N-Grams are continuous sequences from a text file of length n. They are used to calculate the probability of n words being in a phrase together, which is useful in a language model. In our case the likelihood of a series of words being next to each other helped to identify languages as they all have unique words and phrases that are not present in other languages. This can also be used to identify parts of speech, collocations, and distinct writing styles if trained and tuned correctly.

If there are 105 total words in the vocabulary and the current unigram appears 12 times, its probability would be around 11%. A bigram’s probability with Laplace smoothing is found using the formula: (b +1) / (u + v) where b is the number of times the bigram is present, u is the unigram count of the first word in the bigram, and v is the total vocabulary size.

A language model is only as strong as the text that is used to create it as the diversity of words and phrases used would be able to make a more well rounded model whereas a more specific model would be better for identifying something specific. If the model we created had a more diverse text for each of the languages, there may have been an increase in accuracy levels as the models would’ve had more phrases and words that would be unique to a language. Smoothing makes a model more accurate as it eliminates the chances of inaccurately high or low stats. One way to do so is to add a one in the numerator of the bigram probability as it would eliminate the risk of there being a 0 if the bigram is not present in the training data. Language models can be used to create sentences that make more sense to a reader simply by using bigrams with high probabilities. This fails to account for correct sentence structure a lot of the time and would only be able to create generic sentences that fail to have meaning.

Models are evaluated using perplexity (the multiplicative inverse of the test sets probability), bits-per-character, and cross entropy. Googles n-gram viewer shows the occurrence of n-grams throughout history and a graph showing the changes in use overtime. It's a great way to visualize how language has changed throughout history, as words that you may not expect to see being used in the 17th century such as “computer” were commonplace.