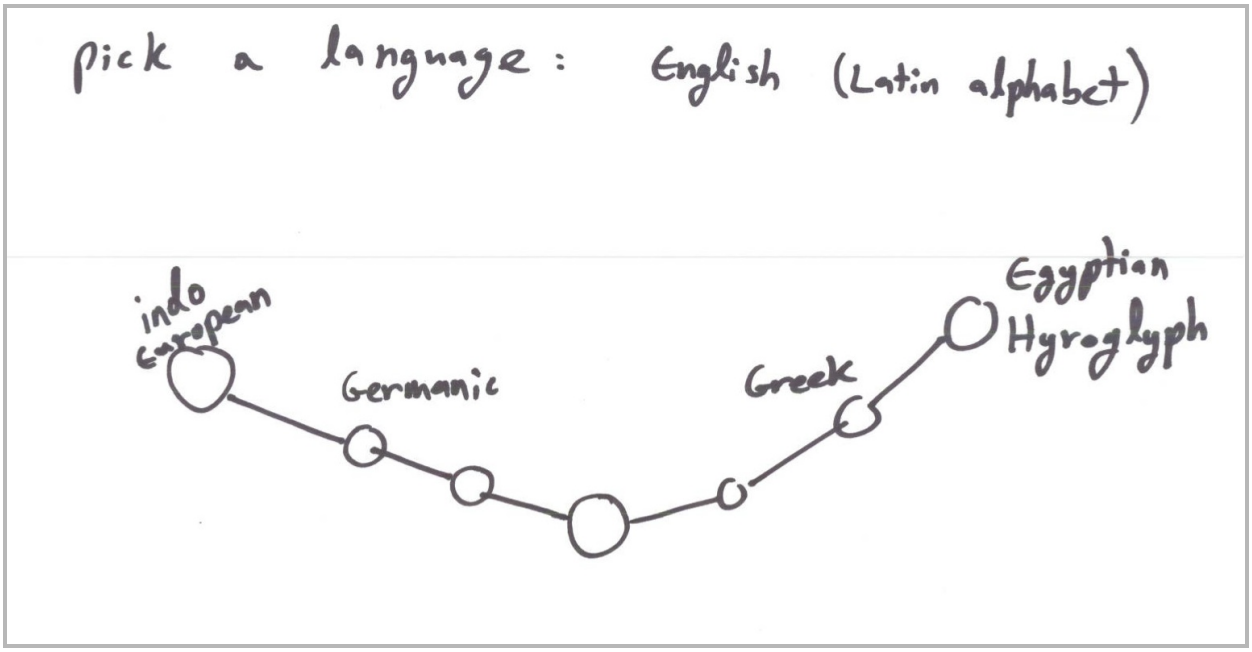
Following project 2 I started thinking of the connection between spoken language and the written one. Considering the evolution of each, I realized as I was sketching the trees below, there are two separate evolution paths for spoken vs. written languages, though we tend to consider them together.



