The focus of this thesis project is to explore visual themes of different scripts. The word script is used to refer to the character set used by different writing systems to encode spoken language; i.e. the Latin alphabet, is a script used by many different writing systems to express languages such as English, German, French and Malay. In the field of linguistics some research exists in the development of specific scripts (see related work section xx). However the goal of this project is not to explore the evolution from past to present, but to compare different current scripts in light of their historical connections.

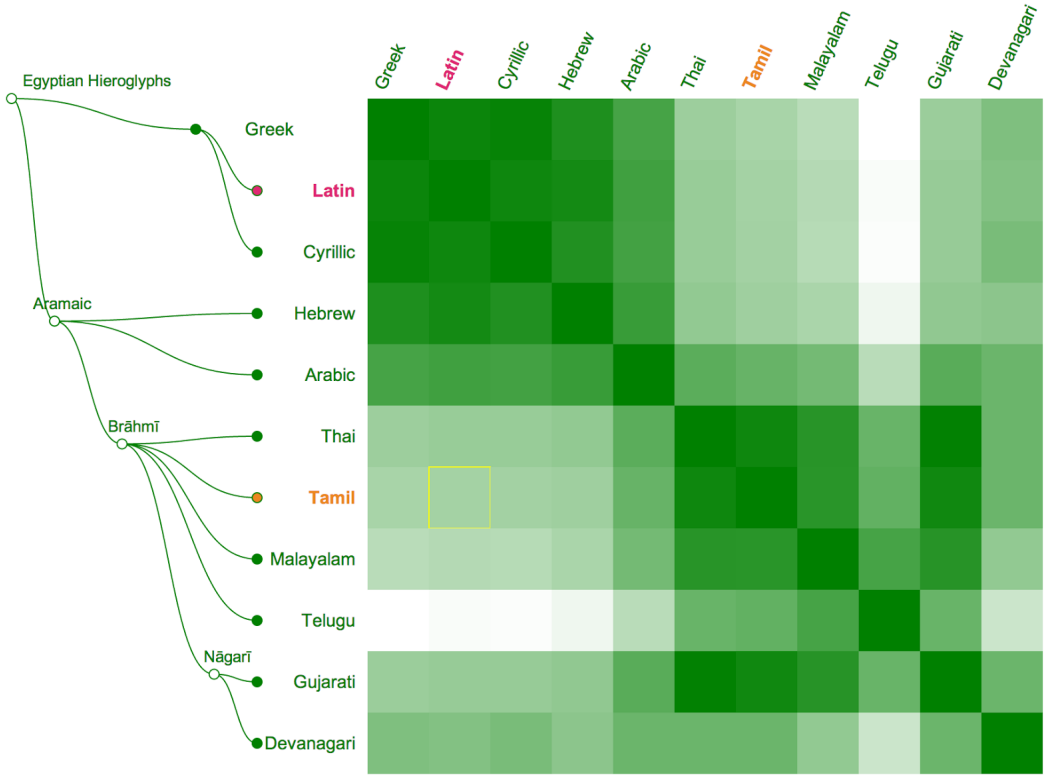
Background

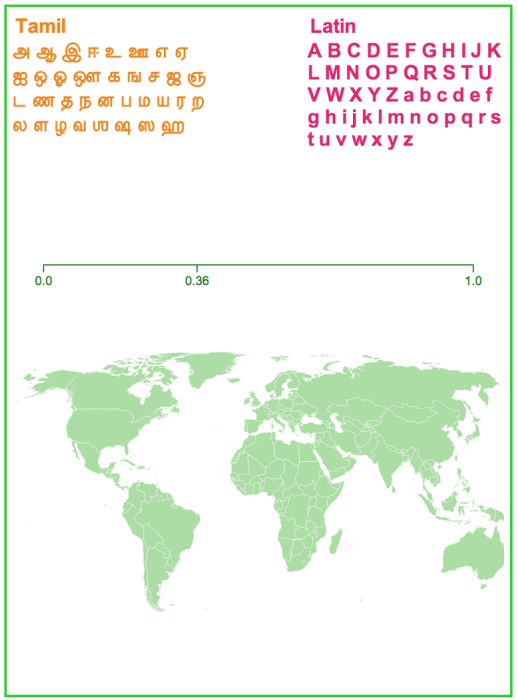
The visual aspect of written language played an important role in the history of writing. Before written language as we know it today was *proto-writing*, a picture writing system. Proto-writing uses ideograms or pictograms - graphic symbols that represent ideas or objects respectively - and does not directly translate to specific words from spoken language. The shape of the symbols conveyed the information and a reader would not necessarily need to know the spoken language of the writer in order to gather the information contained in the symbols. When most *true-writing* evolved (true-writing being a system in which the entire content of spoken language can be encoded), symbols that represent whole words were used to encode information (logograms). Again the shape of the symbol was related to the information, but unlike in picture-writing systems, they represented specific words of the writer's language. From there, the phonetic system stemmed, in which symbols represent sounds that are combined to phonetically construct words from spoken language. The shape of the symbol no longer has a specific meaning; rather it can be combined to create many different words. Some scripts, such as the chinese characters or egyptian hieroglyphs, preserved both systems such that symbols can function both as logograms and as phonemes. In general, most scripts lost the connection that used to exist between the visual shape and the meaning. However the visual aspect of the scripts we use today stemmed from a long history of evolution.

Result

The end result of this project is a web based interactive visualization that compares 11 scripts that are in use today, all of which stemmed from the Egyptian Hieroglyph origin. The visualization displays the evolutionary tree of the scripts on the left (as scraped from wikipedia *reference to*

chapter 3), followed by a heatmap mapping the similarity between all scripts (see figure 1). The rectangles in the diagonal that represent the intersection of a script to itself (identity) show a small scatter plot of the letters distribution and 3 clusters. All scripts are scaled the same so this small multiple can also be used to compare scripts by letters distribution. Selecting one of the rectangles representing the similarity between two scripts highlights the two scripts, displays an extra data section shows a larger version of the scatter plots (TBD). The extra data section shows the full character set, the similarity score on a scale and a world map with current script distribution (based on Wikipedia *add reference to article*).


Greek
Latin
Cyrillic
Hebrew
Arabic
Thai
Tamil
Malayalam
Telugu
Gujarati
Devanagari
Egyptian Hieroglyphs
Aramaic
Brāhmī
Nāgarī


Tamil
அ ஆ இ ஈ உ ஊ எ ஏ
ஐ ஒ ஓ ஔ க ங ச ஜ ஞ
ட ண த ந ன ப ம ய ர ற
ல ள ழ வ ஶ ஷ ஸ ஹ
Latin
A B C D E F G H I J K
L M N O P Q R S T U
V W X Y Z a b c d e f
g h i j k l m n o p q r s
t u v w x y z
0.0
0.36
1.0

Organization of this document

The related work in chapter 2 covers work done both in the field of linguistics and in the field of data visualization. Chapter 3 describes initial work, which includes two projects relating to letters space distribution and mapping of spoken languages and scripts. Chapter 4 will detail the data collection algorithm and analysis, including the validation procedure. The visualization design and implementation will be discussed in chapter 5.