Creating Fake Words New Languages

After using a Markov model to explore phonotactics in English, I decided to do a similar analysis looking at simply letters. For languages that have irregular orthographies (ie, you can't tell exactly how a word is pronounced from its spelling, like English or French), this is just an analysis of orthography. While exploring only orthography doesn't tell you everything about a language, it still tells you a lot about that languages spelling rules and history. It is also very useful in Scrabble, if you're into that.

For languages that have phonemic orthographies, like Spanish, this analysis is just as valid for analyzing phonology as it is for spelling.

I got my data from the awesome github repo FrequencyWords, which generates wordlists in 59 different languages based on subtitles from TV and film. I also recycled a lot of code from my previous project, in my Markov-Phonology github repo.

Here are some samples of outputs in languages that I'm interested in

#Spanish

#English Words

#Swahili

#Tagalog

#Indonesian

#French

A big challenge was getting the data into ASCII