

SCHOOL OF  
ELECTRICAL AND  
ELECTRONIC ENGINEERING



THE UNIVERSITY  
*of* ADELAIDE

# **Project 141:**

## **Cracking the Voynich manuscript code**

### **(Final thesis)**

Student name: Ruihang Feng  
Yaxin Hu

## **ELEC ENG 7076 A/B MASTERS PROJECT**

**M.E. in Electrical and Electronic Engineering**

Each student at Level IV in the School of Electrical and Electronic Engineering is required to complete a final-year design or masters project. The course involves approximately 300 hours of project work over the whole academic year. Students are assessed on their performance in the project, the quality of their outcomes, two progress reports, a final report, two seminars and a project exhibition.

Date submitted: 31 Oct 2016

Supervisor: Prof. Derek Abbott

Adviser: Dr. Brian Ng

## **Acknowledgement**

First of all, we want to express our gratitude to supervisor Professor Derek Abbott. With his help, our project can be completed on time. In the course of the project, he can always give us kindly suggestions.

Secondly, we want to express our gratitude to co-supervisor Doctor Brian Ng. In the course of our project, he provided many useful methods.

### **Abstract**

The aim of this project is to crack the Voynich manuscript which is an unknown hand-written book. This book is considered to be an unknown language, cipher code or hoax. Thesis proposal is aimed to provide methods in determining possible features of the Voynich manuscript. All the methods are related to data mining, computer coding and statistical methods. There will be specific explanation of the methods that will be carried out in the whole project. Furthermore, this document provides the management of this project. In the final part, some possible hypotheses were given according to the whole searching so that it will provide breakpoint in cracking the Voynich manuscript

## Content

<b>1. Introduction</b> .....	<b>6</b>
1.1 Background.....	6
1.2 Aim .....	6
1.3 Motivation .....	7
1.4 Significance .....	7
1.5 Technical Background .....	7
1.6 Knowledge Gaps .....	8
1.7 Technical Challenges .....	8
<b>2. Related Work (the history of the Voynich manuscript research)</b> .....	<b>9</b>
<b>3. Requirements</b> .....	<b>11</b>
<b>4. Proposed Method</b> .....	<b>12</b>
4.1 Phase 1: Text investigation .....	12
4.2 Phase 2: Illustration investigation .....	12
4.3 Phase 3: Marginal symbol research.....	12
<b>5. Project management</b> .....	<b>14</b>
5.1 Deliverables .....	14
5.2 Work breakdown.....	14
5.3 Timeline .....	15
5.4 Task allocation.....	15
5.5 Management strategy.....	15
5.6 Budget.....	16
5.7 Risk analysis .....	16
5.7.1 Mismanagement of time .....	16
5.7.2 Loss of data or files .....	16
5.7.3 Team member's quit.....	16
5.7.4 Lack of references .....	16
5.7.5 Health issues .....	16
<b>6. Results</b> .....	<b>17</b>
6.1 Phase 1: Text investigation .....	17
6.1.1. The total number of words .....	17
6.1.2. The frequency of words .....	17
6.1.2.1: The frequency and the number of simple letters .....	17
6.1.2.2: The frequency of words.....	19
6.1.2.3: Comparing the Voynich manuscript with other known languages .....	22
6.1.3. Digits.....	33
6.2 Phase 2: Illustration investigation .....	36
6.2.1 Searching initial numbers and possible numerical words .....	36
6.2.2 Mapping all initial numbers and numerical words .....	37
6.3 Phase 3: Marginal symbol research.....	45

6.3.1 Statistics for marginal stars of each page .....	45
6.3.2 Digits mining .....	45
6.3.3 Conclusion .....	47
<b>7. Comment on progress.....</b>	<b>50</b>
<b>8. Conclusion .....</b>	<b>51</b>
<b>9. References .....</b>	<b>52</b>
<b>10. Appendix .....</b>	<b>54</b>

## 1. Introduction

### 1.1 Background

The Voynich manuscript is a document written in unknown alphabets that was found by Wilfrid Voynich (1865-1930) in 1912 [1]. Because of the Voynich manuscript's long history, some pages of manuscript were missing. As the result, there are almost 240 pages remaining [2]. In addition, the folios of the manuscript were numbered from f1 to f116 and each folio involved two pages, r and v.

It is divided into six different sections by illustrations with different styles and images:

a) **Herbal:**

There are one or more plants on each page, which is a format of European herbals.

b) **Astronomical**

There are circular diagrams such as suns, moons, and stars which suggest this part as something about astronomy or astrology.

c) **Biological**

Mostly naked women show that this part should be biological section.

d) **Cosmological**

Circular diagrams of obscure nature make this section as cosmological section.

e) **Pharmaceutical**

Drawings of isolated plants parts and objects resembling apothecary jars show that this section should be something about pharmaceutical.

f) **Recipes**

These parts are full pages of text in short paragraphs.

Generally, the Voynich manuscript was made up of three parts: text, illustrations and marginal symbols.

### 1.2 Aim

The aim of project is using the statistics and comparison to infer that the Voynich manuscript is code, nature languages, constructed languages, cipher code or hoax from the perspective of digits.

In addition, the aims of this project also involve cracking the initial digits of the

Voynich manuscript and determining the possible letters which may stand for digits.

Due to the massive number of words and illustrations in the manuscript, it is unnecessary to solve the whole manuscript in a one year project.

### **1.3 Motivation**

In the field of linguistics, the Voynich manuscript is a representative. Researchers deem that there is a kind of useful information among the mysterious alphabets of manuscript.

In the course of this project, statistics and comparison will be applied to crack the Voynich manuscript. If the manuscript can be cracked successfully, the results of this project will be useful for linguists to compare other unknown languages.

### **1.4 Significance**

There are many guesses about the Voynich manuscript. Because of the manuscript's long history, many historians believe that the mysterious alphabets of the Voynich manuscript are related to ancient civilizations [3]. If manuscript can be cracked, the Voynich manuscript will be helpful for historians to explore the culture of ancient society.

In addition, the statistical method which will be used in this project is also useful in other fields, such as engineering, finance and architecture. Moreover, comparison is widely used, such as Turn-It-In, Google translate, Grammarly and Bing.

### **1.5 Technical Background**

The major technique which will be applied in this project is data mining. Data mining is an effective method to search laws among the massive number of data and has a fantastic performance. The two major methods of data mining are statistics and comparison. Statistics is used to count the frequency of the occurrence of some special words. Comparison is served to find out relations between two languages.

In the field of linguistics, European Voynich Alphabet (EVA) is a representative digital transcription of the Voynich manuscript [4]. Then a Japanese linguist Takahashi organised the whole Voynich manuscript by using EVA [5].

Therefore, major data will be extracted from the transcription of Takahashi in the process of this project.

Moreover, other resources will be considered, such as expressions of some representative ancient languages.

## 1.6 Knowledge Gaps

Due to the massive amount of data in the Voynich manuscript, the project requires skilled data processing technique and software programming capabilities; however, no one in this project team has ever dealt with so much data. Hence members should develop data processing ability and software programming skills.

On the other hand, the project requires particular knowledge about statistics, so members must be adept at sorting data.

## 1.7 Technical Challenges

Technical challenges of this project involve two aspects.

First of all, it is very difficult to infer which language the author used. The language of the manuscript does not belong to any known languages [6] and even this language may have been extinct. What is more, due to the long history of the Voynich manuscript, some important information is nowhere to be searched, such as exact information about author. In that case, it is difficult to infer which language the author used from the author's nationality. In order to solve the above problem, members must search many different languages as references and compare those languages with the language of the manuscript.

Secondly, references of cracking the Voynich manuscript are limited. Because of unknown language and mysterious illustrations in manuscript, it is difficult to crack the whole manuscript. Although there are very few words have been cracked by researchers, on one can guarantee that the results are right. In the field of linguistics, there are not recognized correct results about cracking the Voynich manuscript. In that case, it is hard to find reliable references. So members must search references from different ways and find out enough accurate references.



## **2. Related Work (the history of the Voynich manuscript research)**

In the past few years, many researchers had tried to crack the Voynich manuscript by using different methods.

### **Mary E. D’Imperio:**

In 1975, Mary E. D’Imperio was introduced to the problem of the Voynich manuscript by John Tiltman [7]. In the following years, she summed up different features of the Voynich manuscript text [8].

### **Nick Pelling:**

Nick Pelling published his book ‘The course of the Voynich’ at 2006. Based on the illustrations in the rosettes folio of the Voynich manuscript, he believed that the manuscript originated from Milan [9].

### **William Ralph Bennett:**

William Ralph Bennett, a Yale professor, searched the Voynich manuscript with computer. He focused on the research of text by using statistical method. Probably he was the first to note the low entropy of the Voynich manuscript text. As the result, the only language he found with entropy similar to the Voynich manuscript was Hawaiian [10].

### **John Tiltman:**

John Tiltman was a British intelligence specialist. He cracked the text part of the Voynich manuscript with William Friedman. At last, Tiltman and Friedman suggested that the text of manuscript was a kind of artificial (constructed) language [11].

### **Feely:**

Joseph Martin Feely was a Rochester lawyer. In 1943, Feely published a book which involved some solutions of cracking the Voynich manuscript. His solutions showed a viable method to use Latin to replace some words in the manuscript [12].

### **First study group:**

The first study group (FSG) was founded at 1944, dissolved at 1946 [13]. Members of this organization involve:

- Robert A.Caldwell
- G. E. McCracken

- Tomas A. Miller
- Frances Puckett, later Frances Wilbur
- Mark Rhoads
- William M. Seaman

Under the joint efforts of those researchers, the FSG transcribed most parts of the Voynich manuscript and devised a transcription alphabet [14]. The details of the transcription alphabet are as shown in the Appendix section A.1.

### 3. Requirements

Although it is not necessary to crack the whole manuscript, there are some basic requirements as following:

- Text investigation: find out linguistic laws from some paragraphs of the Voynich manuscript. Such as the total number of words, the frequency of some special words and the frequency of some special single letters. Then the Voynich manuscript will be compared with other known languages.
- Illustration research: look for laws from some illustrations from the perspective of digits. Such as statistics for illustrations of each page and digits analysis.
- Marginal symbols investigation: make a thorough inquiry about marginal symbols from the perspective of digits.
- Code run smoothly.
- Evaluation for results.
- Make some assumptions which are helpful for the further research.

