Voynich Extra Credit

Jack Ellenberger

June 11, 2015

Method

In order to analyze the Voynich Manuscript, I relied on the work done for Homework 6, Successor Frequency. I was interested in the numerical data that could be pulled from words split where their successor frequency is greater than 1. These splits effectively identified common runs in words in addition to prefixes and suffixes. While not a perfect metric for comparing languages, these splits allowed dialects to be contrasted in a general way: what language is more prefix-focused, what language has unique word constructions, and what languages may be related at a language deeper than vocabulary. I'm acting on the assumption that if languages have similar distributions for certain basic metrics, then they may have developed from a similar source. This theory is tested to the best of my non-linguist abilities in the next section, where I contrast known language.
Specifically, the metrics I used to compare on are:

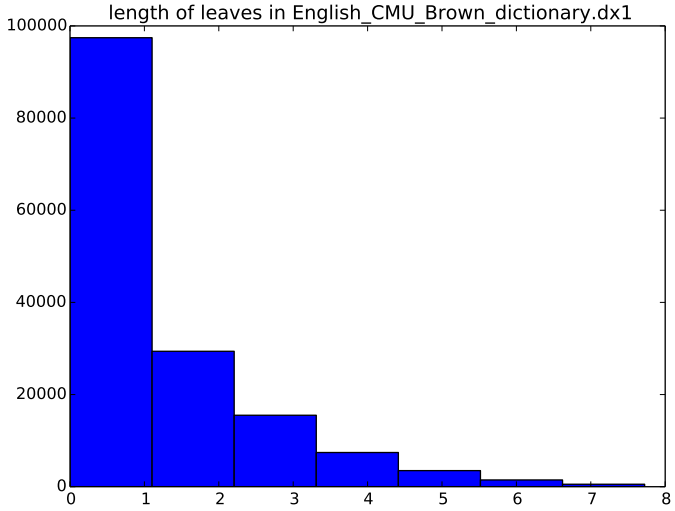
- numUsages: for each prefix/root, the number of words in the corpus that use that root
- usageLens: for each usage of a root, the number of split parts. These are most often referred to as chunks in this analysis for simplicity
- leafLens: for each chunk, the number of letters

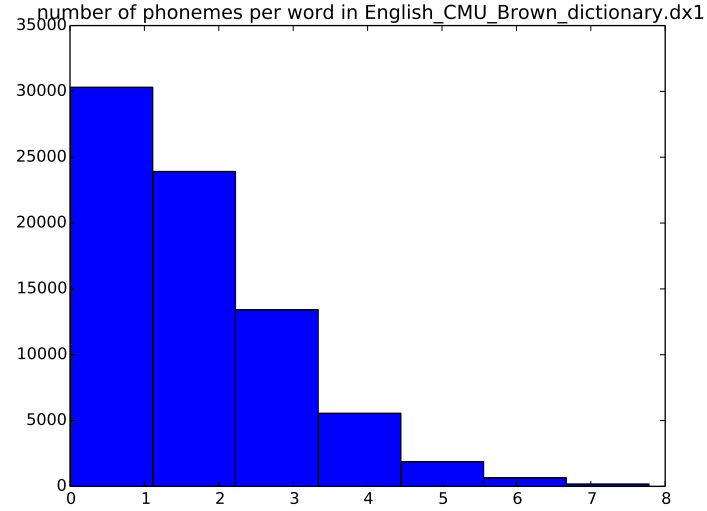
To compare these distributions, I created histograms for a very basic shape-analysis. These were most useful to throw out Voynich interpretations that didn't have enough data to be significant. I will mostly present Quantile-Quantile plots of the distributions which gives a good visual indication of similarities, given that the lengths of each dataset is unequal, ruling out interpretations such as a Kolmogorov-Smirnov test.
I will only present a fraction of the data I have gathered. The rest is available at
https://github.com/jackellenberger/ComputationalLinguistics2015Voynich.

Initial Comparisons

We will start with the most absolute of comparisons: Comparing english with itself. We have a corpus of 1000 words, as well as the CMU dictionary. While the latter is much more respected, we can gather enough data from the former to see if our comparisons have any meaning at all. We will include the histograms for this comparison, but soon drop them in favor of the more distinguishing QQplot.

