# Determining the relationship between the letters in the Voynich manuscript splitting the text into parts

AUTHORS: Esbolat Sapargali | Iskander Akhmetov | Alexandr Pak | Alexander Gelbukh

## Abstract

The Voynich Manuscript is an illustrated manuscript code that has not yet been defined the structure of the writing and the relationship to other languages. This study investigated the effectiveness of examining point detail versus examining the full picture all at once in a single study. In the approach of this study, one of these ways, some letter patterns based on frequency and word length were identified, including

connections at different combinations of consonant and vowel letters by a statistical approach for a latent Markov model. A narrowly directed systematic direction can help lead to the unraveling of the manuscript text in progressive steps.

## Introduction

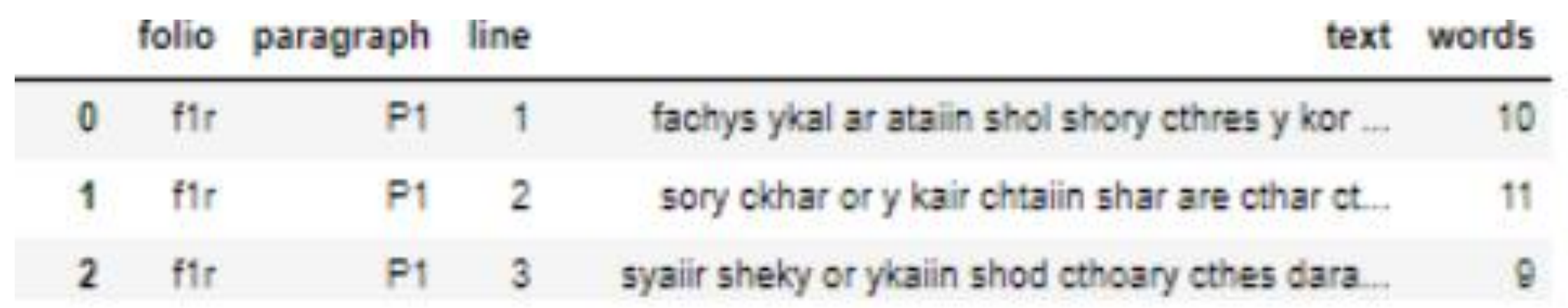
Previous studies (the first attempts still date back to the 16th century) have tried to decipher the text by all known methods of cryptography and linguistics, including advanced technologies such as neural networks. The problem of deciphering it is that scientists have not yet been able to determine the structure of the writing and the relationship to other languages.

The purpose of this article is to determine the effectiveness of studying point details, compared to studying the whole picture at once in a single study. The approach of this study will use methods that allow statistical analysis of words in the text. Here statistical methods will be used for the hidden Markov model. This will be used to find out to what extent there is a pattern in the various parameters of word frequency, word types, and the removal of spaces between words.

## Methodology

The manuscript was studied using a static Hidden Markov Model (HMM). In contrast to entropy and mutual information methods, The HMM can analyze the relations between letters without their prior segmentation, because the sequence of symbols can be extracted from the columns of the scanned symbol image in the same way as from the word image.

The project used scanned images of the manuscript as well as texts divided into different types of parts to determine the relationship of certain letters as well as substitutions for other characters. The data are sorted by several properties: by folio, by line, and by word length.

