Project Proposal: Syllables, Languages, and Geography

Koffi Alavo, Chauncey Becker, Rohit Kunte, Austin Simmer

Introduction

The goal of this project is to analyze syllables that are used in multiple languages, and gain insight into which continents favor which syllables. We will make use of the skills we have learned in this class by combining multiple, disparate datasets, and splicing them together in R to allow better analysis and visualizations.

Data

The data we will be using comes from the UCLA Phonological Segment Inventory Database.

http://www.linguistics.ucla.edu/faciliti/sales/software.htm#upsid

Sadly, this website was defunct at the time of access, but a group of scholars from Goethe University Frankfurt, in Germany, preserved the database here:

http://www.linguistics.ucla.edu/faciliti/sales/upsid.zip

This link will download a ZIP file consisting of two defunct Windows executable files and some .STR files in the root directory, and many .INF files in the folders named UPSID2, 3, and 4. The .STR files can be opened with any text editor; each file contains information about a single language, including the language's name and classification.

Additionally, a tab-delimited text file of 194 languages and syllables used within those languages is located here:

http://web.phonetik.uni-frankfurt.de/upsid_matrix.txt.zip

We will combine the .STR files' classification/family of the language with the tab-delimited list of languages and syllables. This combination is needed because the tab-delimited list does not include the languages' classification/family. Additionally, the classification/family of a language can be mapped to a region easier than an individual language can. Such a mapping can be found here:

https://en.wikipedia.org/wiki/List_of_language_families#Language_families_(non-sign)

The Big Question

Which languages and locations favor specific syllables?