



Cracking the Voynich Code

Project Group 31: Andrew McInnes and Lifei Wang

Supervisor: Prof. Derek Abbott

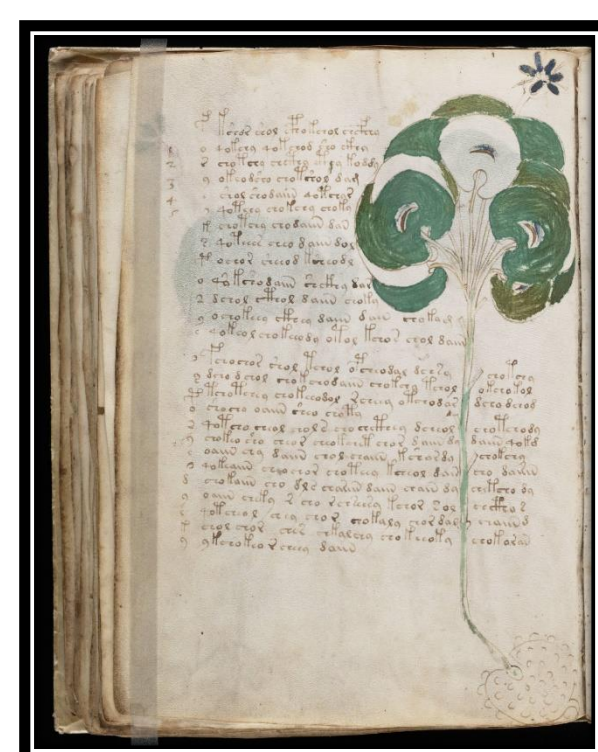
Co-Supervisors: Dr. Brian Ng and Maryam Ebrahimpour

Aim

Determine possible features and relationships of the Voynich Manuscript through data-mining and analyses of basic linguistic features.

Background

The Voynich Manuscript is an unknown script that has been carbon dated back to the early 15th century [1]. Named after Wilfrid Voynich the manuscript has become a well-known mystery within linguistics and cryptology. It is divided into several different sections based on the nature of the illustrations.



Significance

Generally, the project deals with data acquisition, through data-mining, and analyses. Actual analyses can be expanded to be used in plagiarism detection, machine-runnable translators and search engines.