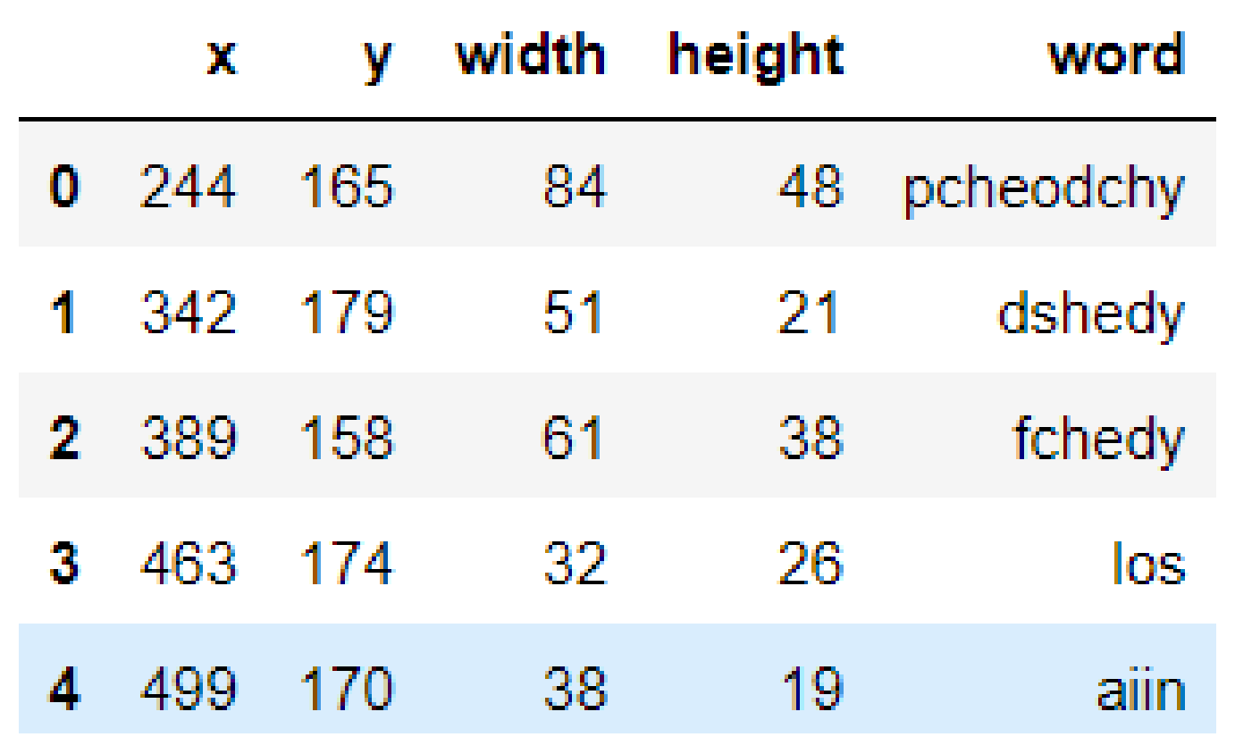


| | folio | paragraph | line | text | words |
|---|---|---|---|---|---|
| 0 | f1r | P1 | 1 | fachys ykal ar ataiin shol shory cthres y kor ... | 10 |
| 1 | f1r | P1 | 2 | sory ckhar or y kair chtaiin shar are cthar ct... | 11 |
| 2 | f1r | P1 | 3 | syaiir sheky or ykaiin shod cthoary cthes dara... | 9 |

Each HMM parameter has a symbolic code that can be used to configure its initialization and evaluation. The input data for training is a matrix of combined observation sequences along with sequence lengths.

During tokenization the text is an array. The function written for tokenization selects strings longer than 4 words. The input, of course, is the list of words we want to tokenize. Then a two-part HMM fitting takes place, and we get the tags. The tags themselves are highlighted in green, and all other occurrences are highlighted in blue.

Word matches in sentences are called bigram frequencies. As you know, for bigrams, the context window is asymmetric by one word to the right of the current word when counting occurrences together. The morpheme boundary runs between the morphemes that make up the word. In some cases, the free stem and suffix are connected by a morpheme boundary, but in most cases the bases are also connected. In the first case, the free base, and in the second case, the connected base.