## **4. Proposed Method**

As shown in the Appendix section A.2, the proposed methods of this project are divided into three phases.

### **4.1 Phase 1: Text investigation**

There are two parts in this phase: words and digits.

During the process of words research, Matlab will be used as an essential tool. Team members will attempt to search laws from three aspects:

- The total number of words in the Voynich manuscript.
- The characters and words which may stand for digits from some paragraphs of the manuscript.
- The frequency of special characters and words.

On the other hand, in the course of digits investigation, team members will search for different kinds of known expressions of digits and make a comparison with the words in the Voynich manuscript. For example, the expression of digits in Roman is as shown in the Appendix section A.3. The word which is as shown in the Appendix section A.4 is extracted from the Voynich manuscript, it is obvious that the form of the word in the Appendix section A.4 is like “\*##”. According to the method of comparison mentioned above, this word may mean ‘seven’ in Roman.

### **4.2 Phase 2: Illustration investigation**

An illustration which is extracted from the Voynich manuscript is as shown in the Appendix section A.5.

In this phase, illustrations will be analysed by using Matlab. Generally, there are three aspects which are needed to be completed:

- The number of different elements in the illustrations.
- The characters which may stand for digits.
- Match the characters and digits.

### **4.3 Phase 3: Marginal symbol research**

A page which contains marginal symbols is as shown in the Appendix section A.6.

This phase also requires proficiency in programming by using Matlab. During the process of this phase, there are four major aspects:

- Ordering and quantitative features of the marginal symbols of each page.
- Search the characters which may stand for digits.
- The differences between marginal symbols in each page.
- Match the characters and digits and make inference about the relationship between characters and digits.

## 5. Project management

### 5.1 Deliverables

As shown in Table 1, deliverables involve eleven parts.

Table 1: Deliverables

Deliverable	Deadline
Proposal seminar	4st of April, 2016
Project wiki (introduction)	Semester 1, week 5.
Thesis (1 <sup>st</sup> draft)	22th of April, 2016.
Thesis (2 <sup>nd</sup> draft)	Semester 1, week 12.
Master thesis (final)	Semester 2, week 11.
Expo Poster	Semester 2, week 11.
Project wiki (full)	Semester 2, week 12.
Expo presentation	Semester 2, week 12.
YouTube video	Semester 2, week 12.
USB flash drive (all codes and works)	Semester 2, week 12.
Final seminar	Semester 2, week 13.

### 5.2 Work breakdown

The details about tasks are as shown in the Appendix Section A.2. The key tasks involve three aspects:

- Text investigation (digits).
- Illustrations research.
- Marginal symbols investigation.

### 5.3 Timeline

Timeline of project involves six parts. The specific details are as shown in the Table 2.

Table 2: Timeline

NO.	Task	Week
Semester 1		
1	Background research	1
2	Phase 1: Text analysis	6
3	Phase 2: Illustration investigation	8
4	Phase 3: Marginal symbol research	10
Semester 2		
5	Phase 2: Illustration investigation	5-10
6	Phase 3: Marginal symbol research	5-10

### 5.4 Task allocation

Task allocation is divided into six parts:

Table 3: Task allocation

No	Task	Student
Semester 1		
1	Background research	Ruihang Feng, Yaxin Hu
2	Phase 1: Text analysis	Ruihang Feng, Yaxin Hu
3	Phase 2: Illustration investigation	Yaxin Hu
4	Phase 3: Marginal research	Ruihang Feng
Semester 2		
5	Phase 2: Illustration investigation	Yaxin Hu
6	Phase 3: Marginal research	Ruihang Feng

### 5.5 Management strategy

Team members will be managed through a minimum of two internal meetings every week, and a minimum of one fortnightly meeting with supervisors. In addition, the preparation for each meeting involves three aspects:

- Achievements in the past two weeks.
- Questions about the work of the past two weeks.
- Plan for next two weeks.

After meeting, there are two tasks:

- Meeting review.

- Code modification.

## 5.6 Budget

Budget involves four aspects:

- 500 AUS dollars for team members.
- Research need to be carried out further research.
- All programs that need to be used are available on university system.
- All major works can be achieved by using computer.

## 5.7 Risk analysis

Details of risk analysis are as shown in the Table 4.

Tabl4 4: Risk analysis

No.	Risk	Probability	Impact
1	Mismanagement of time	Moderate	High
2	Loss of data or files	Low	High
3	Team member's quit	Low	High
4	Lack of references	High	High
5	Health issues	Moderate	Moderate

### 5.7.1 Mismanagement of time

Due to other works in daily life, the mismanagement of time may occur. Hence each member should arrange the time in advance to avoid time clash.

### 5.7.2 Loss of data or files

During the process of project, there may be some accidents, such as code lost or failure of files storage. In order to avoid that kind of situation, team members should buy two or more USB flash drive to store the backup files.

### 5.7.3 Team member's quit

In order to avoid this case, team members should keep frequent contact with each other.

### 5.7.4 Lack of references

As the mentioned before, the references of the Voynich manuscript are limited. So members should expand the scope of research, such as Bing, Grammarly and other websites.

### 5.7.5 Health issues

Members should pay attention to regular work and break to prevent health problems.



## 6. Results

### 6.1 Phase 1: Text investigation

As the introduction in the section 5.4, ‘Text investigation’ is a cooperative task.

#### 6.1.1. The total number of words

In this stage, Matlab is used to count the total number of words in the Voynich manuscript. The results are shown in the Table 5.

Table 5: The total number of the Voynich manuscript

Book name	Total characters number	Total words number (TWN)	Unique words number (UWN)	Ratio (UWN/TWN)	Average number of characters per word
The Voynich manuscript	234507	37104	8486	0.229	6.32

According to the Table 5, the total characters number of the Voynich manuscript is 234507. The total words number is 37104. The unique words number is 8486. The average number of characters per word is 6.32.

#### 6.1.2. The frequency of words

In this stage, Matlab is used to count the frequency of words in the Voynich manuscript and statistics is used to analyse the characteristics of the manuscript. In addition, this phase is divided into three parts:

- The frequency and the number of simple letters.
- The frequency of words.
- Comparing the Voynich manuscript with other known languages.

##### 6.1.2.1: The frequency and the number of simple letters

The results are shown in the Figure 1, Figure 2 and Figure 3.