Figure 1: CMU Brown English Corpus





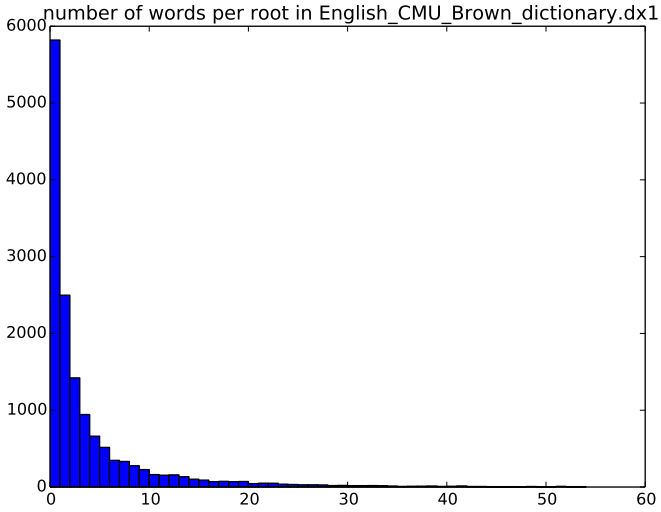
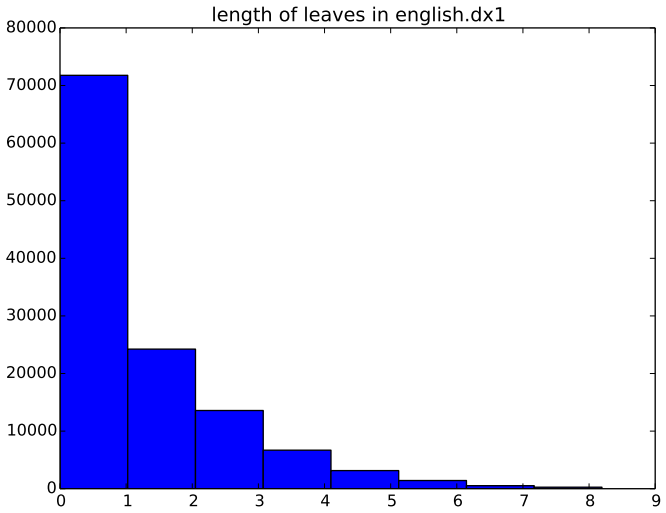
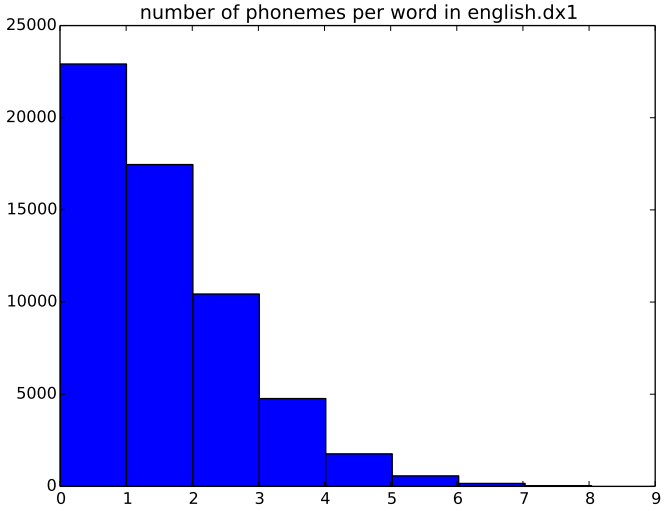
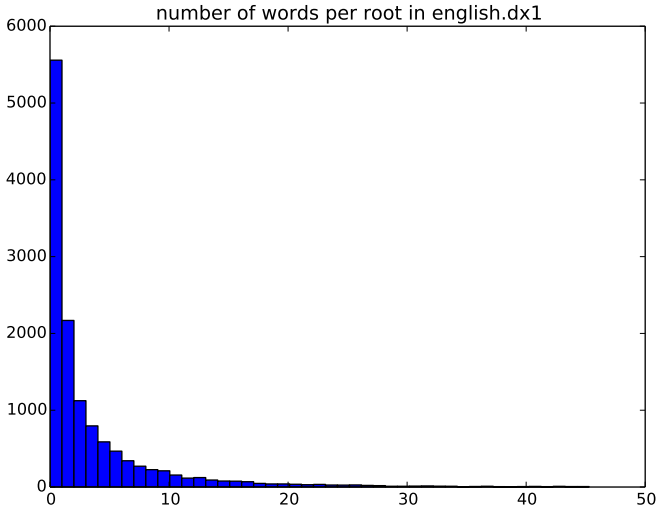