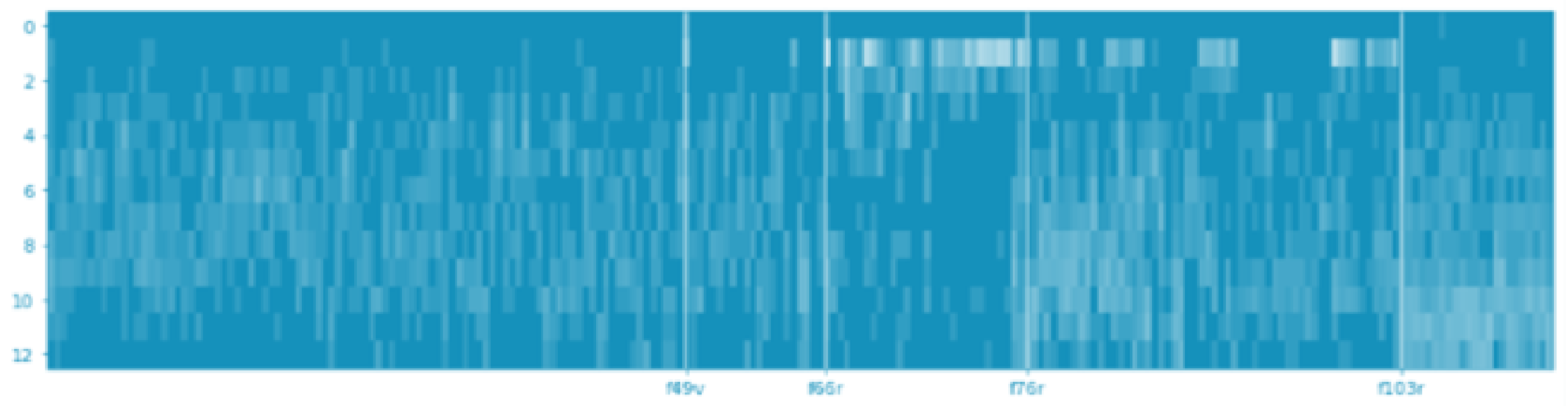


| | x | y | width | height | word |
|---|---|---|---|---|---|
| 0 | 244 | 165 | 84 | 48 | pcheodchy |
| 1 | 342 | 179 | 51 | 21 | dshedy |
| 2 | 389 | 158 | 61 | 38 | fchedy |
| 3 | 463 | 174 | 32 | 26 | los |
| 4 | 499 | 170 | 38 | 19 | aiin |