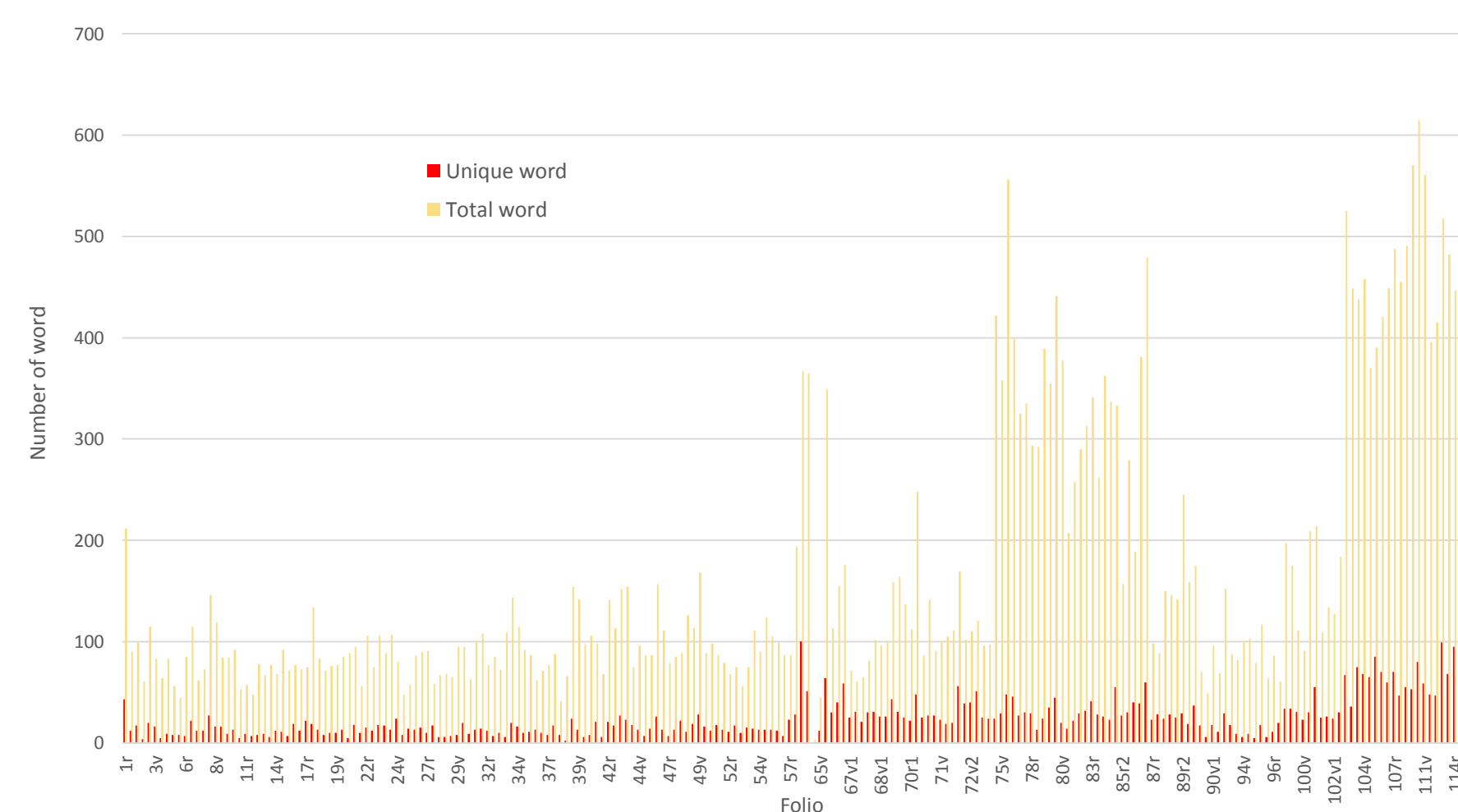
Methods and Processes

Analyses focused around basic linguistics features such as word and character frequencies, character and word bigrams, and character sequences.

Code was developed to be used in MATLAB and C++ environments.

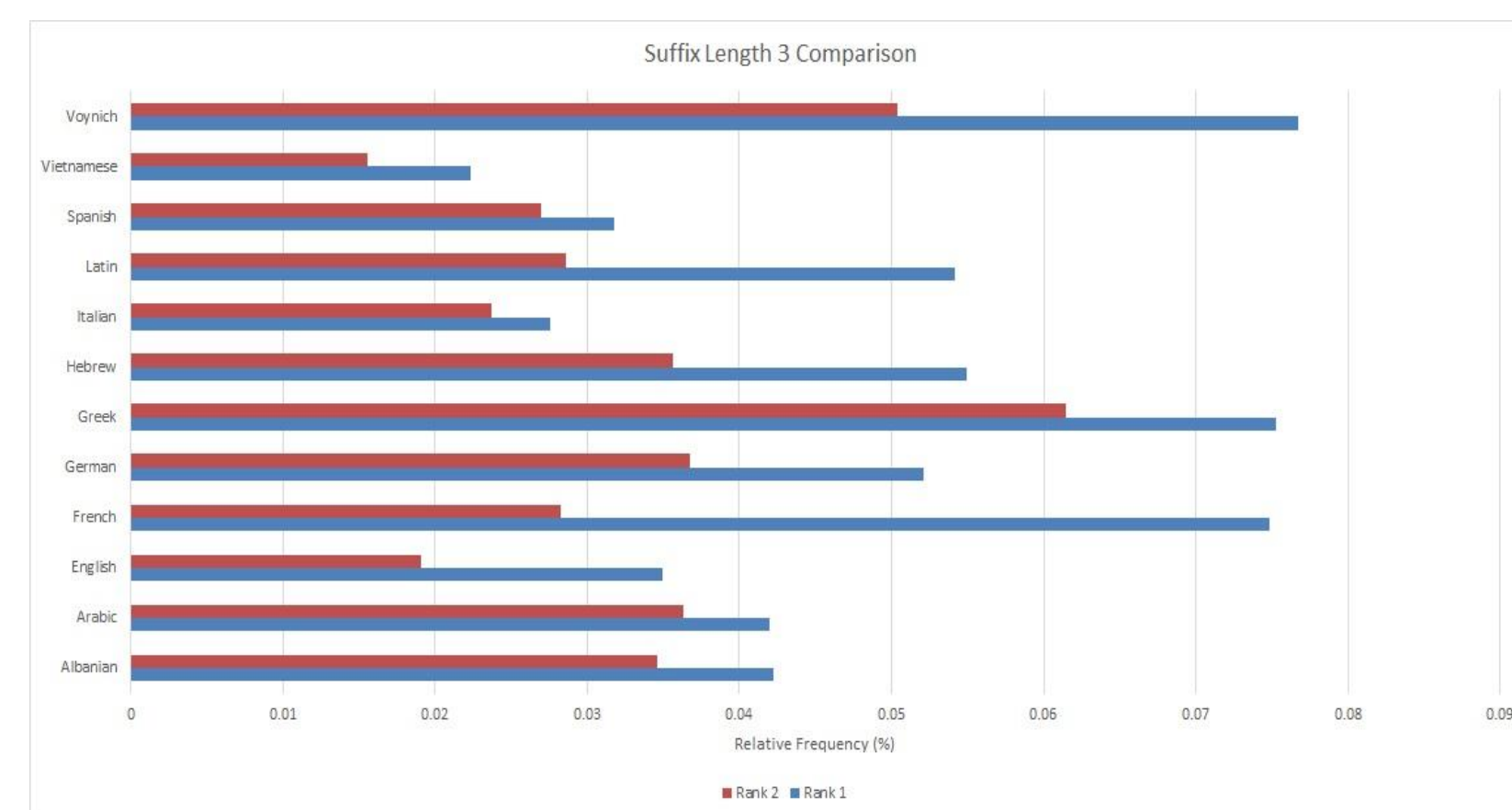
Illustrations

Number of total word VS. number of unique word on each page. The unique word tokens of the page may related to the illustration of the page. The unique word token of each section may correspond to the illustration of the section.