Figure 2: English 1k Corpus

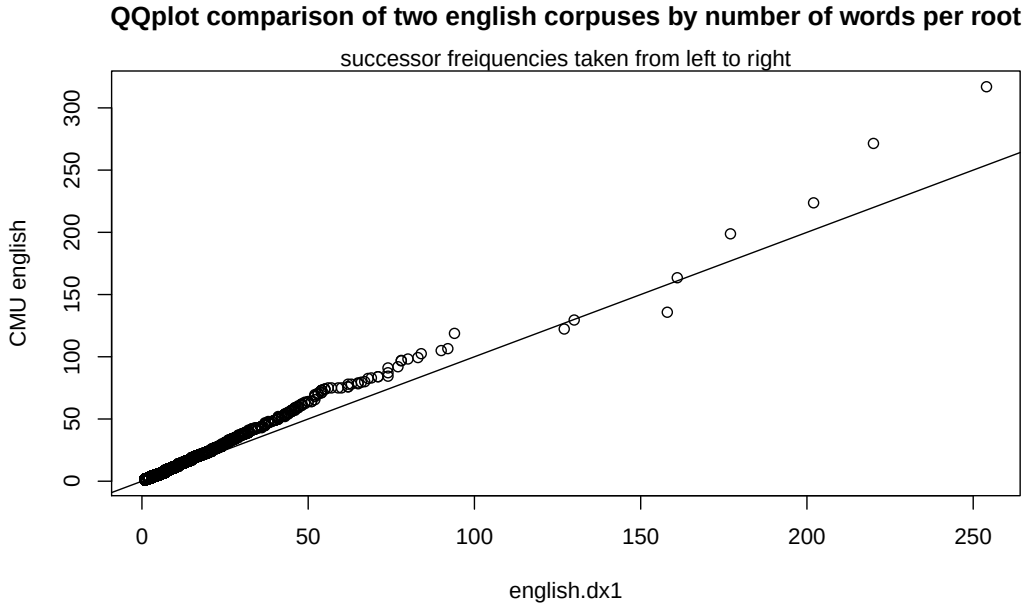


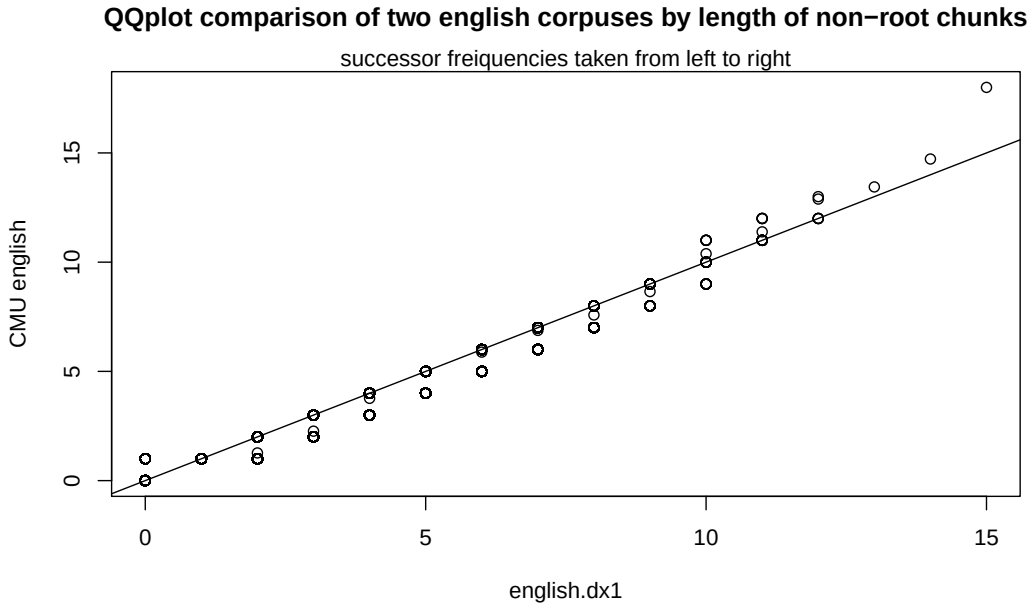


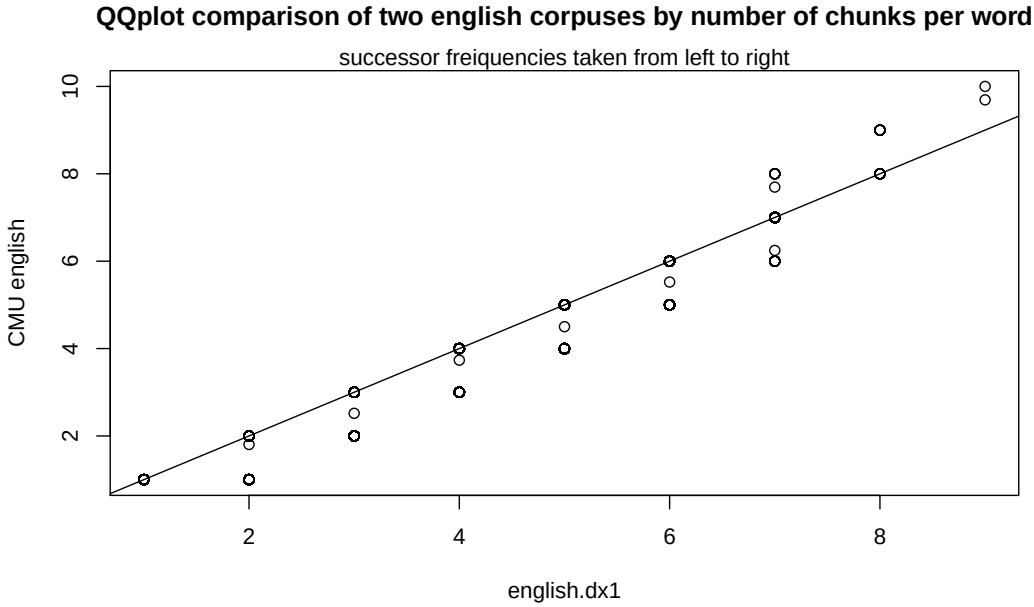


We can see that they are similar as far as the distribution goes, but the values are not identical. This frustrates a numerical comparison, but the Quantile Quantile Plot resolves this by providing a visualization of a numerical similarity between unequal elements. Note that all QQplots feature the line y=x, and all analysis is done in the left to right direction. With more time, a right to left version could be easily developed.