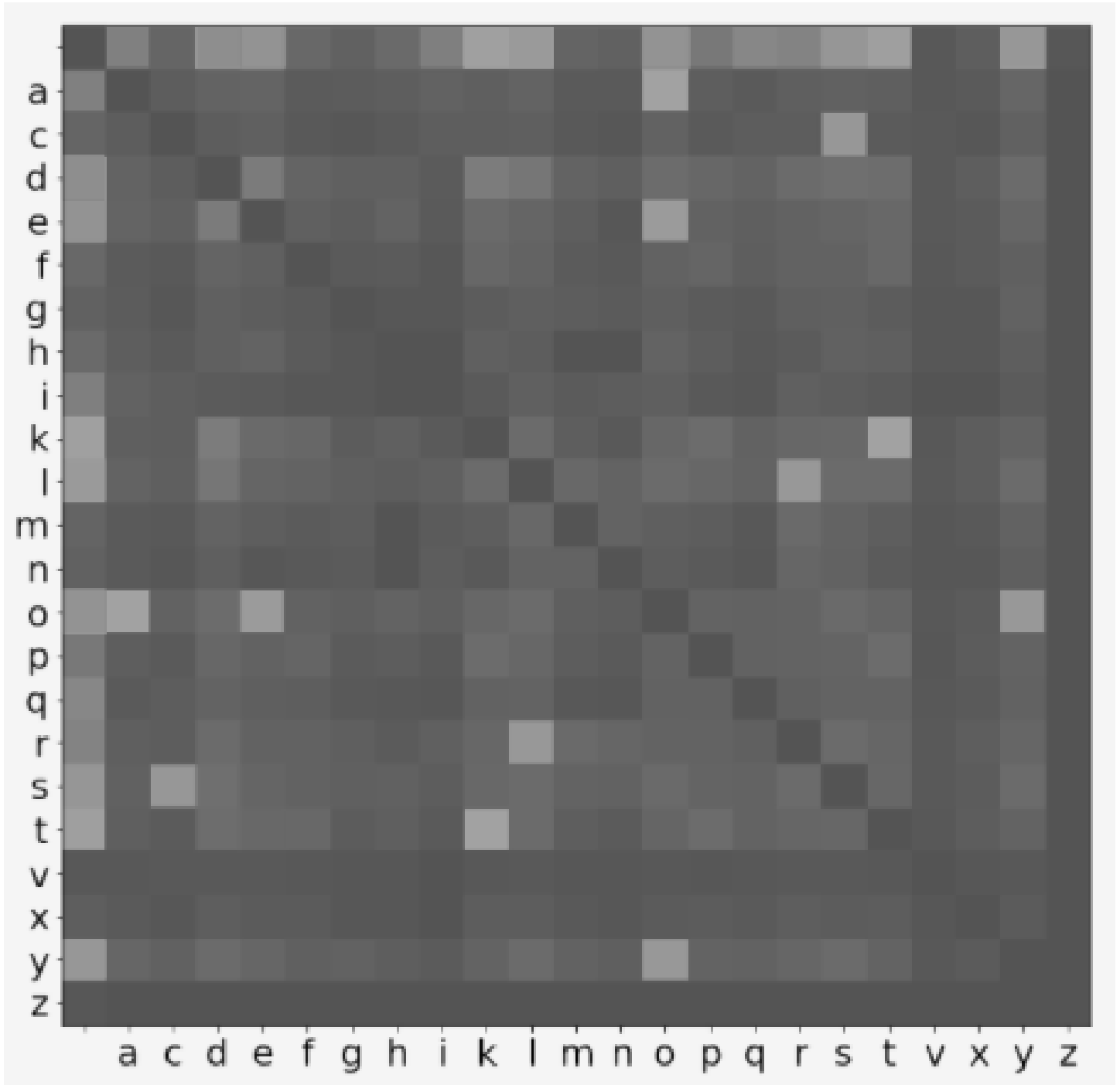
## Results

As a result of the study, the folio integrity check function showed a match after several check attempts. Searching for specific characters and finding a pair gave a successful result. The number of unique words was determined (8078 out of 37886).

After that, we displayed a graph of frequently encountered pairs. With the help of the graph and table it was possible to understand the patterns with the help of visualization functions, looking at the pictures. The folio consists of an average of 5 lines of 5 words each (a long row of properties).



The configuration file generated the best choice of vowels and consonants (some groups of letters were vowels and some were consonants) showed V=ai, e, o, k, C=a, l, dy, n. Loss of vowels between consonants would be 0, with only consonants or only vowels would be 0.1. When in condition 1, the probability of seeing these words will be frequent. But 5.5 percent when going to state 2 other words.



There are letters that are different. Returns context, letters with state 1 or 2. No difference was detected on the first pass. The first insertion was returned, regardless of editing distance. Some letters (d) replace each other with a higher frequency than others. K, t have relations, o replaces a, e and y in many cases. Some groups of letters turned out to be vowels and some groups turned out to be consonants.