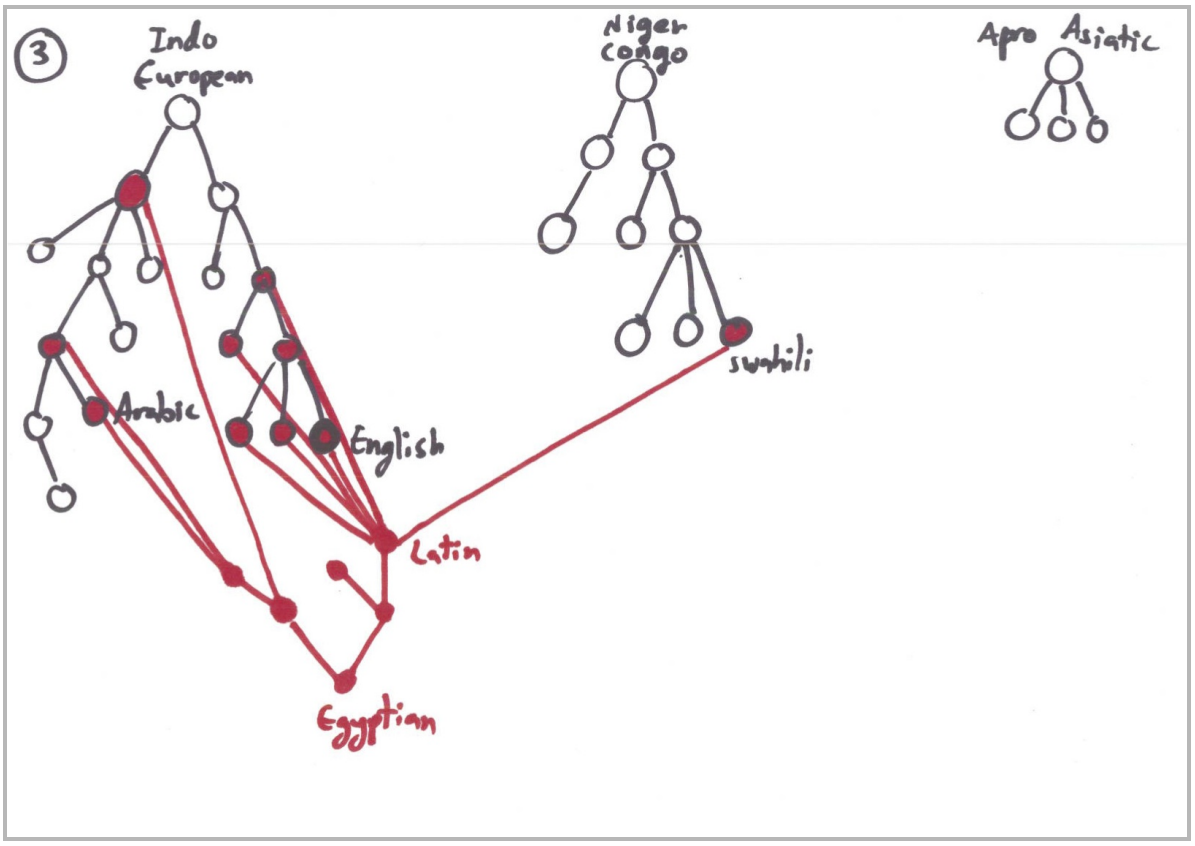
language trees



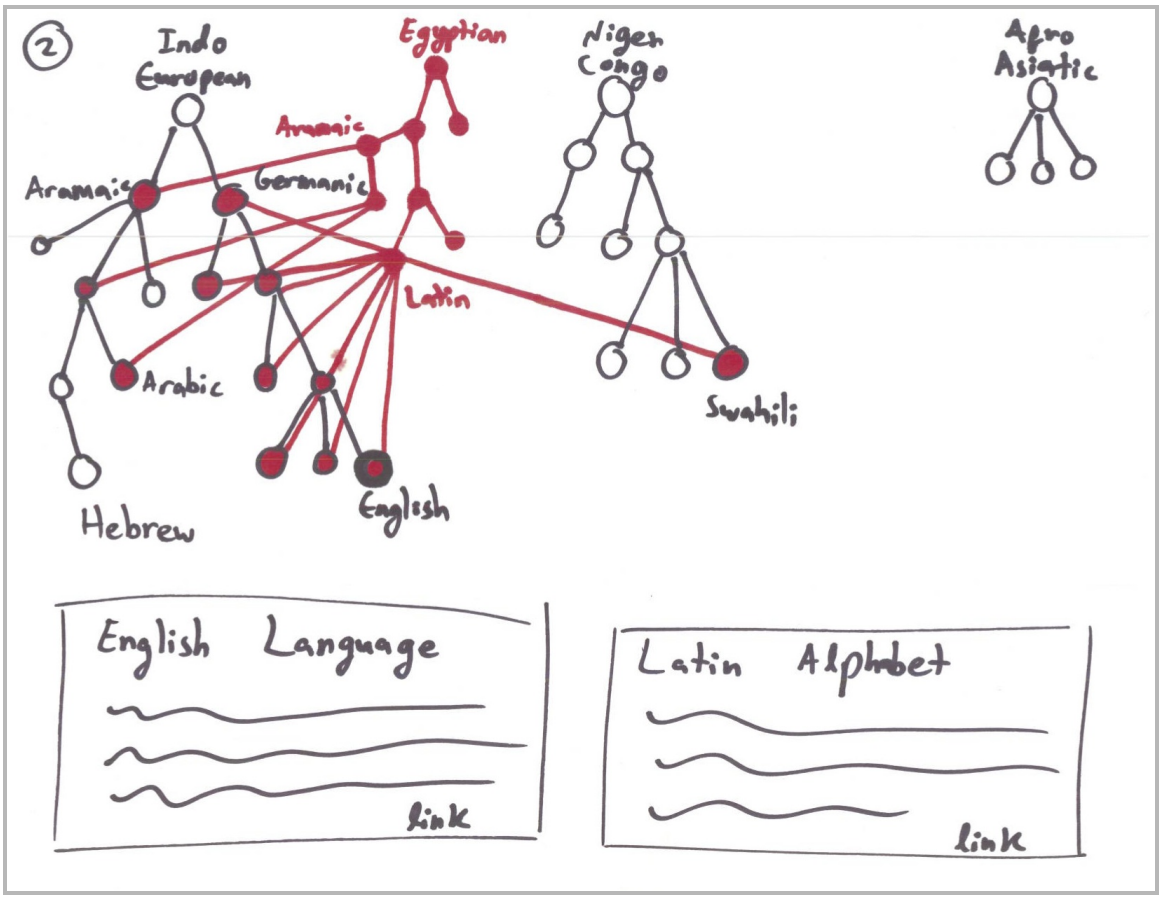
First I considered showing a single language with the evolution path from either side of the language and the script, as can be seen below.



But I wanted to get the full story and variety, such as the multiple languages that use a Latin alphabet. Therefore I decided to display all the language family trees side by side and connect the script tree either from below or as an overlay.