Figure 3: Comparison of two English Corpuses

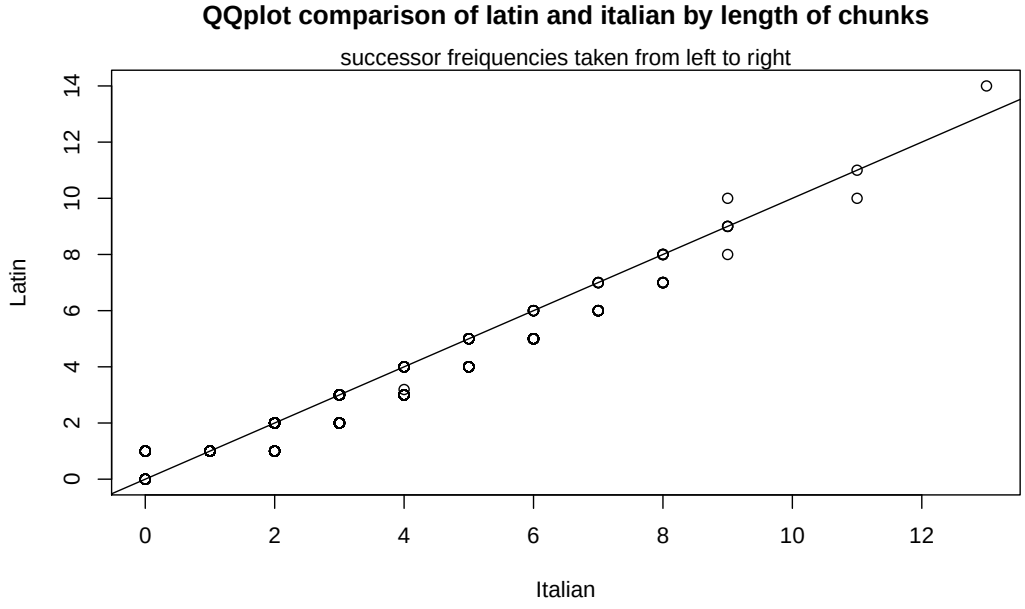


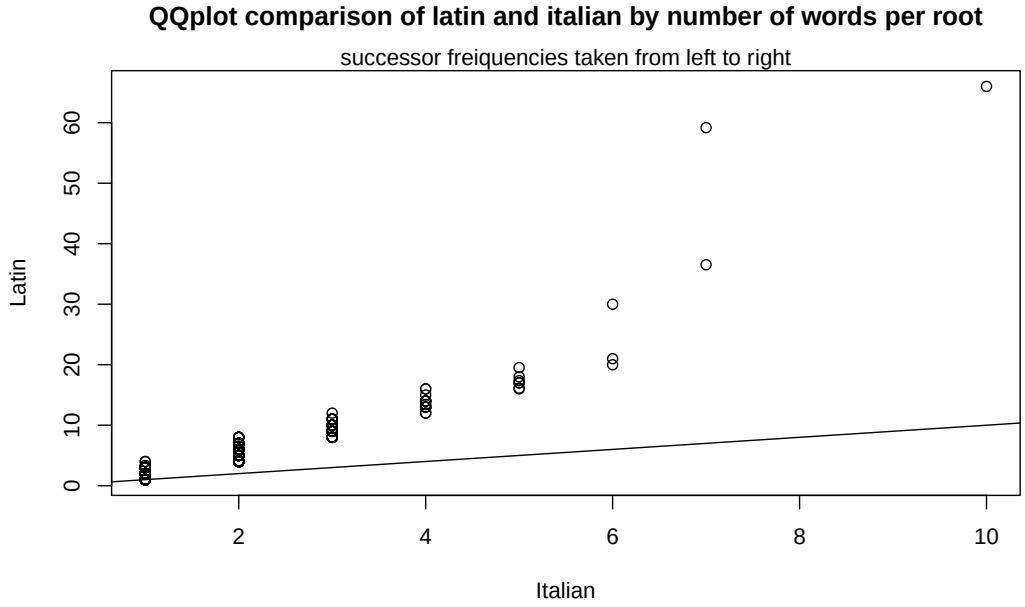


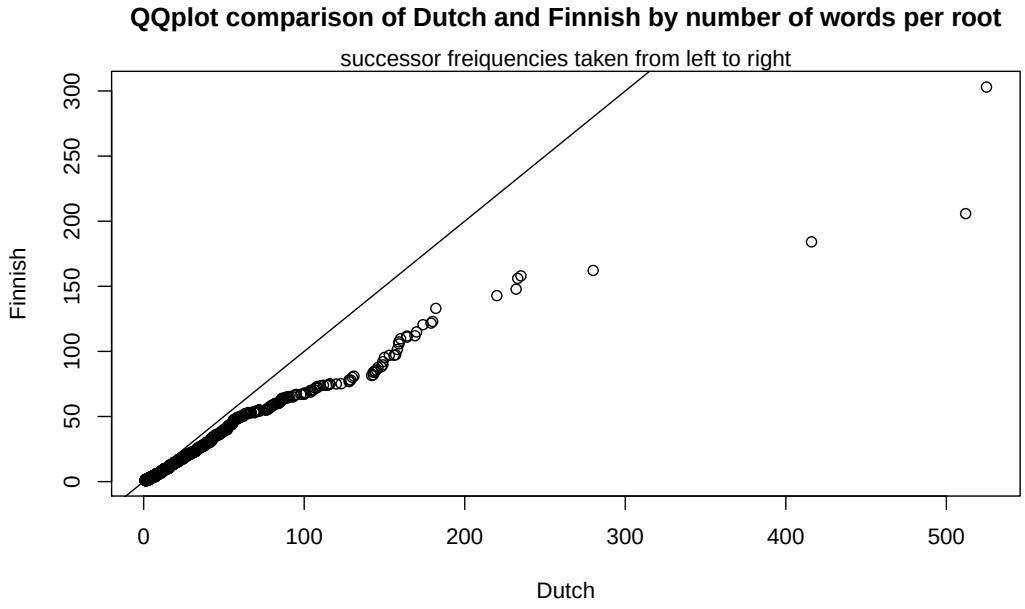


We see that although the data is a bit sparse, it follows the y=x trend fairly closely.
Now we will look at a pair of known-similar language that share some features, where they diverge in others. These languages are Latin and Italian They are not mutually intelligible, but one develops directly from the other so their structure should be similar.

