### Voynich Manuscript

Malika, Grace, George

### Motivation

The Voynich Manuscript provides an interesting opportunity to apply the techniques that we have learned in this class to a dataset that is not a confirmed language. In this paper, we will analyze the language in the Voynich manuscript using the following techniques:

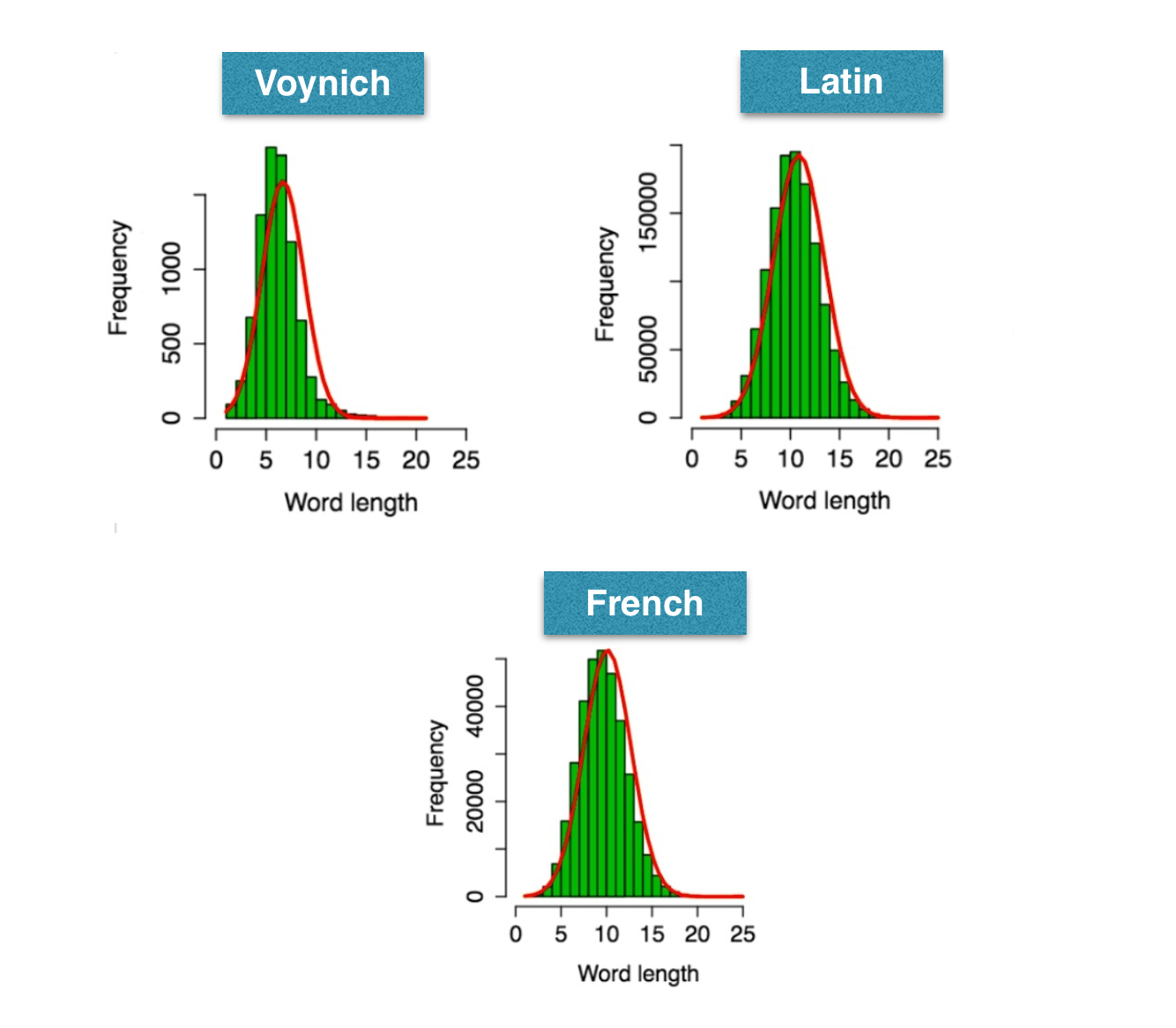
1. word length comparisons

2. hidden Markov models

3. arithmetic compression.