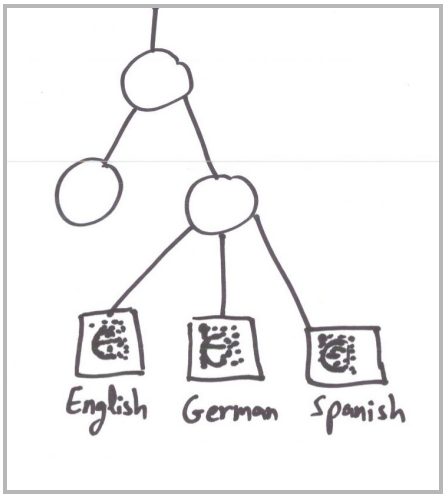


linking trees from below



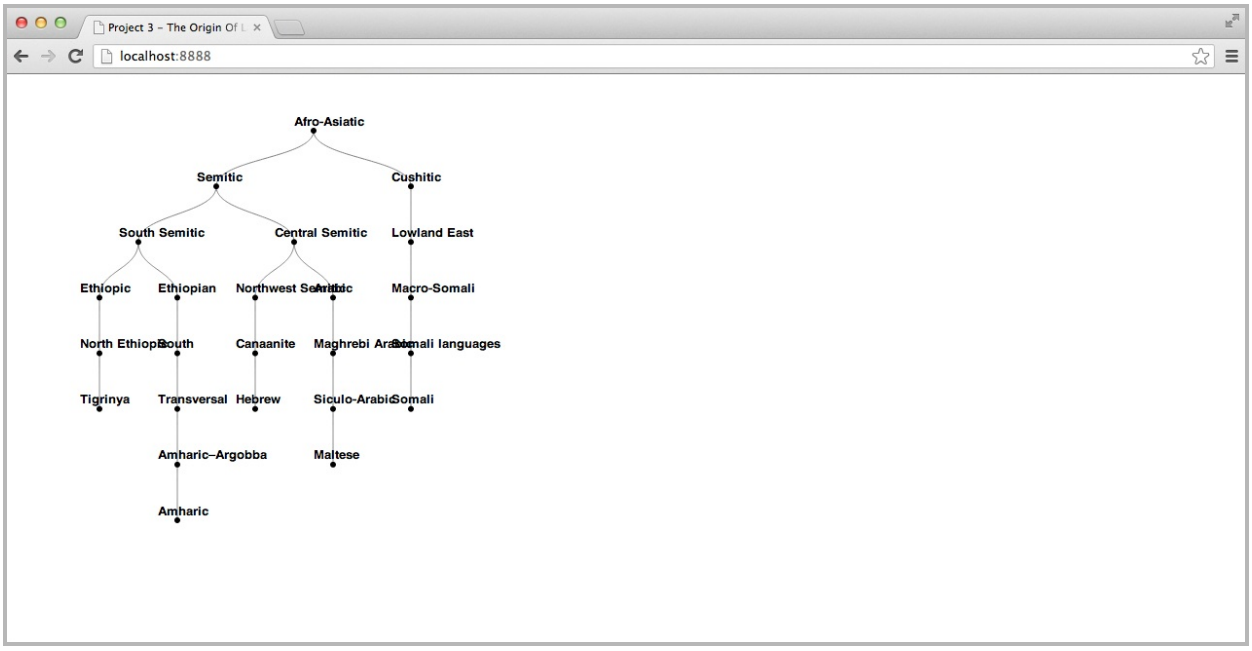
linking trees in overlay

I decided to go with an overlay since I did not want the page to be too long to obfuscate the connections entirely. I also considered adding the space distribution I did for project 2 as the nodes. The wikipedia articles at the bottom were added to provide further insight to the languages and scripts selected, allowing the users to delve into the stories that lie behind.



space distribution nodes

Once I started coding though, I realized the trees are far too large to allow for a side by side display. For example, you can see below the tree for the afro-asiatic family (which is not the largest family), takes up most of the page. Considering the fact I have 9 families, I realized I have to rethink my design.