Figure 4: Comparison of two Mediterranean Languages





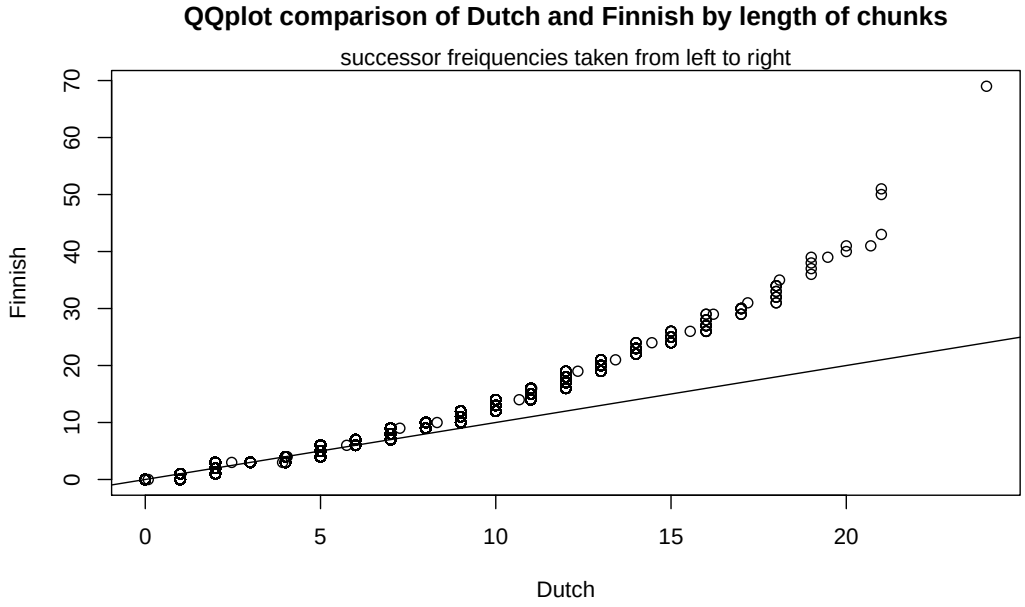


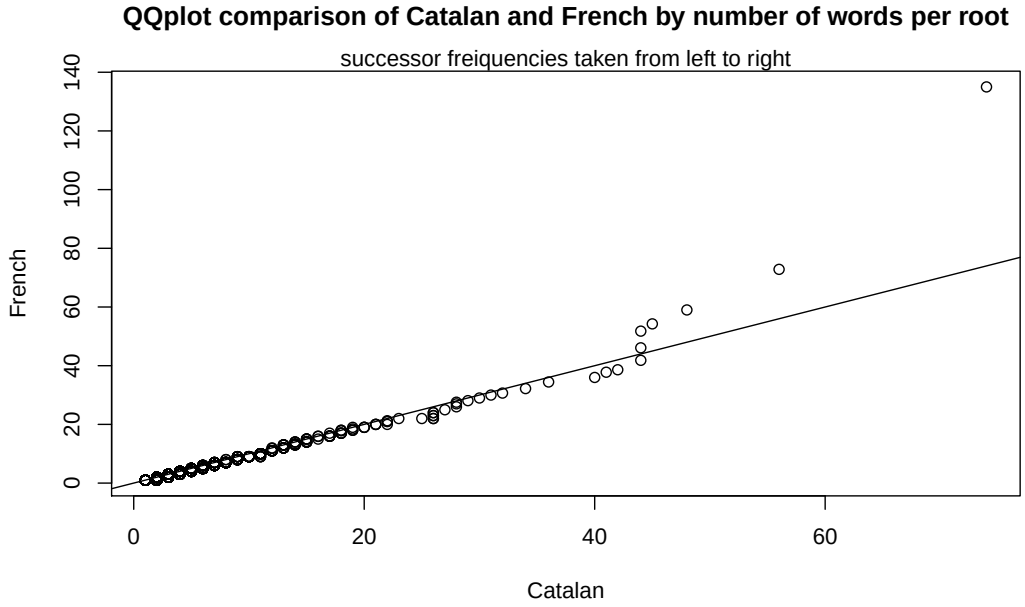
We see that a comparison of the length of chunks and chunks per word are very similar, perhaps meaning that the structure of the language is similar - prefixes are often 3 characters long, tense suffixes are 3 letters, superlative endings are 4 letters long, etc. However, on number of words per root diverge widely. This could be an indication of the increased grammatical complexity of latin, or it could be an indication on the larger more diverse vocabulary of Italian.

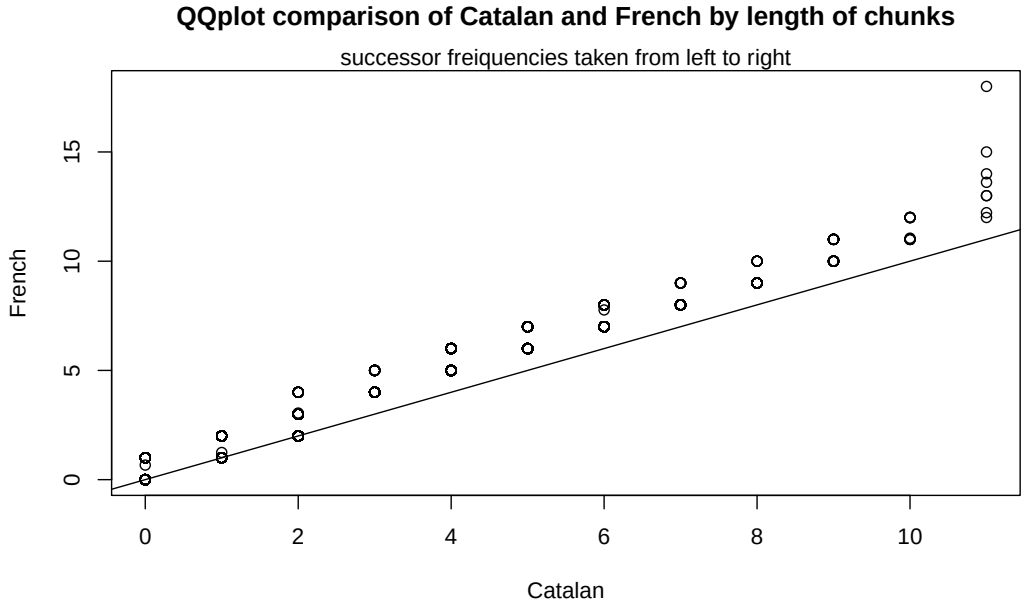
In the full list of plots there are some interestingly similar languages - Catalan and French, for instance, are very similar despite their fierce political differences.