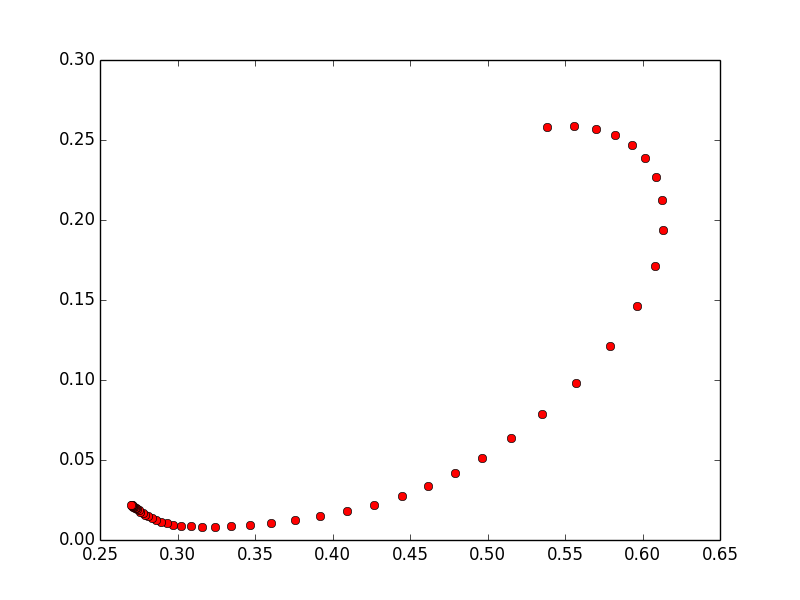
By comparing the above analyses run on the Voynich manuscript to known languages, we believe that there is some evidence that the Voynich manuscript really be a valid language.

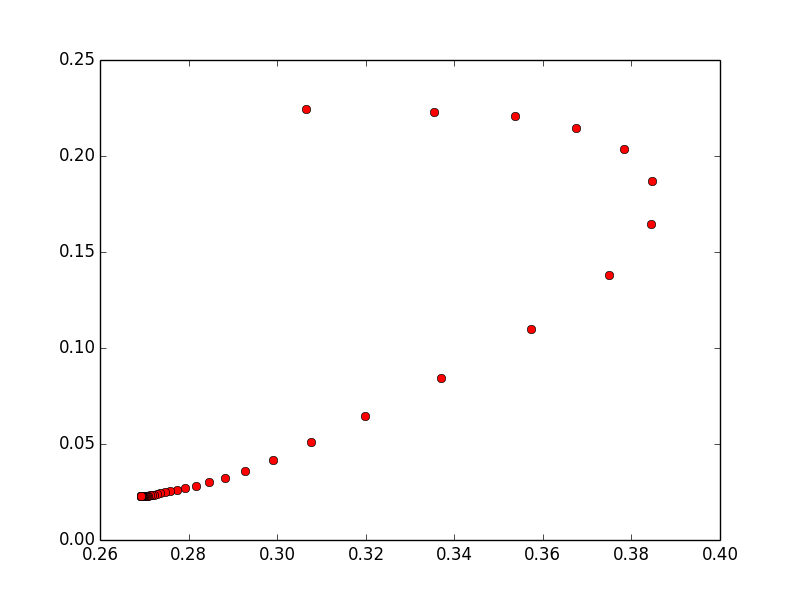
### Word Length Analysis

We begin by looking the distribution of word lengths in Voynich. We compare to the frequency of word lengths in Latin and French.