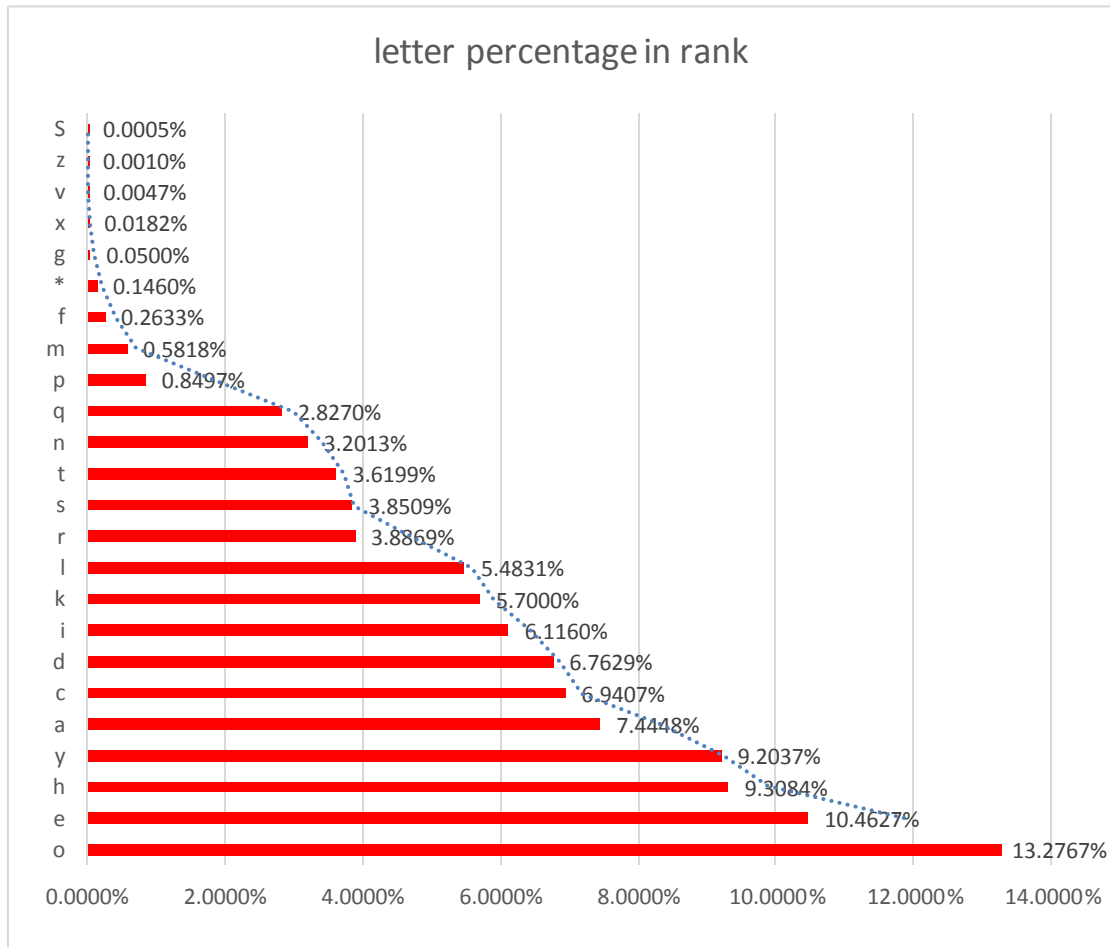


Figure 1

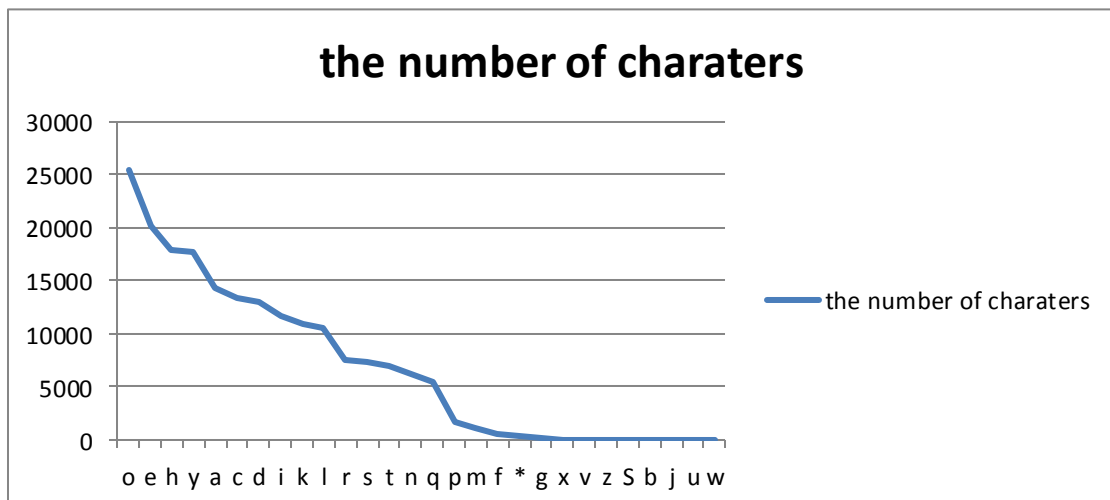


Figure 2

characters	the number of letters	frequency of letters
o	25468	13.2767%
e	20070	10.4627%
h	17856	9.3084%
y	17655	9.2037%
a	14281	7.4448%
c	13314	6.9407%
d	12973	6.7629%
i	11732	6.1160%
k	10934	5.7000%
l	10518	5.4831%
r	7456	3.8869%
s	7387	3.8509%
t	6944	3.6199%
n	6141	3.2013%
q	5423	2.8270%
p	1630	0.8497%
m	1116	0.5818%
f	505	0.2633%
*	280	0.1460%
g	96	0.0500%
x	35	0.0182%
v	9	0.0047%
z	2	0.0010%
S	1	0.0005%
b	0	0.0000%
j	0	0.0000%
u	0	0.0000%
w	0	0.0000%

Figure 3

As shown in the figures above, it is obvious that the frequencies of the simple letter 'b', 'j', 'u' and 'w' equal to zero, which means these letters have never appeared in the Voynich manuscript. In addition, the letter with the highest frequency (0.133) is 'o'.

#### 6.1.2.2: The frequency of words

The results are shown in the Figure 4.

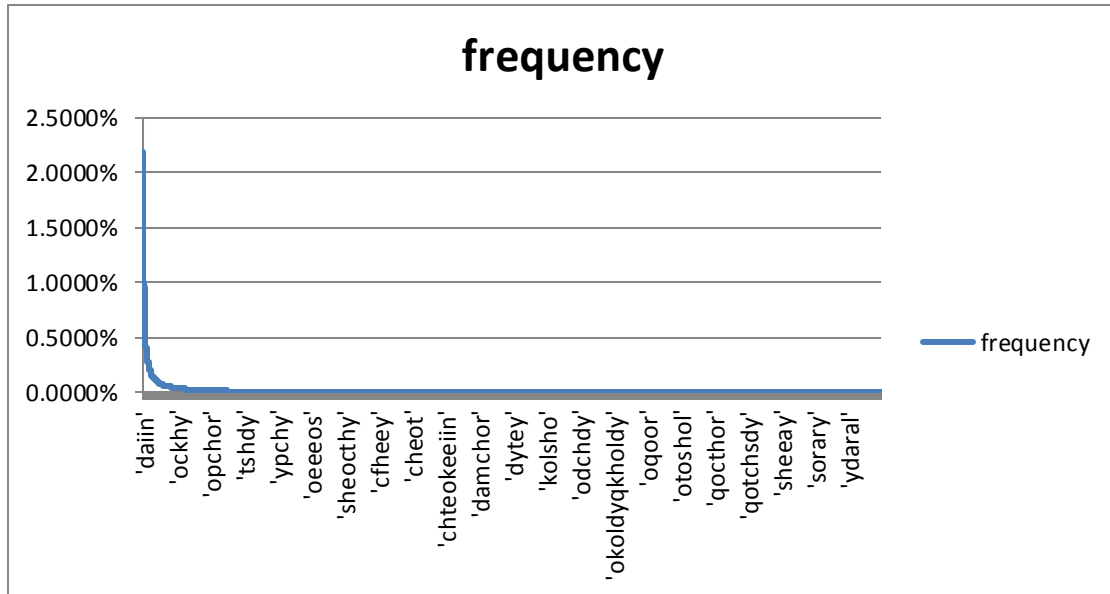


Figure 4

In the Figure 4, x axis means the words in the manuscript, y axis means the frequency. Because there are almost 8486 unique words in the Voynich manuscript, so the x axis in the Figure 4 can't show every word. In order to analyse the words with high frequency accurately, we try to extract the first 100 words. The results are shown in the Figure 5.

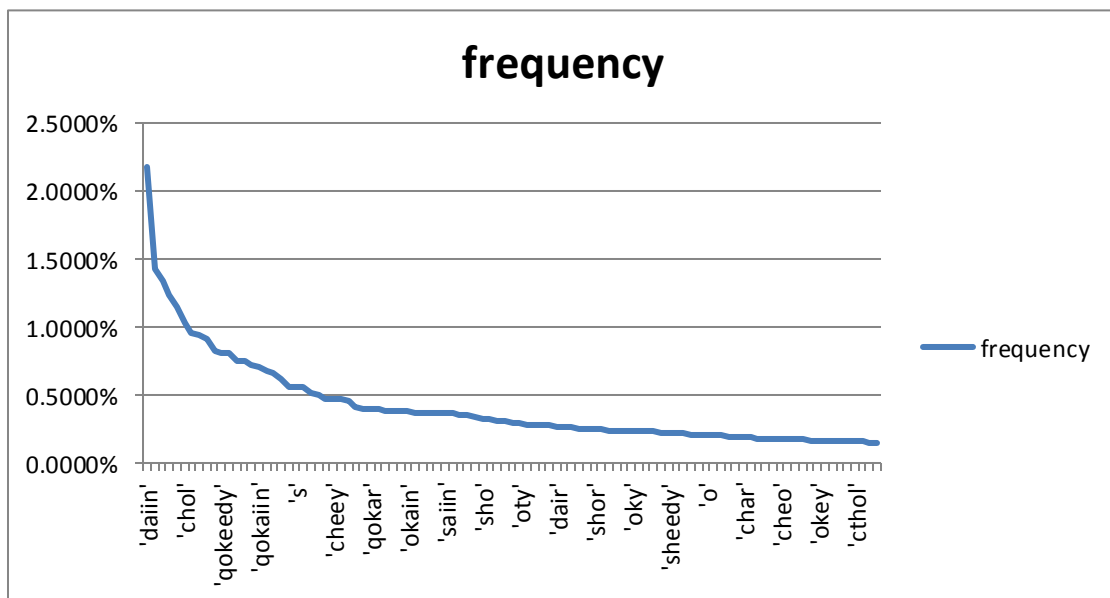


Figure 5

As shown in the Figure 5, the line keeps a downward trend and tends to be stable, which means the frequencies of the last few words are very low. But the x axis still

can't show every word. In that case, we extract the first 20 words. The results are shown in the Figure 6 and Figure 7.

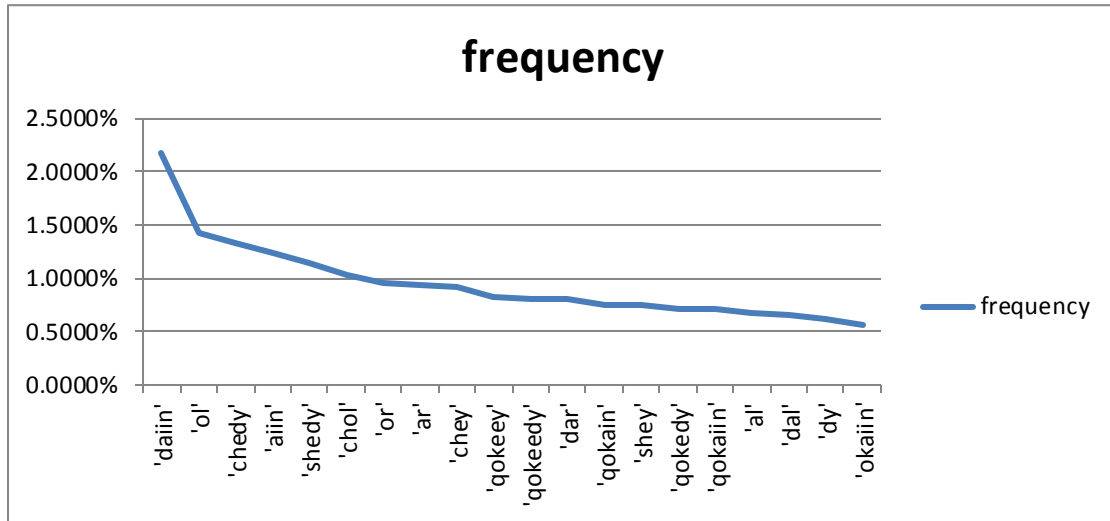


Figure 6

1	words	occurrence number	frequency
2	'daiin'	807	2.1750%
3	'ol'	528	1.4230%
4	'chedy'	495	1.3341%
5	'aiin'	457	1.2317%
6	'shedy'	424	1.1427%
7	'chol'	381	1.0268%
8	'or'	354	0.9541%
9	'ar'	348	0.9379%
10	'chey'	339	0.9136%
11	'qokeey'	308	0.8301%
12	'qokeedy'	301	0.8112%
13	'dar'	298	0.8031%
14	'qokain'	277	0.7466%
15	'shey'	276	0.7439%
16	'qokedy'	265	0.7142%
17	'qokaiin'	262	0.7061%
18	'al'	253	0.6819%
19	'dal'	243	0.6549%
20	'dy'	229	0.6172%
21	'okaiin'	209	0.5633%

Figure 7

From the figures above, the word with the highest frequency (0.022) is 'daiin'.

### 6.1.2.3: Comparing the Voynich manuscript with other known languages

As the introduction in the section 1.1, the Voynich manuscript was found in 1912. During the period of 17 Century to 18 Century, the most commonly languages are Latin, English, French and German [15]. So in this section, the project team search some references about the frequency of commonly used letters in those four languages and compare the Voynich manuscript with those four kinds of languages [16].