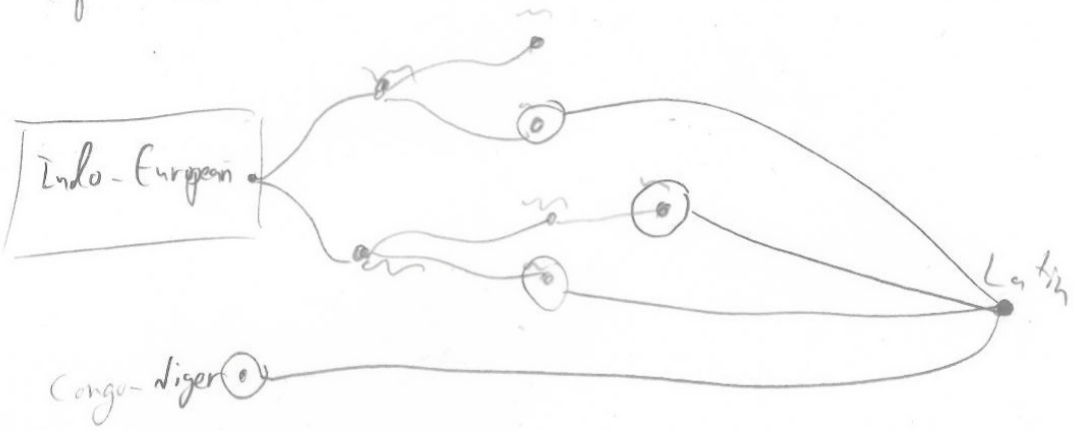


Afro-Asiatic tree in the browser

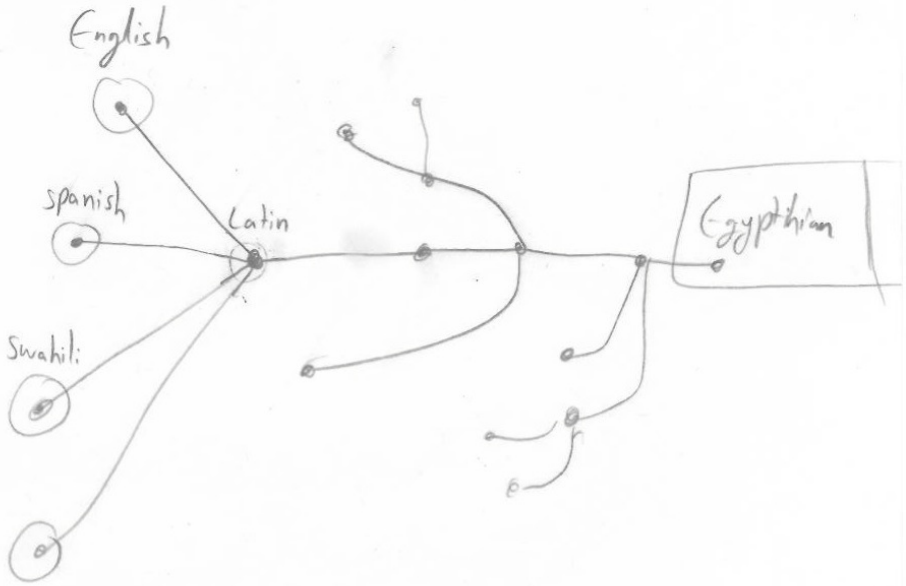
After discussing this issue with Chi and Alex, I decided to filter the data based on language and script origins and display only two relevant trees facing each other with their connections. However, the data was larger still, not allowing for two trees to be displayed horizontally together. Eventually I have decided on the final solution which is to only display the relevant node and origin on selection. The sketches below were the final ones which were translated into the visualization described above.

Finally, inspired by the lectures and examples of storytelling, I wanted to add a stronger storytelling aspect to the visualization. Therefore popups were added to create a walk through that highlights some interesting points and familiarize the user with how the visualization works. The walkthrough can be dismissed at any point, once the user interacts in a different way than suggested. This was done to create a story without limiting the interactivity of the visualization.

Afro-Asiatic

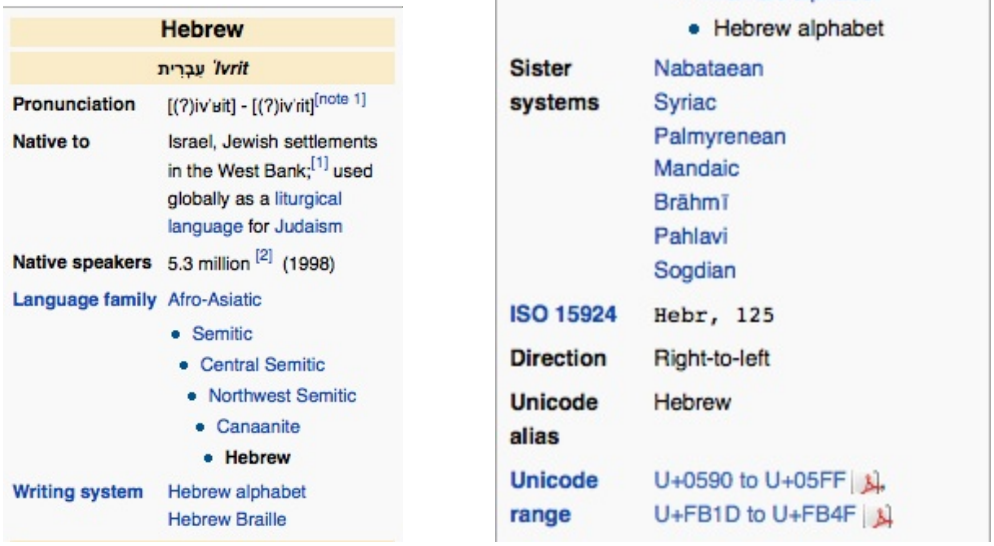


Afro-Asiatic



Data and technical details:

Wikipedia was the source to gather the information and form the trees of the languages and scripts. The codes for all the languages that have a wikipedia article about wikipedia were scraped using the Pattern library Wikipedia API. Then the language name was matched using an external source (see file in data folder). That name was used to search the Wikipedia article about that language, which was in turn scraped and the linked writing system was followed. Articles about languages and scripts have specific structure for the origins as can be seen in the images below. Languages that did not follow this structure were ignored. The trees were constructed by the python script, starting for each origin going down, checking if the node existed for each level. The data was stored in json files. A file was created for each language origin and for each writing system origin. Three more helper json files were generated - a list of language origins and a list of script origins which were used to generate the menu and indicate the json file name, and a json file listing each script and the languages that use it. After the generation of the data by the script, it was manually cleaned. There were some inconsistencies with trees starting points (for example some scripts ended their parent systems in the proto-sinaitic origin, though that is a child of the egyptian hieroglyphs) as well as references that use slightly different names (such as Latin script vs. Latin alphabet). Also, to reduce the overwhelming amount of data, some of the esoteric language trees and branches were removed. To narrow down the tree size, nodes that have a single child were ignored (single line of connection).



Hebrew	
עִבְרִית *'Ivrit*	
Pronunciation	[(ʔ)iv'ʁit] - [(ʔ)iv'rit][note 1]
Native to	Israel, Jewish settlements in the West Bank;[1] used globally as a liturgical language for Judaism
Native speakers	5.3 million [2] (1998)
Language family	Afro-Asiatic • Semitic • Central Semitic • Northwest Semitic • Canaanite • **Hebrew**
Writing system	Hebrew alphabet Hebrew Braille

Hebrew alphabet	
אָלֶפְבֵּית עִבְרִי	
Type	Abjad (for Hebrew, Aramaic, and Judeo-Arabic) True Alphabet (for Yiddish)
Languages	Hebrew, Yiddish, Ladino, and Judeo-Arabic (see Jewish languages)
Time period	3rd century BCE to present
Parent systems	Egyptian hieroglyphs • Proto-Sinaitic • Phoenician alphabet • Aramaic alphabet • Hebrew alphabet
Sister systems	Nabataean Syriac Palmyrenean Mandaic Brāhmī Pahlavi Sogdian
ISO 15924	Hebr, 125
Direction	Right-to-left
Unicode alias	Hebrew
Unicode range	U+0590 to U+05FF U+FB1D to U+FB4F

Details from wikipedia on the hebrew language and the hebrew script respectively

Once the data was ready, the pages were created using javascript and the d3 library to generate the visualization. The trees were based on the example by Pavan Podila (Jul 20th, 2011) in his blog Pixel-In-Gene. The dynamic size calculation was based of the answer in StackOverflow by user SuperBoggly (Jan 3rd, 2013).

Analysis of results and thoughts:

Many interesting connections were revealed by this visualization. The Indo-European tree contains the most languages that have a wikipedia article and Latin is by far the most used alphabet by the languages represented. If we look at the script trees we find that Egyptian Hieroglyphs origin encompasses all the writing systems of the languages displayed except for Chinese and Japanese. Those use the chinese characters which come from the Oracle Bone Script origin. On the other hand the shortness of Oracle Bone Script system reveals that the chinese characters are the most ancient writing system still in use. A different story is that of Bantu languages, from the Congo-Niger tree. All three languages which are represented in this tree, use the Latin alphabet, which is most likely a result of the unique history of the black continent.

Thinking back on Project 2 results, Chinese space distribution was distinctly different than the other languages. In light of the data displayed in this visualization I assume the reason for this is that Chinese writing system has a different origin than the other languages analysed in project 2.

Given more time I would have liked to connect the data to a map of the world and possibly a timeline. That way the story of the spread of languages and writing systems would have had the geographical and historical connection.

Special Thanks to Chi and Alex for their feedback and thoughts.