As evident from the graphs above, we see that the frequency of word length appears roughly similar in the 3 languages. Our analysis identifies the most frequent word length in the Voynich manuscript is around 6 characters. This appears to contrast with the same statistic for Latin and French, which have the most frequent word length at around 9 or 10 characters. While this simple statistic seems already suggest that Voynich may not be a true language, by considering the distribution more holistically (as presented in the diagrams above), we see that the fitted (red) curve for Voynich appears to be similar, though not exactly the same, as Latin and French. Indeed, statistically, simply considering statistics like the mean of the histogram may not be best model to fit to the data, because many real distributions cannot be expressed well by parametric models. Thus a non parametric (“holistic”) consideration may be more appropriate. Therefore, we do not rule out the possibility that the language in Voynich is a real language, and examine other analytic methods.

### HMM

We now run the HMM algorithm for the Voynich data.



The two graphs compare the state transitions a(0,1) and a(1,0) after runs of 50 iterations each. Similar to the results run over a dictionary of English words, the plot moves toward the upper right corner as the maximization algorithm finds optimal probabilities for each of the two states. This suggests the possibility that the letters in the Voynich manuscript could be split into two states - similar to the consonant/vowel split in English. To examine this in greater detail, we look at the log ratio of the emissions from the two states that have been identified.

### Log ratios of emissions from the 2 states

### Positive:

### q 82.3436572959

### 4 63.3855617313

### 5 13.1085603862

### c 9.06943759915

### z 8.90469510507

### p 7.94272359696

### f 5.80681761037

### h 5.46311573334

### s 4.99603600306

### t 4.63075251894

### d 3.64349546882

### o 2.79787590094

### 3 1.28370990834

### 0.775906189335

### - 0.168477602574

### Negative:

### w -243.894698224

### m -112.886720195

### n -107.920317862

### ? -88.6771630732

### # -68.2031683708

### i -63.6249715347

### l -61.3645495985

### j -60.7669051777

### 9 -47.7265553148

### k -44.6522382539

### 7 -4.71462327515

### r -3.22548245287

### a -3.02556248326

### 6 -2.91718501653

### x -2.72971983557

### g -1.22569951688

### y -1.10215798215

### 8 -0.595375433845

### = -0.526399311092

### e -0.210765957404

### 2 -0.0867359585459

By looking at the log ratios, we see that, in Voynich, there appears to be a fairly equal number of letters in each state (unlike English where there are significantly more consonants than vowels). While this does not rule out the possibility that it is a language, it is once difference between Voynich and English, and an important structural characteristic in the language.

We examine one final aspect of the HMM model: the probability of an individual word given the emission and transition probability generated by the HMM.