#### Part 1: The Voynich versus Latin.

The results of the occurrence frequency of letters in the Voynich manuscript and Latin are shown in the Figure 8 and Figure 9.

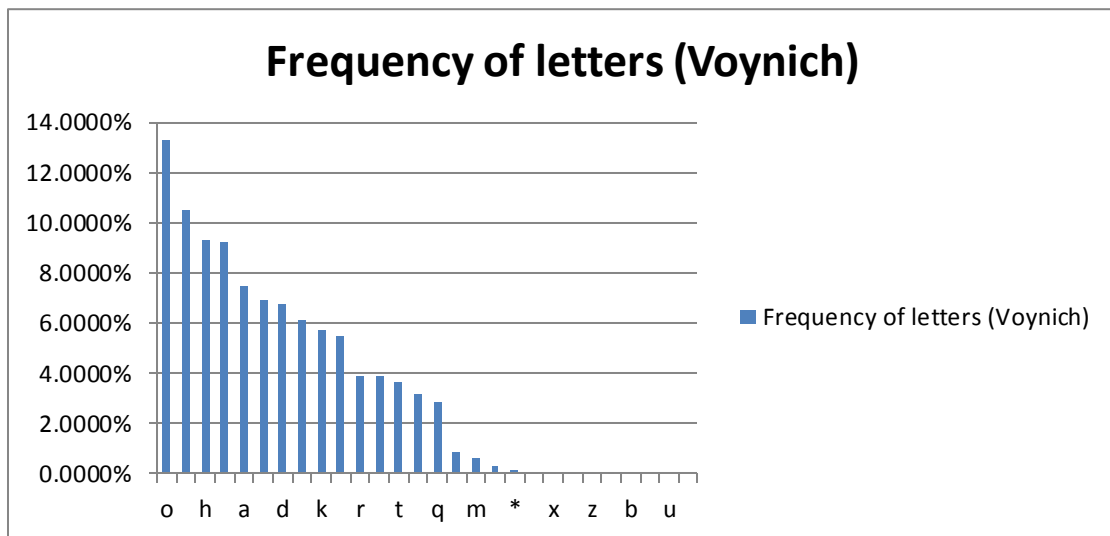


Figure 8

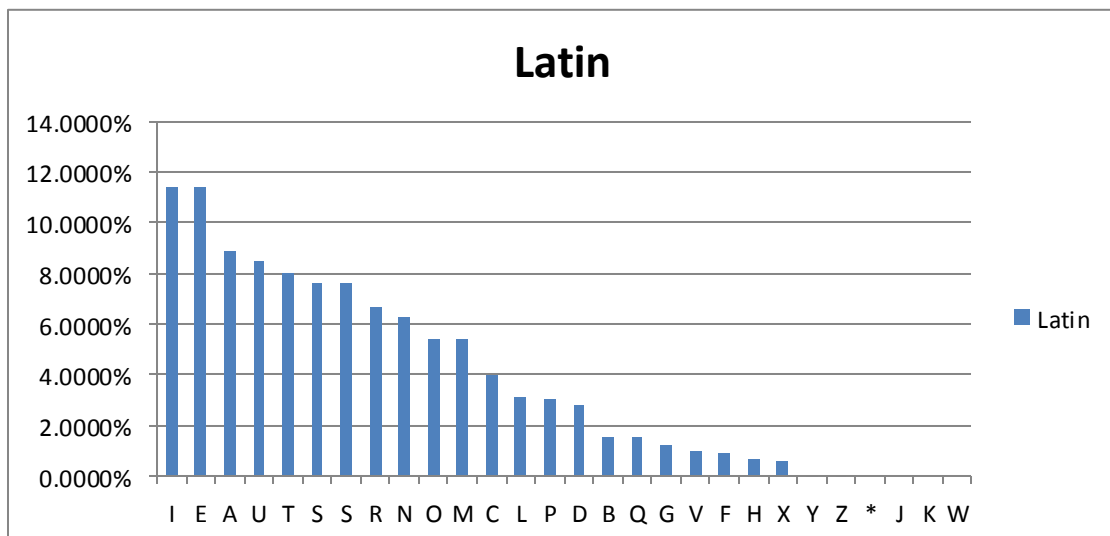


Figure 9

In order to analyse conveniently, the Figure 8 and Figure 9 are changed to the form of proportion, which is shown in the Figure 10.

Characters (Voynich)	Frequency of letters (Voynich)	Characters Latin	Frequency of letters (Latin)
o	13.2766%	I	11.4400%
e	10.4626%	E	11.3800%
h	9.3084%	A	8.8900%
y	9.2037%	U	8.4600%
a	7.4448%	T	8.0000%
c	6.9407%	S	7.6000%
d	6.7629%	S	7.6000%
i	6.1160%	R	6.6700%
k	5.7000%	N	6.2800%
l	5.4831%	O	5.4000%
r	3.8869%	M	5.3800%
s	3.8509%	C	3.9900%
t	3.6199%	L	3.1500%
n	3.2013%	P	3.0300%
q	2.8270%	D	2.7700%
p	0.8497%	B	1.5800%
m	0.5818%	Q	1.5100%
f	0.2633%	G	1.2100%
*	0.1460%	V	0.9600%
g	0.0500%	F	0.9300%
x	0.0182%	H	0.6900%
v	0.0047%	X	0.6000%
z	0.0010%	Y	0.0700%
S	0.0005%	Z	0.0100%
b	0.0000%	*	0.0000%
j	0.0000%	J	0.0000%
u	0.0000%	K	0.0000%
w	0.0000%	W	0.0000%

Figure 10

According to the Figure 10, it is obvious that the commonly used letters in Latin are all the capitals. Because the Takahashi edition is a transcript from the Voynich manuscript, which means the letter 'o' in the Takahashi edition may does not mean 'o', it just looks like 'o' in the Voynich manuscript. So in order to get the results, the correlation between the Voynich and Latin is calculated, the result is 98.60%, which means the 'o' in the Takahashi edition may stand for 'I' in Latin. In the same way, there are potential relationships between 'e' and 'E', 'h' and 'A', 'y' and 'U'.

## Part 2: The Voynich versus English.

The occurrence frequencies of letters in the Voynich manuscript and English are shown in the Figure 11 and Figure 12.

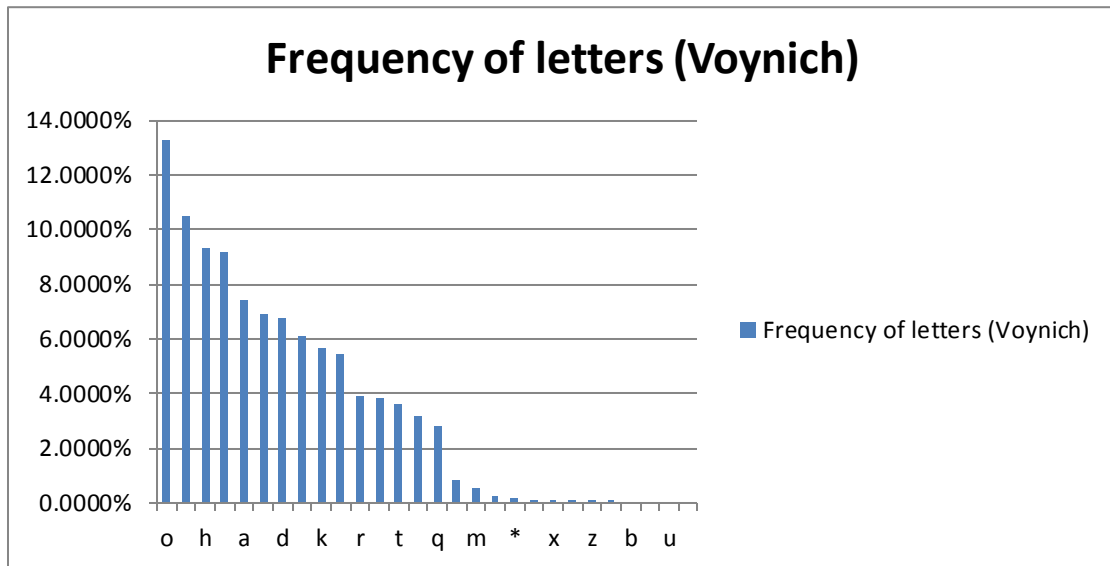


Figure 11

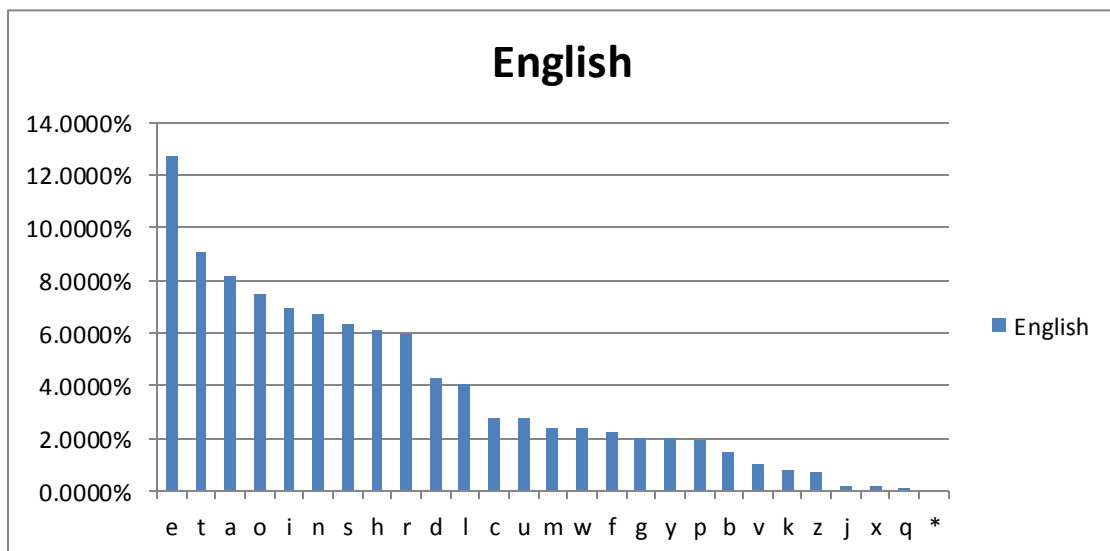


Figure 12

The form of proportion is shown in the Figure 13.