To show that this method differentiates differing languages, we next compare Dutch and Finnish. A comparison of Dutch and Norwegian may yield similar plots, but Finnish, being closer to Russian (I think) and other Slavic languages rather than Germanic means that they diverge greatly.

Figure 5: Comparison of Dutch and Finnish





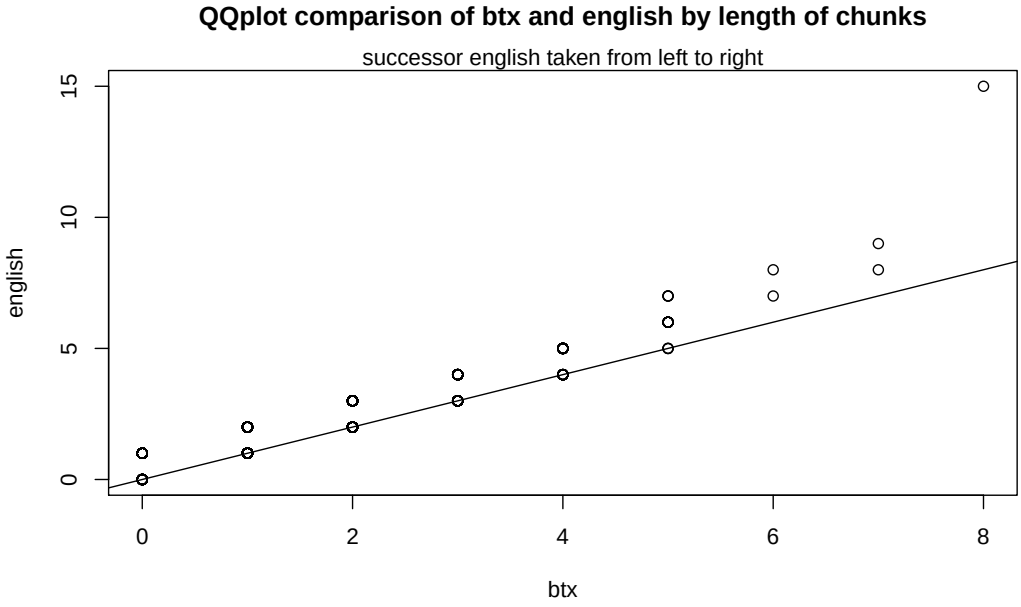


On no metic do Dutch and Finnish appear to be similar. We will now see if we can apply this method to the Voynich manuscript to determine its linguistic similarities.

Voynich

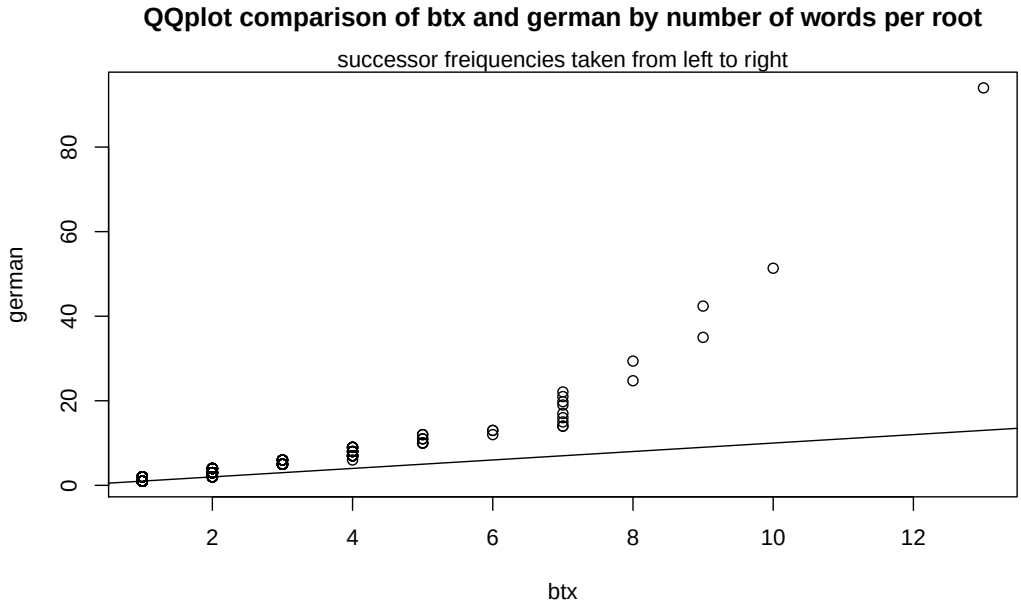
The first thing to note is that there are several interpretations of the Voynich manuscript, and some are better than others. There was a very persistent problem with with majority of the interpretations not being long enough to draw any conclusions from, so I have only made plots for interpretations coded BTX, BFX, BF2, BC2, AL4, AF2, AF1, and AC1. The interpretations of these three letter codes in included in the problem description of homework 1.
I made the most through comparison with BTX, and with the best language contenders I made further comparisons against other interpretations. Notably, I found BTX to be similar in some ways to English:

Figure 6: Comparison of BTX and English



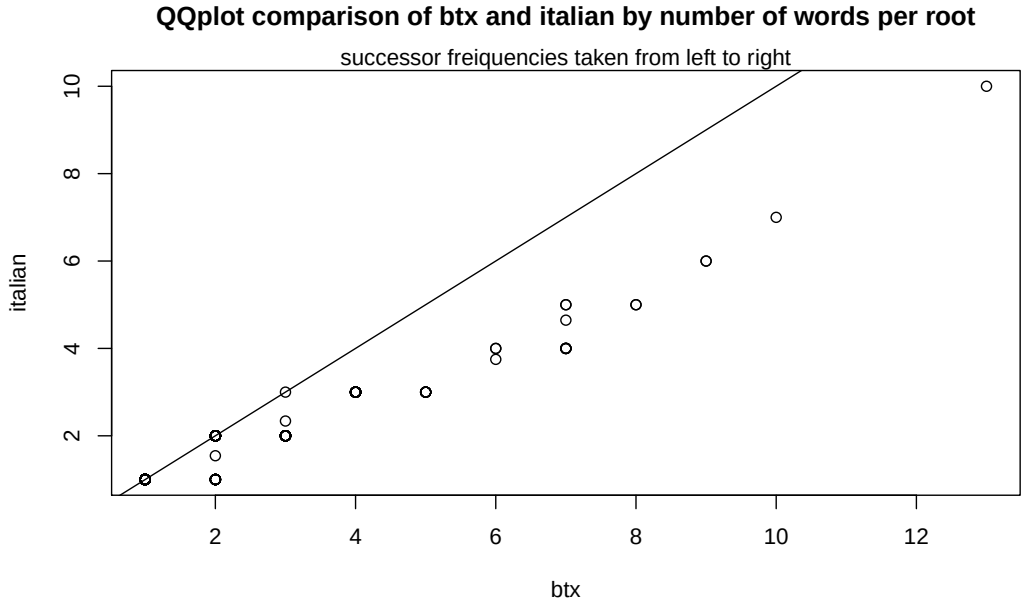
This didn't hold up, however. They may be similar in length of chunks, but they differ greatly in words per root. As we saw in latin v italian, this may be indication of a common linguistic ancestor. With this on my mind I compared BTX to Modern German and was disappointed with their dissimilarity.

Figure 7: Comparison of BTX and German



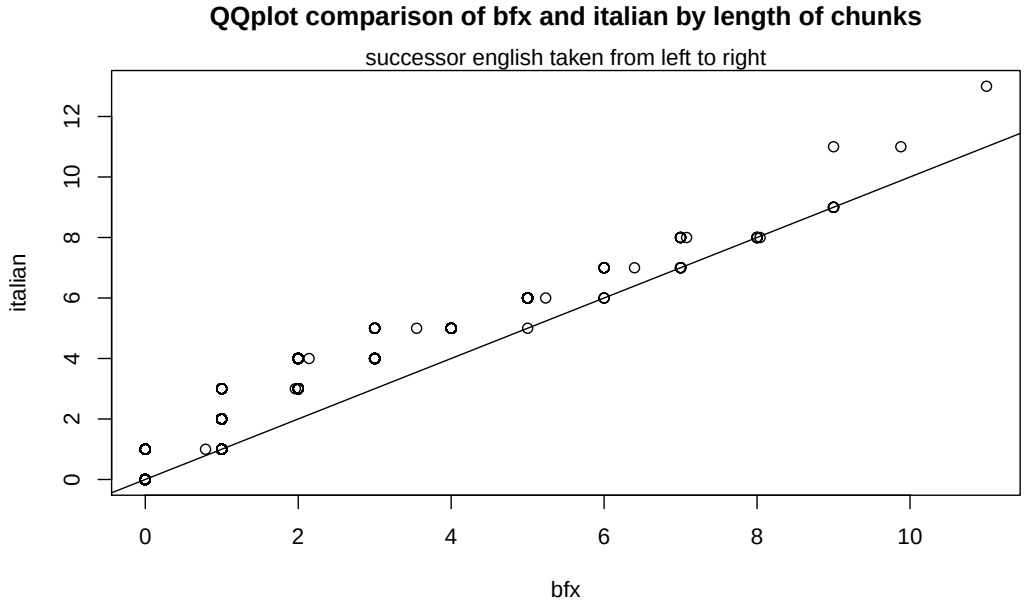
Modern Italian, too, seemed very dissimilar.