---------------------

Sample of Results Table (without plog) – probability of each word

---------------------

-4ODCOE# 2.4693970113e-10

-4ODOE# 1.86560202665e-09

-8TOR# 2.19520087882e-08

-GDCIIIR# 2.07185794879e-12

-ODAN# 8.11957725704e-09

-ODCC# 8.22659301245e-08

-ODCCG# 1.1889502434e-08

-SCDG# 2.09420332248e-08

2# 0.00258577136387

2-2# 5.50002370659e-07

2-2SO# 6.14389786614e-10

2-8AN# 2.85142075445e-09

2-8G-8G-GDCCTG# 1.83229592137e-21

2-AM-SAPTC8G# 1.57055670205e-20

2-GHOM# 6.45723789934e-11

2-OD# 7.70156850407e-08

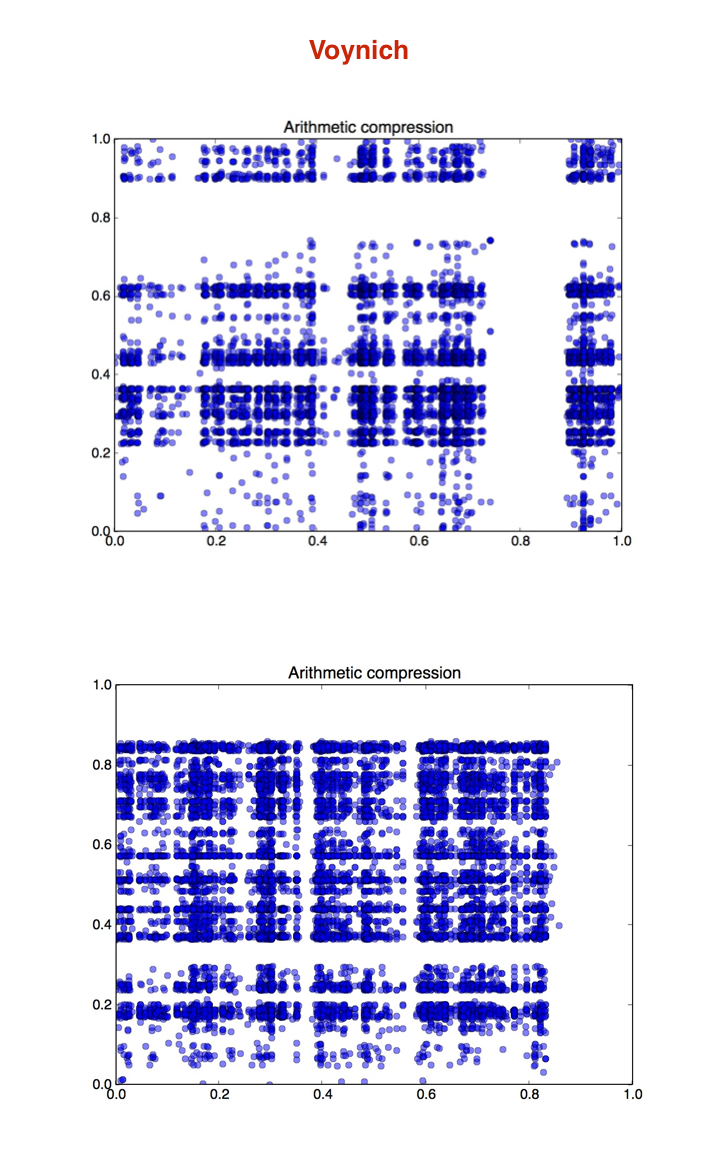
2-ODOE# 1.1121047553e-09

2-OHAM# 2.20418568349e-10

2-SCDG# 3.00146353279e-10

The sample of the final results table of probability of each word further shows the similarly of the Voynich language to a real one given the words all have low probabilities yet a wide range of values, showing some to be more frequent than others without one that truly dominates the set.

### Arithmetic Compression

We finally turn our attention to examining common starting and ending letters in the Voynich manuscript. The first plot is the arithmetic compression for the Voynich manuscript and the second plot is for the English language.

The x-axis of the plots indicates the probability of starting with that letter or phoneme, while y axis indicates probability of ending with it. By looking at both plots, we immediately see some obvious similarities. Namely, that there are clear horizontal and vertical bands in both plots. This indicates that Voynich, like English, has common starting and ending characters. This is additional evidence that Voynich may be a true language, because it appears to have the notion of a suffix (such as “ed” in English) that may modify words. Similarly, the vertical lines may indicate the presence of prefixes in the language, again, suggestive of a syntactic structure similar to many other languages.

### Conclusion

When comparing Voynich to common languages at a static statistical level (e.g., looking at frequency of word length and other static statistics), it appears that there may be evidence that Voynich is not a true language. These statistics are measures for how similar Voynich is to the reference language, but they are not as good at identifying underlying hidden structure in an unknown language that may suggest that it comes from a coherent system that has different “statistics” to known languages. For this sort of analysis, we use the HMM model trained on Voynich and the arithmetic compression analysis. These analyses seem to indicate the Voynich has a strong coherent hidden structure, which while different from common European languages, is sufficient to suggest that it could be a real language.