Characters	Frequency of letters (Voynich)	Characters	English
o	13.2766%	e	12.7020%
e	10.4626%	t	9.0560%
h	9.3084%	a	8.1670%
y	9.2037%	o	7.5070%
a	7.4448%	i	6.9660%
c	6.9407%	n	6.7490%
d	6.7629%	s	6.3270%
i	6.1160%	h	6.0940%
k	5.7000%	r	5.9870%
l	5.4831%	d	4.2530%
r	3.8869%	l	4.0250%
s	3.8509%	c	2.7820%
t	3.6199%	u	2.7580%
n	3.2013%	m	2.4060%
q	2.8270%	w	2.3610%
p	0.8497%	f	2.2280%
m	0.5818%	g	2.0150%
f	0.2633%	y	1.9740%
*	0.1460%	p	1.9290%
g	0.0500%	b	1.4920%
x	0.0182%	v	0.9780%
v	0.0047%	k	0.7720%
z	0.0010%	z	0.7400%
S	0.0005%	j	0.1530%
b	0.0000%	x	0.1500%
j	0.0000%	q	0.0950%
u	0.0000%	*	0.0000%
w	0.0000%	S	N/A

Figure 13

According to the Figure 13, the correlation between the Voynich and Latin is calculated, the result is 97.76%.

In order to search the exact correlation between the Voynich and English, the next step is to compare the Voynich with other books which were written in English and the results are shown in the part 5.

### Part 3: The Voynich versus French.

The occurrence frequencies of letters in the Voynich manuscript and French are shown in the Figure 14 and Figure 15.

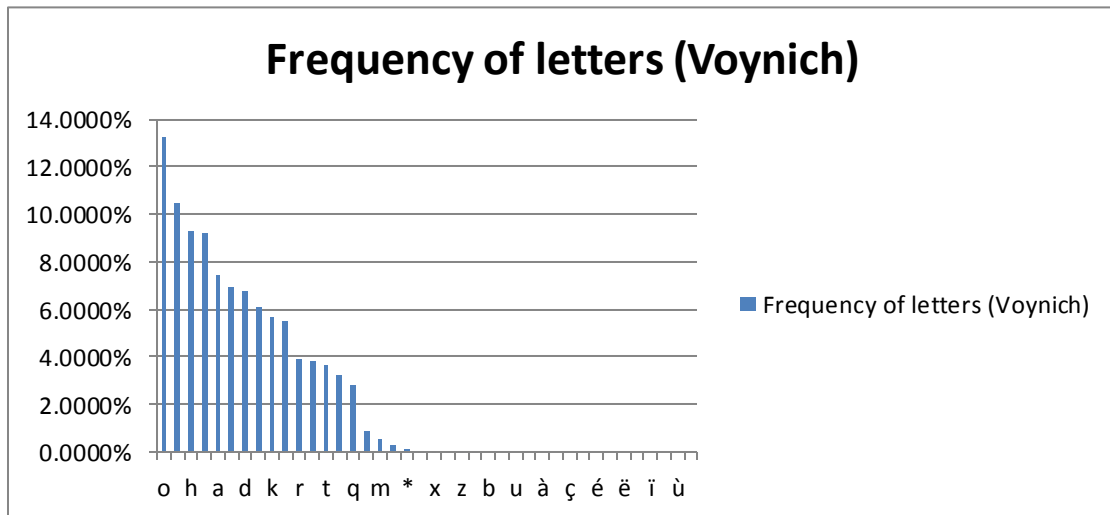


Figure 14

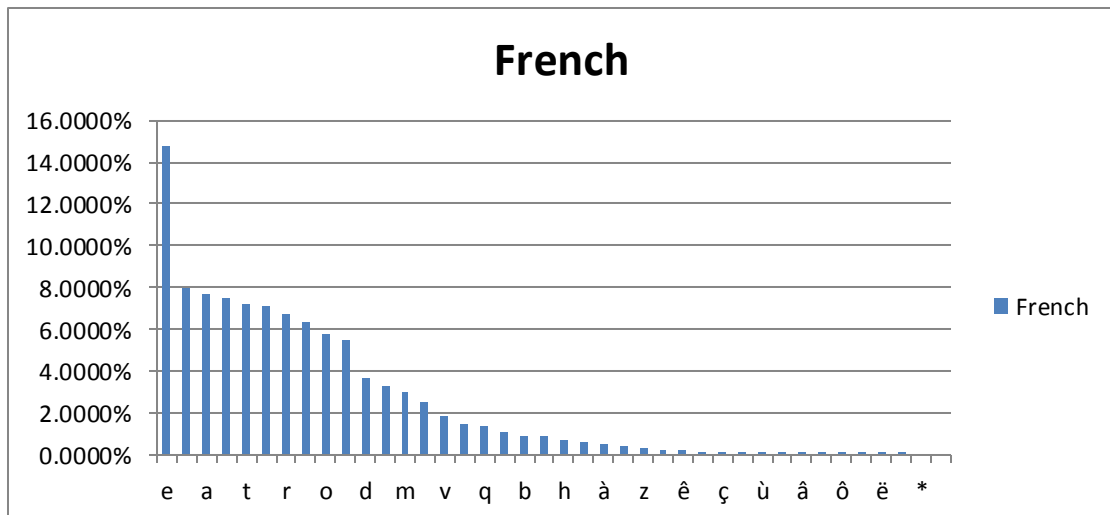


Figure 15

The form of proportion is shown in the Figure 16.

Characters	Frequency of letters (Voynich)	Characters	French
o	13.2766%	e	14.7200%
e	10.4626%	s	7.9500%
h	9.3084%	a	7.6400%
y	9.2037%	i	7.5300%
a	7.4448%	t	7.2400%
c	6.9407%	n	7.1000%
d	6.7629%	r	6.6900%
i	6.1160%	u	6.3100%
k	5.7000%	o	5.8000%
l	5.4831%	l	5.4600%
r	3.8869%	d	3.6700%
s	3.8509%	c	3.2600%
t	3.6199%	m	2.9700%
n	3.2013%	p	2.5200%
q	2.8270%	v	1.8400%
p	0.8497%	é	1.5000%
m	0.5818%	q	1.3600%
f	0.2633%	f	1.0700%
*	0.1460%	b	0.9000%
g	0.0500%	g	0.8700%
x	0.0182%	h	0.7400%
v	0.0047%	j	0.6100%
z	0.0010%	à	0.4900%
S	0.0005%	x	0.4300%
b	0.0000%	z	0.3300%
j	0.0000%	è	0.2700%
u	0.0000%	ê	0.2200%
w	0.0000%	y	0.1300%
à	0.0000%	ç	0.0900%
â	0.0000%	w	0.0700%
ç	0.0000%	ù	0.0600%
è	0.0000%	k	0.0500%
é	0.0000%	â	0.0500%
ê	0.0000%	î	0.0500%
ë	0.0000%	ó	0.0200%
î	0.0000%	œ	0.0200%
ī	0.0000%	ë	0.0100%
ó	0.0000%	ĩ	0.0100%
ù	0.0000%	*	0.0000%
œ	0.0000%	S	0.0000%

Figure 16

According to the Figure 16, the correlation between the Voynich and Latin is 98.11%.

Though there are some similarities between the Voynich and French as the analysis above, there are much more differences. For example, as shown in the Figure 16, there are 38 letters in total in French, only 24 letters in the Voynich manuscript. Therefore, there are many differences between the Voynich and French.

**Part 4: The Voynich versus German.**

The occurrence frequencies of letters in the Voynich manuscript and German are shown in the Figure 17 and Figure 18.