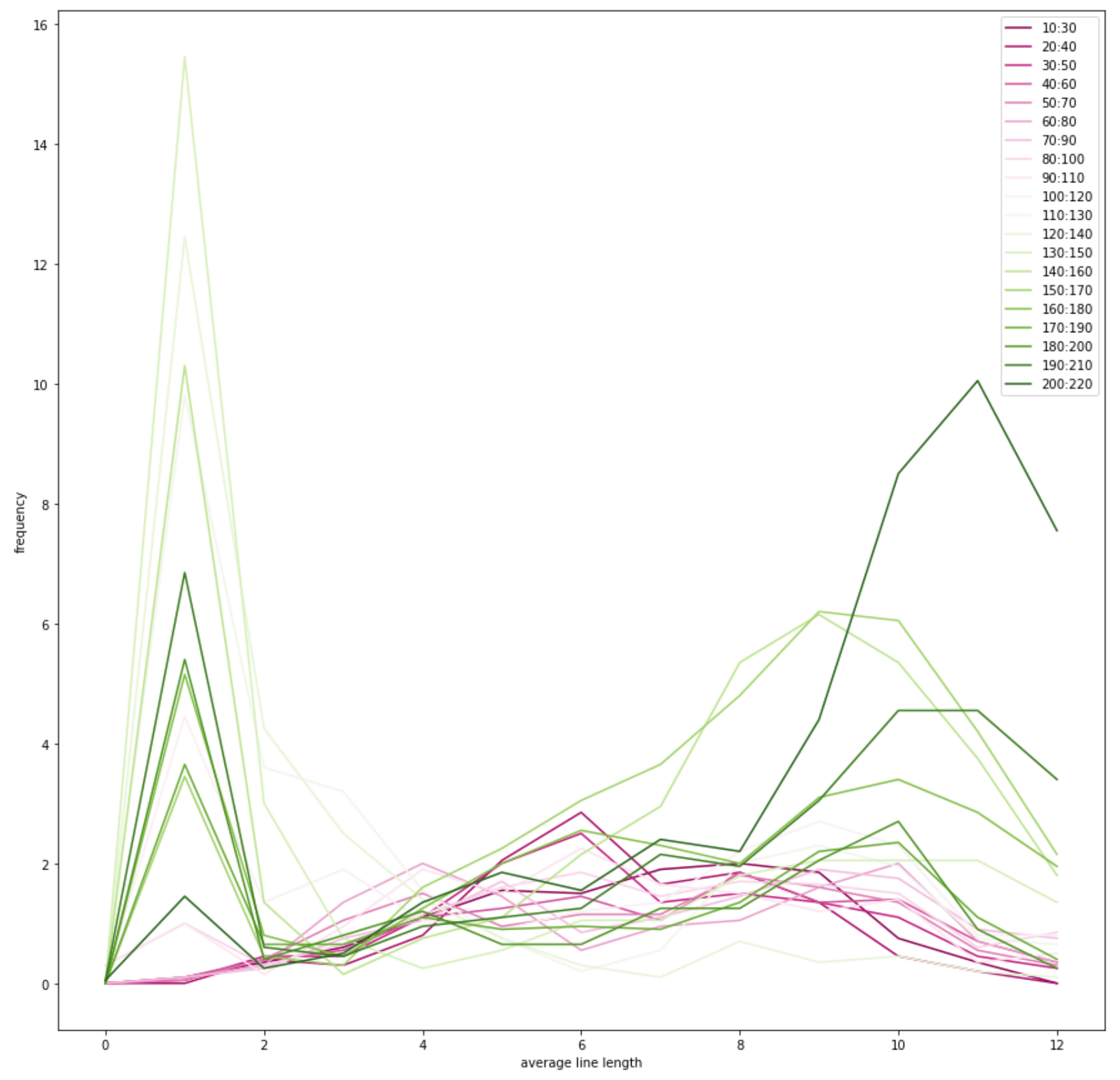
## Conclusion

In this study, scanned images of symbols and words from the manuscript as well as texts divided into different types of parts were examined.

Letter patterns (ratios of certain letters) from the frequency and length of words were determined, including the occurrence of certain letters in English as well as the occurrence of consonants and vowels. Some letters are substituted for each other with greater frequency than others and also have a substituting relationship in some cases.

Some letter patterns based on word frequency and length have been identified, including relationships at different combinations of consonant and vowel letters.

Compared to previous research, it has become clear that different features are required for different applications. The ways in which the Voynich manuscript is handled are varied and depend on applications and languages, which means that they cannot be restricted to any general framework in the subsequent study.



Sliding window with 21 sheets averaged over the entire manuscript

## Acknowledgements

We gratefully acknowledge the financial support of the Ministry of Education and Sciences,Republic of Kazakhstan (Grant num. AP09260670 \Development of methods and algorithms for augmentation of input data for modifying vector embeddings of words").

The work was done with partial support from the Mexican Government through the grant A1-S-47854 of the CONACYT, Mexico and grants 20211784, 20211884, and 20211178 of the Secretar´ıa de Investigaci´on y Posgrado of the Instituto Polit´ecnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnolog´ıas del Lenguaje of the Laboratorio de Superc´omputo of the INAOE, Mexico.