Cracking the Voynich Code

Peter Roush, Bryce Shi (Group 44)
Supervised by Prof. Derek Abbott, Dr. Brian Ng and Maryam Ebrahimpour

THE UNIVERSITY of ADELAIDE

SCHOOL OF ELECTRICAL AND ELECTRONIC ENGINEERING

Purpose

To use computer analysis to compare the features of the unknown language in the Voynich Manuscript with known languages, and develop new analysis techniques for future projects.

Background

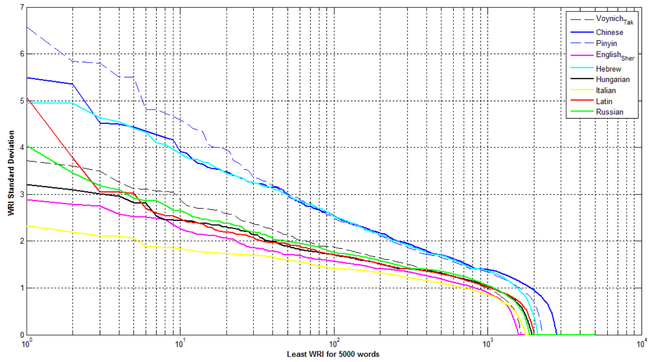
The Voynich Manuscript is a 15th century book written in an unknown alphabet by an unknown author. The illustrations indicate that the book is a herbal or medicinal manual, and also include diagrams of the cosmos. Despite a century of research, no one has deciphered the text.

Importance

The techniques we have worked with are also used for plagiarism detection, author identification, and search engine algorithms.

Word Recurrence Interval (WRI)

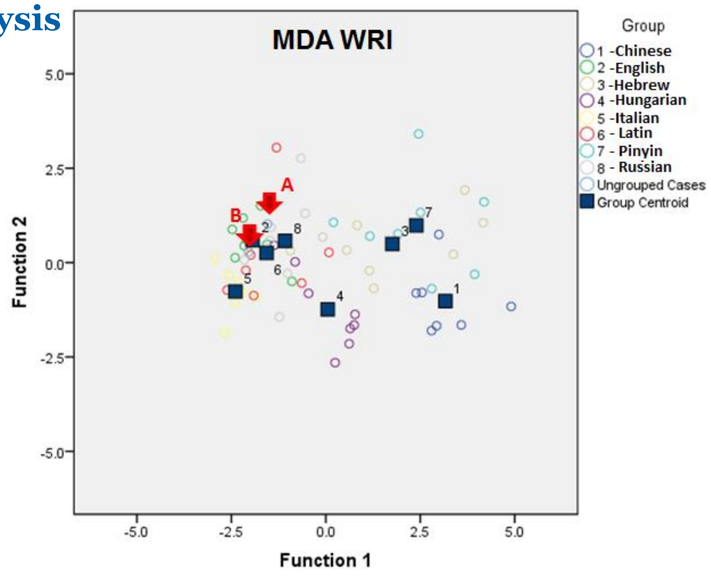
WRI is the number of words between successive repeating words. It is a useful statistic for producing sets of data against other comparison texts as it is language independent. The graph below indicates similarities between the Voynich and European languages



Multiple Discriminant Analysis

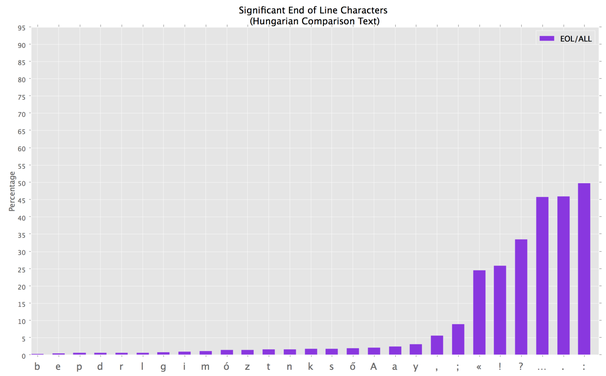
MDA is a group classifier. It does this by setting a rule which provides the most meaningful separation between data.

MDA classified two excerpts from the Voynich manuscript (marked with red arrows) as similar to English based on the WRI, but the close clustering indicates that WRI is a poor classification feature for this data.

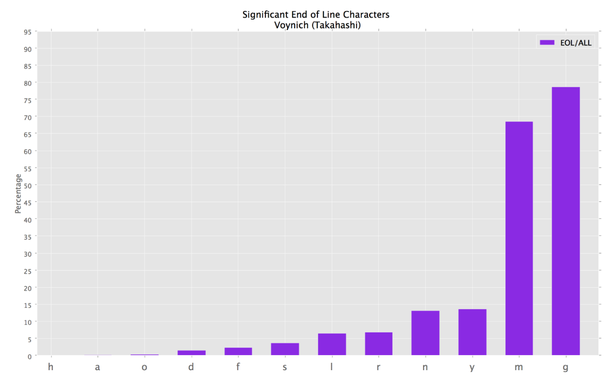


Line Break Characteristics

Based on testing with known languages, such as Hungarian (shown below), it was identified that only punctuation characters occur often at line breaks (i.e. the end of a line). This characteristic was the same for all tested natural languages.



However, in the Voynich manuscript, which has no clear punctuation, the characters with significant line break relationships (shown below) are '**m**' and '**g**'. The manuscript may use these characters as a form of punctuation, or this may indicate that the text is a code instead of a natural language.



Methods and Processes:

Analysis was focused on WRI, word order relationships, and supervised learning algorithms. The large body of previous research on the manuscript was used to develop new analysis techniques, such as line break characteristics.

The majority of code was developed in Python and MATLAB, and documented for use in future projects. MDA comparisons were developed with IBM's SPSS software.

Conclusions

-The text appears as a European language such as English based on the WRI and MDA methods.

-Based on the line characteristics, there may be a form of punctuation present, or characteristics of a code.

-The text has low word order, which may indicate that it is encoded.

Future Research

-Compare the manuscript with transcriptions of known 15th century ciphers.

-Expand research into authorship.

-Develop new features for use with the MDA algorithm.