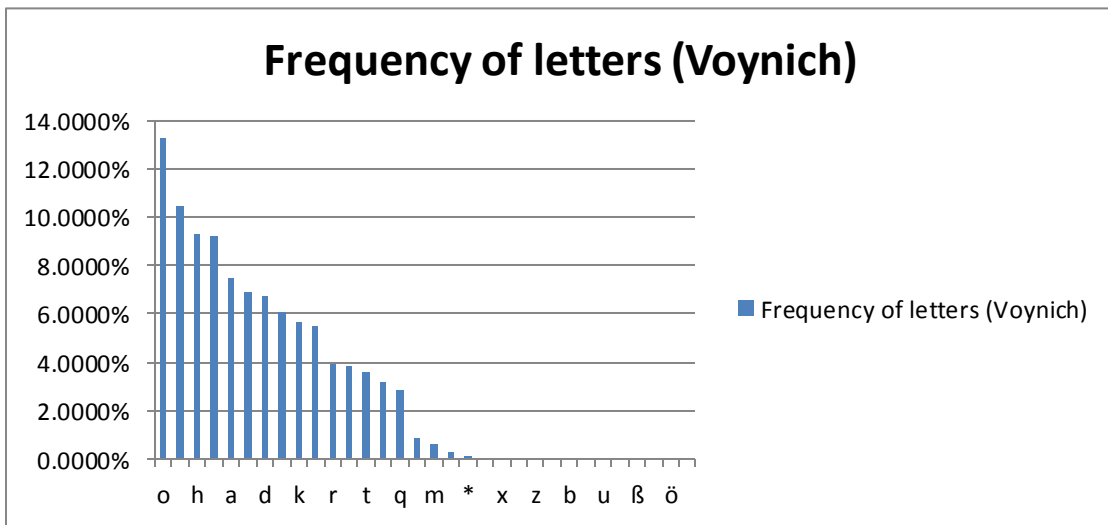


Figure 17

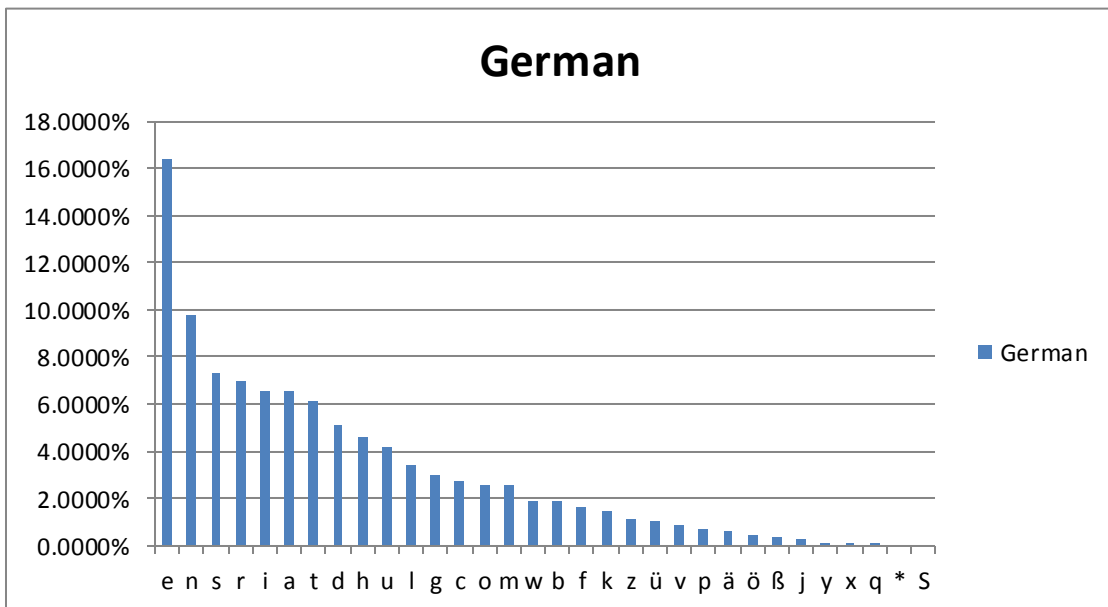


Figure 18

The form of proportion is shown in the Figure 19.

Characters	Frequency of letters (Voynich)	Characters	German
o	13.2766%	e	16.4000%
e	10.4626%	n	9.7800%
h	9.3084%	s	7.2700%
y	9.2037%	r	7.0000%
a	7.4448%	i	6.5500%
c	6.9407%	a	6.5200%
d	6.7629%	t	6.1500%
i	6.1160%	d	5.0800%
k	5.7000%	h	4.5800%
l	5.4831%	u	4.1700%
r	3.8869%	l	3.4400%
s	3.8509%	g	3.0100%
t	3.6199%	c	2.7300%
n	3.2013%	o	2.5900%
q	2.8270%	m	2.5300%
p	0.8497%	w	1.9200%
m	0.5818%	b	1.8900%
f	0.2633%	f	1.6600%
*	0.1460%	k	1.4200%
g	0.0500%	z	1.1300%
x	0.0182%	ü	1.0000%
v	0.0047%	v	0.8500%
z	0.0010%	p	0.6700%
S	0.0005%	ä	0.5800%
b	0.0000%	ö	0.4400%
j	0.0000%	ß	0.3100%
u	0.0000%	ÿ	0.2700%
w	0.0000%	y	0.0400%
ß	0.0000%	x	0.0300%
ä	0.0000%	q	0.0200%
ö	0.0000%	*	0.0000%
ÿ	0.0000%	S	0.0000%

Figure 19

According to the Figure 19, the correlation between the Voynich and Latin is 95.86%.

Though there are some similarities between the Voynich and German as the analysis above, there are some differences. For example, as shown in the Figure 19, there are 30 letters in total in German, only 24 letters in the Voynich manuscript.

In order to search the exact correlation between the Voynich and German, the next step is to compare the Voynich with other books which were written in German and the result is shown in the part 5.

### Part 5: Comparing the Voynich with other books which are written in the known languages.

In order to ensure the accuracy of the results, project team search some literary classics which were written by English, French and German and compared the Voynich manuscript with those books. In order to compare them conveniently, project team extract the same number of words from every book. The results are shown in the Figure 20, Figure 21 and Figure 22.

Book name	Total characters number	Total words number (Twn)	Unique words number (Uwn)
Voynich	234507	37104	8486
Pride and prejudice	210748	36433	3706
Sherlock	229037	37095	5397
The three presents	211675	36874	4073
ALPHONSE DE	219073	37467	6366
magog_les_buveurs_d_ocean_source	235942	37457	6859
karr_sous_les_tilleuls_source	213019	37484	5371
Faust	195835	30654	6079
FAUST: EINE TRAG	229801	37082	7242
Carlos Ruiz Zafón	238584	37270	7423

Figure 20

Book name	Ratio(Uwn/Twn)	characters per word	Language
Voynich	0.229	6.32	
Pride and prejudice	0.102	5.78	<b>English</b>
Sherlock	0.145	6.17	<b>English</b>
The three presents	0.110	5.74	<b>English</b>
ALPHONSE DE	0.170	5.85	<b>French</b>
magog_les_buveurs_d_ocean_source	0.183	6.30	<b>French</b>
karr_sous_les_tilleuls_source	0.143	5.68	<b>French</b>
Faust	0.198	6.39	<b>German</b>
FAUST: EINE TRAG	0.195	6.20	<b>German</b>
Carlos Ruiz Zafón	0.199	6.40	<b>German</b>

Figure 21

Book name	number of words that appear once	Ratio (words appear once/Uwn)
Voynich	6012	70.85%
Pride and prejudice	1747	47.14%
Sherlock	2913	53.97%
The three presents	2006	49.25%
ALPHONSE DE	3833	60.21%
magog_les_buveurs_d_ocean_sourc	4060	59.19%
karr_sous_les_tilleuls_source	3002	55.89%
Faust	3722	61.23%
FAUST: EINE TRAG	4415	60.96%
Carlos Ruiz Zafón	4797	64.62%

Figure 22

In order to compare the Voynich manuscript with those books conveniently, line charts which can show the potential relationship between the Voynich manuscript and these books are made. The results are shown in the Figure 23, Figure 24 and Figure 25.

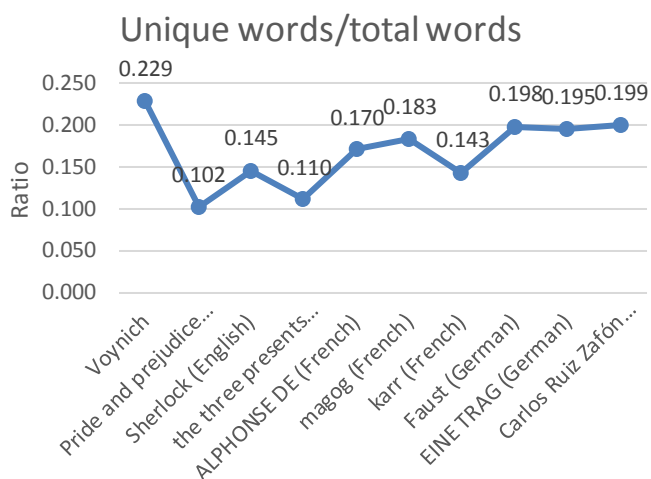


Figure 23

Figure 23 shows the percentage of unique words/total words. There is significant difference between the Voynich manuscript and English books (47.9%) or French books (27.7%). However, there is no significant difference between the Voynich manuscript and German (13.6%).

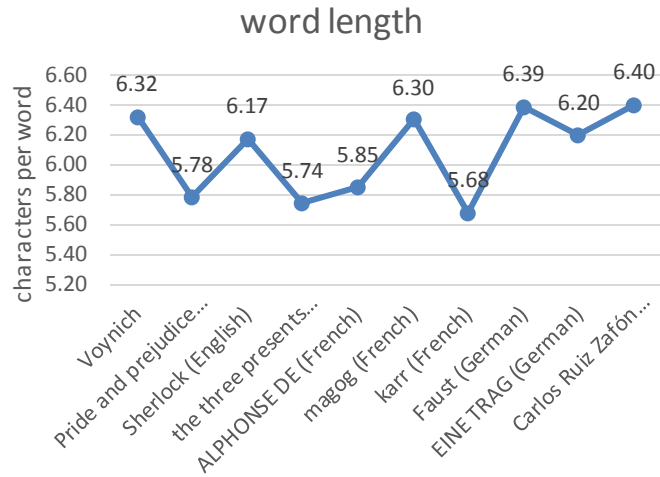


Figure 24

Figure 24 shows the word length the Voynich, English, French and German. There is small difference for the word length between the Voynich manuscript and English (6.7%) or French (6.0%). Furthermore, there is no significant difference for the word length between the Voynich manuscript and German (0.1%).

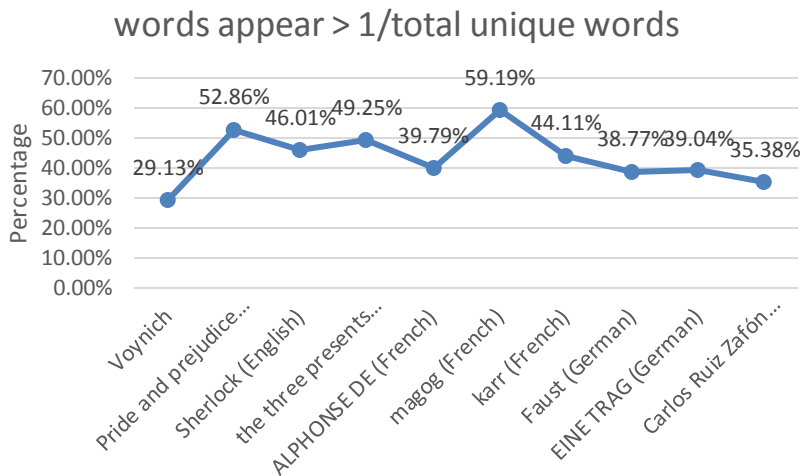


Figure 25

Figure 25 shows the percentage of words appear more than once /total unique words were compared. There is large difference between the Voynich manuscript and English (41.0%) or French (38.9%) or German (22.8%). However, the difference between the Voynich manuscript and German books is the smallest difference among these differences.



In addition, as the analysis in the part 4 section 6.1.2.3, the correlation between the Voynich manuscript and German is high (95.86%). So maybe the language which was used in the Voynich manuscript is a branch of German. As the result, there are potential relationships between the Voynich and German.

### 6.1.3. Digits

According to the introduction in the section 1.1, the Voynich manuscript was found in 1912. During the period of 17 Century to 18 Century, the most commonly used method of expressing digits is using Roman [14]. The method of expressing digits in Roman is shown in the Appendix section A.3. In addition, the method is introduced in the section 4.1.