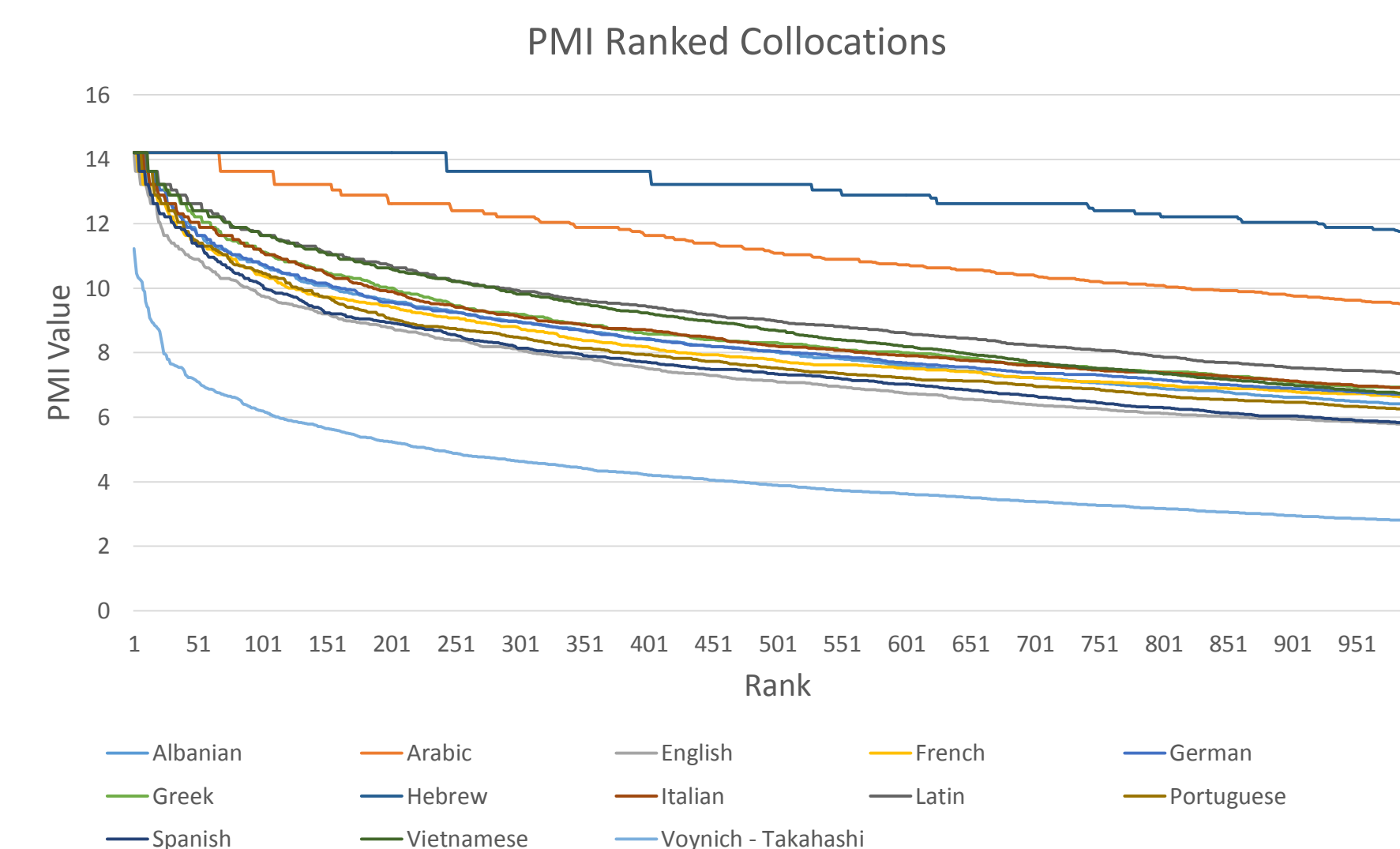
Suffix Frequencies

Character sequences of various lengths at the end of words from various languages were ranked and compared. Greek and French showed the closest relationship from a frequency perspective.



Collocations

Collocations give a quantitative measure of word association. We applied a Pointwise Mutual Information metric to rank the word-pairs of the Voynich and other languages. The Voynich Manuscript shows a low measure of word association with not close relationship to any tested language.



Conclusions

Due to the unknown nature and small sample size of the Voynich Manuscript, it is difficult to definitively make any conclusions without further in-depth research.

- Unique words may relate to specific illustrations
- Greek and French show the closest relationship in regards to suffix frequencies
- The Voynich has a weak measure of word association suggesting it may be related to a type of code or be a hoax.

References

[1] D. Stolte, "Experts determine age of book 'nobody can read'," 10 February 2011. [Online]. Available: <http://phys.org/news/2011-02-experts-age.html>. [Accessed 12 March 2015]