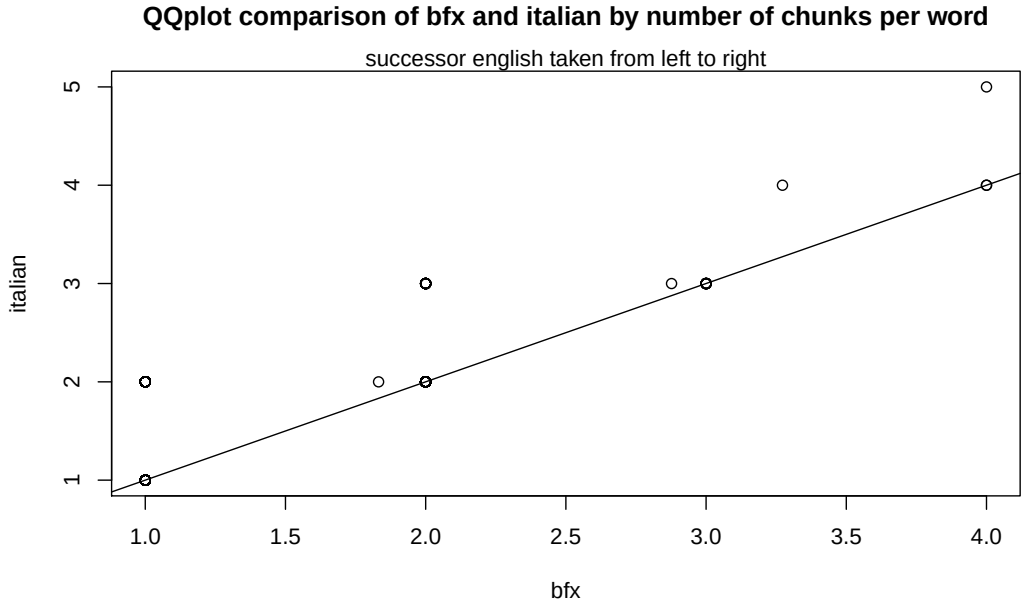
Figure 8: Comparison of BTX and Italian



However, other interpretations were more favorable to italian, especially BFX:

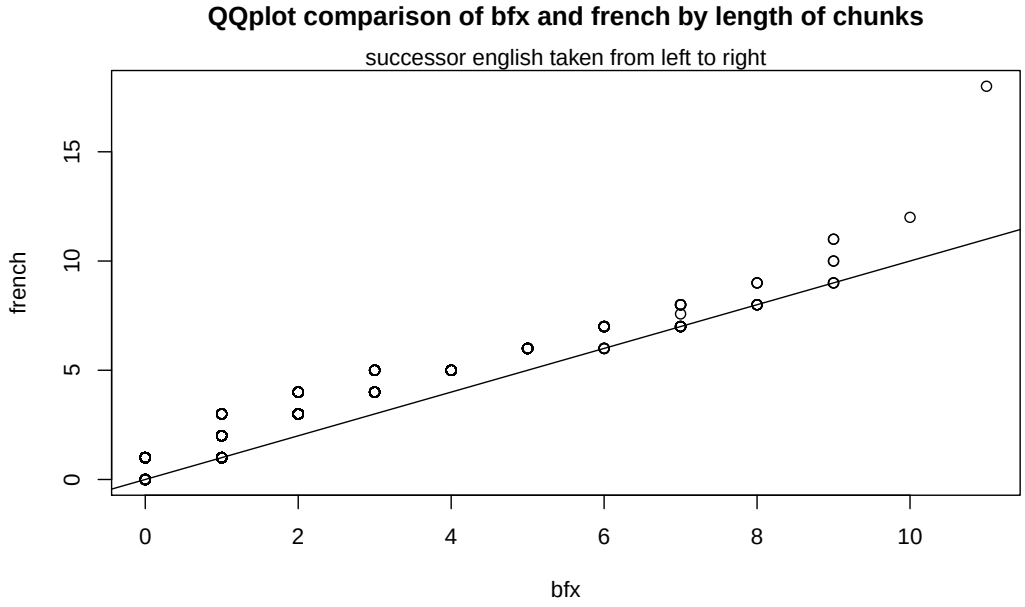
Figure 9: Comparison of BFX and Italian

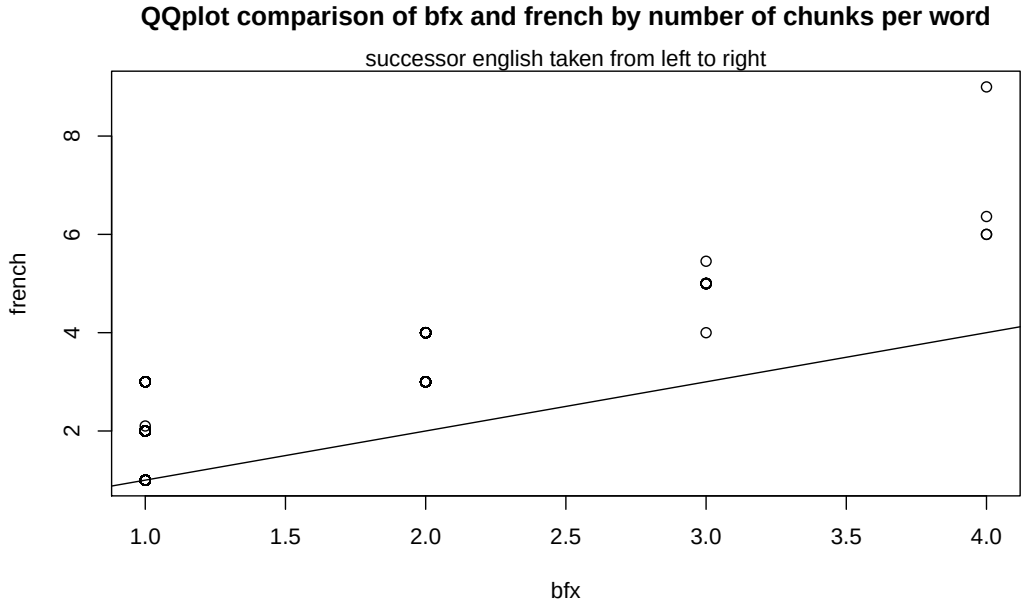




Surprisingly, French also agreed with BFX:

Figure 10: Comparison of BFX and French





In conclusion, if the metrics here are to be trusted, the linguistic origin of the Voynich manuscript appears to be northern Italy, some time between the fall or the roman empire and Italian Unification / the formation of Modern Italian. This is all corroborated by the geological evidence surrounding the Voynich Manuscript, so it is not exactly new information, but I did find it rather interesting!