The results of searching the characters with the form ‘\*##’ in the Voynich manuscript are shown in the Figure 26.

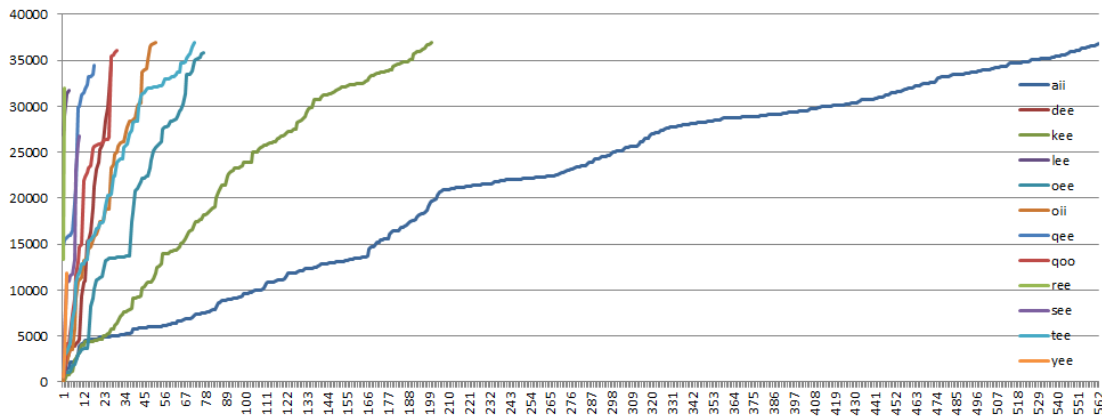


Figure 26

As shown in the Figure 26, the words with the form ‘\*##’ in the Voynich manuscript involve: ‘aai’, ‘dee’, ‘kee’, ‘lee’, ‘oee’, ‘oii’, ‘qee’, ‘qoo’, ‘ree’, ‘see’, ‘tee’ and ‘yee’. X axis means the occurrence number of each word. As shown in the Figure 26, the most commonly used word is ‘aai’ and the occurrence number of ‘aai’ is 563. In addition, the occurrence number of ‘ree’ is the smallest, which are 2. Y axis means the positions of each word. For example, the position of 563<sup>rd</sup> ‘aai’ is 36821, which means this word is the 36821<sup>st</sup> word in the Voynich manuscript.

As the analysis above, ‘aai’ may stand for ‘seven’ in Roman (VII).

Then, the words with the form ‘\*###’ are extracted by using the same method. The results are shown in the Figure 27.

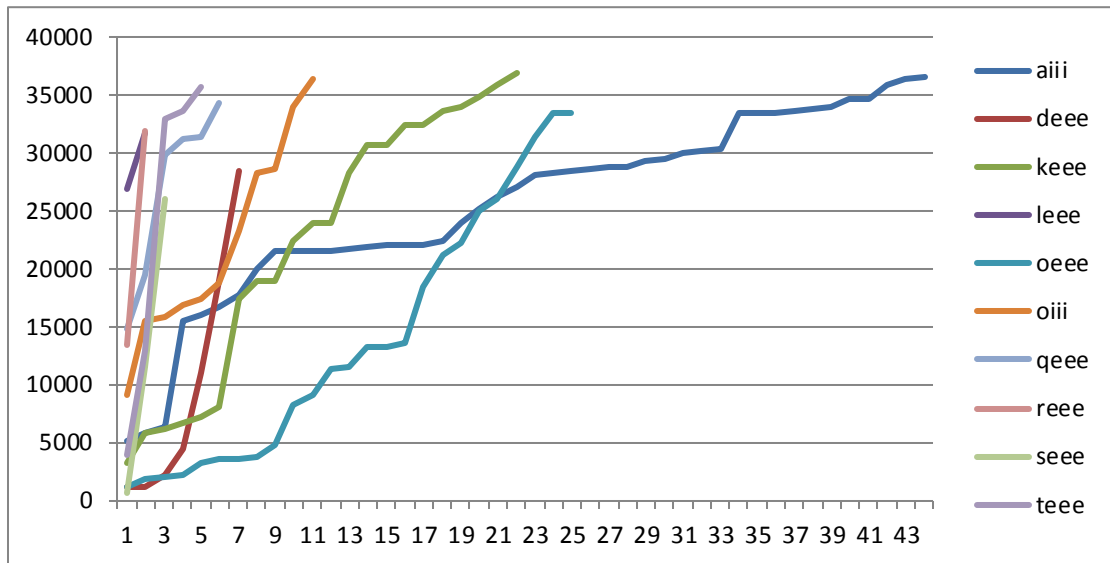


Figure 27

From the Figure 27, it is obvious that the occurrence frequency of ‘aiii’ is the highest, so maybe ‘aiii’ means ‘eight’ in Roman (VIII). The exact data is shown in the Appendix section A.7.

Then, these triple words are compared with other triple words which come from the known languages: English, German and Russian. The results are shown in the Table 6, Table 7 and Table 8.

Table 6: English

Word	Part of speech	Letters
sss	noun (onomatopoeia)	3
zzz	noun (onomatopoeia)	3
zzzs	noun	4
ohmmm	noun (onomatopoeia)	5
Aaaaba	proper noun	6
illit	adjective	6
gillless	adjective	8
wallless	adjective	8
bulllike	adjective, adverb	8
hilllike	adjective	8
Aaadonta	proper noun	8

willless	adjective	8
shellless	adjective	9
skillless	adjective	9
skulllike	adjective	9
Amerikkka	proper noun	9
Amerikkkan	proper noun	10
goddessship	noun	11
hostessship	noun	11
willlessness	noun	12
headmistressship	noun	16

Table 7: German

Words	Meaning
Schneeeule	snow
Teeei	tea

Table 8: Russian

Words	Meaning
длинношеее	having a long neck
змееед	a name of a bird, 'snake-eater'.
дооооновский	'pre-UN'
зоообъединение	'zoos' association

According to the tables above, it is obvious that triple 'l' and triple 's' are the most commonly used in English. In addition, triple 'e' and triple 'i' are the most commonly used in the Voynich manuscript. As the result, it can be inferred that there are potential relationships among 'l', 's' in English and 'e' and 'i' in the Voynich.

In addition, it is obvious that triple 'e' is the most commonly used in German from the Table 7 above. Compare with The Voynich, it is obvious that there are potential relationship among 'e' in German and 'e', 'i' in the Voynich.

Moreover, triple ‘o’ is the most commonly used in Russian. As the result, it can be inferred that there are relationships among ‘o’ in Russian and ‘e’, ‘i’ in the Voynich.

### 6.2 Phase 2: Illustration investigation

This phase includes three parts: statistics for illustration of each page, digits mining and conclusion. According to the section 5.4 task allocation, this part is completed by Yaxin Hu.

#### 6.2.1 Searching initial numbers and possible numerical words inside images

The first part of this section is to find all initial numbers inside the images of the whole Voynich manuscript. There is a list of some part of the initial numbers below:

Table 9. A part of the initial numbers

page	f1r	f1v	f2r	f2v	f3r	f3v	f4r	f4v	f5r	f5v	f6r	f6v	f7r	f7v
initial number	14	1	3	1	13	2	1	1	1	4	3	1	2	2
		12	5	4	27	6	3	6	4	5	4	5	4	3
		14	6				5	7	9	6	8	6	8	7
		26					13	8			16	7		18
								9				35		
								18						
								29						

In order to make a comparison and mapping between initial numbers and the Voynich manuscript, all possible words that may stand for numbers. There is a list of some part of the possible words below:

Table 10. A part of the possible words

page	f1r	f1v	f2r	f2v	f3r	f3v	f4r	f4v	f5r	f5v	f6r	f6v	f7r	f7v
possible words	o	ol	s	o	s	r	ol	s	s	qo	s	s	NA	ty
	s	or			ol	s		ol	or	sy	y	y		ol

	y	qo			or	or					od	os		
	ol										ol	or		
	or										or			
											os			

**6.2.2 Mapping all initial numbers and numerical words**

When we compare the initial numbers and possible words, there can be seen some potential relationship between them, such as there are a lot of ‘s’ and ‘2’ appear in the same page (54 pairs), ‘o’ and ‘1’ for 24 pairs, ‘ol’ and ‘10’ for 14 pairs. Therefore, in order to make it simple to compare, mapping between initial numbers and possible words are made to show whether there is any relationship between them. There is a list of a part of the mapping pairs below, the other parts are shown in the section Appendix A.8.

Table 11. Mapping pairs for letter m

letter	number	frequency
m	1	5
	2	5
	3	3
	8	3
	4	2
	5	2
	9	2
	6	1
	7	1

Table 12. Mapping pairs for letter n

letter	number	frequency
n	2	2
	4	2
	1	1
	3	1
	5	1
	6	1
	8	1
	9	1
	7	0

Table 13. Mapping pairs for letter o

letter	number	frequency
o	1	24
	2	20
	3	16
	4	14
	5	12
	6	8
	9	8
	8	6
	7	4

word	number	frequency
ol	10	14
	13	12
	12	11
	11	9
	14	7
	15	7
	20	6
	18	5
	29	5
	24	4
	19	3
	16	2
	17	2
	22	2
	27	2
	21	1
	25	1
	26	1
	32	1
	33	1

	35	1
	37	1
	39	1
	40	1
	46	1

word	number	frequency
or	10	19
	12	13
	13	12
	14	6
	15	5
	19	5
	20	5
	11	4

Table 14. Mapping pairs for letter r

letter	number	frequency
r	1	48
	2	26
	3	21
	5	19



	4	14
	6	9
	8	6
	7	5
	9	4

Table 15. Mapping pairs for letter s

letter	number	frequency
s	2	54
	1	46
	3	41
	5	32
	4	26
	6	22
	8	19
	9	15
	7	10

Table 16. Mapping pairs for letter y

letter	number	frequency
y	2	36
	1	30

	3	29
	5	20
	4	15
	7	13
	8	12
	6	11
	9	7

In order to make it simple to find a more possible relationship among them, we choose the most frequency pairs for each pair, and made a new list, which is shown below, other parts are shown in the section Appendix A.9.

Table 17. The most frequency pairs for each pair

letter	number	frequency
o	1	24
	2	20
	3	16
	4	14
	5	12
ol	10	14
	13	12
	12	11
or	10	19
	12	13

	13	12
os	19	11
	12	4
	13	4
r	1	48
	2	26
	3	21
	5	19
	4	14
s	2	54
	1	46
	3	41
	5	32
y	2	36
	1	30
	3	29
	5	20

There can be easily seen that ‘o’ and ‘1’ appears together for 24 times. Furthermore, there are a lot of ‘ol’ and ‘10’ (14 times), ‘ol’ and ‘13’ (12 times), ‘ol’ and ‘12’ (11 times), ‘or’ and ‘10’ (19 times), ‘or’ and ‘12’ (13 times), ‘or’ and ‘13’ (12 times), ‘os’ and ‘19’ (11 times) appear together. Therefore, there is a potential relationship between ‘o’ in the Voynich manuscript and number ‘1’.

Furthermore, there are ‘r’ and ‘1’ for 48 times, ‘r’ and ‘2’ for 26 times, ‘r’ and ‘3’ for 21 times, ‘s’ and ‘2’ appear together for 54 times, ‘s’ and ‘1’ for 46 times, ‘s’ and ‘3’ for 41 times, ‘s’ and ‘5’ for 32 times, ‘y’ and ‘2’ for 36 times, ‘y’ and ‘1’ for 30 times,

'y' and '3' for 29 times, 'y' and '5' for 20 times. There may exist potential relationship among them, which need further investigation.

In order to make it simple to see and compare, there is a list that all the possible pairs shown below:

Table 18. Possible relationship for letters in the Voynich manuscript and numbers

possible relationship	
letter in the Voynich manuscript	number
o	1
ol	10
	13
	12
or	10
	12
	13
os	19
r	1
	2
s	2
	1
	3
y	2
	1
	3

### **6.3 Phase 3: Marginal symbol research**

According to the section 4.3, this phase is divided into three parts: statistics for marginal stars of each page, digits mining and conclusion. According to the section 5.4 task allocation, this phase is completed by Ruihang Feng.

#### **6.3.1 Statistics for marginal stars of each page**

There are 15 pages which involve marginal stars in the Voynich manuscript. As the analysis in the section 4.3, an example is shown in the Appendix section A.6. The results of this part are shown in the Appendix section A.10.

From the A.10, we can find that there are two kinds of marginal stars in the Voynich manuscript: white stars and coloured stars. In addition, A.10 also involves detailed information about the number of stars, arrangement and location in the text.

#### **6.3.2 Digits analysis**

In this phase, first, the number of marginal stars for each page is counted. Then, letters which may stand for digits are extracted. An example (page number: f58r) is shown in the Figure 28.

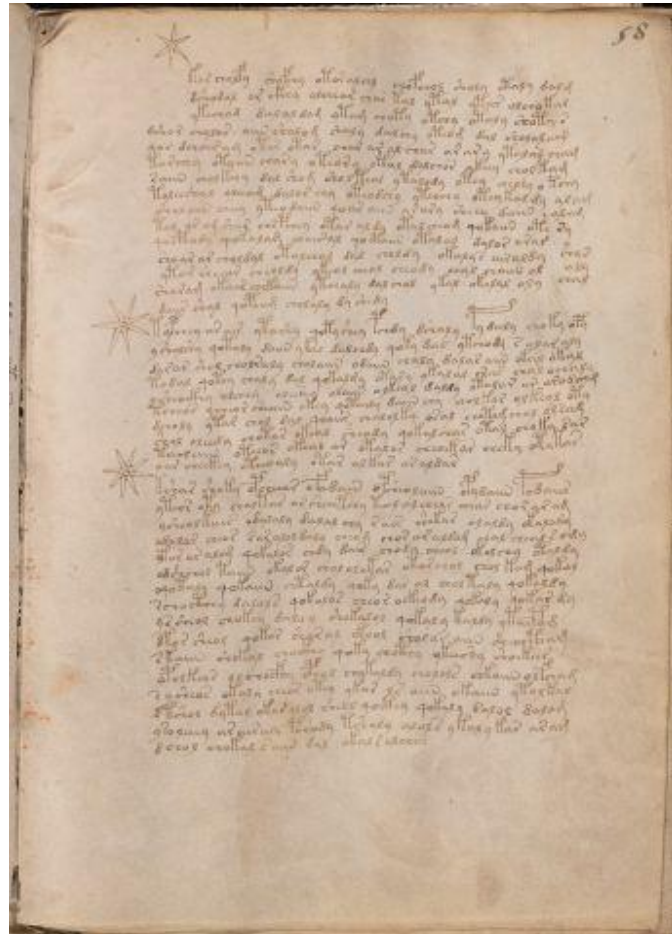


Figure 28

For this page, there are 3 white stars (according the Appendix section A.10) and the single letters which may stand for digits are m, o, r and s. Then all the 25 pages are counted in this way.

As the result, these 25 pages involve 16 kinds of digits: 1 3 4 5 6 7 8 9 10 12 13 14 15 16 17 and 19. Some of them stand for the total number stars of each page; some of them stand for the number of white stars or the number of the coloured stars of each page. The detailed information is shown in the Appendix section A.10.

The results of this phase are shown in the Appendix section A.11. The first column stand for those 16 kinds of digits, the information in brackets mean the number of the pages which involve that digit (for example, for the digit '5', the information in brackets is 3 pages, that means there are 3 pages which involve '5'); the red mark represent the top several letters which has high occurrence frequency; the second column stand for the pages which involve the digits and the last column means the letters which may stand for digits.

### 6.3.3 Conclusion

According to the section 6.3.1 and 6.3.2, the conclusion of the phase ‘marginal symbol research’ is shown in the Table 19.

Table 19: Conclusion

letter	Pages (25)	The digit which letter may stand for	frequency
y	18	5	16.67%
		6	38.89%
		7	55.56%
		8	27.78%
		9	11.11%
l	13	5	38.46%
		7	53.85%
		8	30.77%
		9	15.38%
r	7	5	28.57%
s	10	6	40%
		7	30%
o	8	6	50%
		8	25%
		9	25%
ar	22	10	18.18%
		12	4.46%
		13	27.27%

		14	9.1%
		15	13.64%
		16	9.1%
		19	9.1%
al	21	10	19.05%
		13	28.57%
		14	9.52%
		15	19.05%
		16	9.52%
or	17	10	17.65%
		13	29.41%
		16	11.76%
ol	21	10	14.29%
		12	14.29%
		13	23.81%
		14	9.52%
		15	19.05%
		16	9.52%
am	12	12	16.67%
		19	16.67%
dy	6	14	33.33%
		19	33.33%



---

om	5	16	40%
----	---	----	-----

The letters of the first column are extracted according to the red mark in the Appendix section A.11. The forth column stand for the occurrence frequency of letters. For example, the occurrence frequency of  $y=5$  is equal to  $3/18=16.67\%$ , '3' means there are 3 pages which involve 'y=5' (according to the Appendix section A.10.), '18' means there are 18 pages which involve 'y'.

As the result, according to the figures above, we can find that there are the most possible potential relationships between:

- 'y' and '7'
- 'l' and '7'
- 'r' and '5'
- 's' and '6'
- 'o' and '1'
- 'o' and '6'
- 'ar' and '13'
- 'al' and '13'
- 'or' and '13'
- 'ol' and '13'
- 'am' and 12
- 'am' and '19'
- 'dy' and '14'
- 'dy and '19'
- 'om' and '16'

## **7. Comment on progress**

In the past two semesters, the processes of this project are normal. Though we met many problems in the course of the project, such as references limitation and Matlab code error, we adjusted and modified our origin plan in time. As the result, the whole project schedule is not affected too much.

In general, we finished this project on time and reached the expected goal.

## 8. Conclusion

This project is divided into three phases: text investigation, illustration research and marginal symbol investigation. On the other hand, the major works of this project can be achieved by using computer.

In addition, the goals of this project involve three parts:

- Use statistical method Matlab to search the linguistic laws in the Voynich manuscript.
- Search laws from illustrations from the perspective of digits.
- Investigate laws from marginal symbols from the perspective of digits.

Over the past two semesters, the whole phases have been finished. As the analysis in the section 6.1, we can infer that the language which is used in the Voynich manuscript may be a branch of German.

In addition, we can get the results of the digits analysis from combining the section 6.2 and 6.3:

The possible relationship	
Characters	Digits
s	2 and 6
y	2 and 6
a	1
r	1
The most possible relationship	
o	1
ol	13
or	13

## 9. References

[1] R. Zandbergen (2016). *The Voynich MS-Introduction* [Online]. Available: <http://www.voynich.nu/intro.html>

[2] Kevin Knight, Sravana Reddy, *What We Know About The Voynich Manuscript* [Online]. Available: <http://www.isi.edu/natural-language/people/voynich-11.pdf>

[3] Stojko, John, *Letters to God's Eye: The Voynich Manuscript for the first time deciphered and translated into English*. New York: Vantage Press, 1978.

[4] Joachim Dathe, *The EVA-Transcription* [Online].

Available: <https://voynich2arabic.wordpress.com/eva-transcription/>

[5] Vladimir Sazonov, *Voynich Manuscript* [Online].

Available: <http://voynich.naobum.de/>

[6] Reed Johnson (2013, July 9), *The Unread: The Mystery Of The Voynich Manuscript* [Online].

Available:

<http://www.newyorker.com/books/page-turner/the-unread-the-mystery-of-the-voynich-manuscript>

[7] R. Zandbergen (2016), *History of research of the Voynich MS* [Online]. Available: <http://www.voynich.nu/solvers.html#n01>

[8] Mary E. D'Imperio, *An Application of Cluster Analysis and Multiple Scaling to the Question of "Hands" and "Languages" in the Voynich Manuscript*. Washington, DC, 1992.

[9] Pelling, Nicholas, *The curse of the Voynich; the secret history of the world's most mysterious manuscript*, Compelling Press, Surbiton, 2006.

[10] Bennett, William Ralph, *Scientific and Engineering Problem Solving with the Computer*. Englewood Cliffs: Prentice-Hall, 1976.

[11] Tiltman, John, "The Voynich Manuscript, The Most Mysterious Manuscript in the World". NSA Technical Journal 12 (July 1967), pp.41-85.

[12] Feely, Joseph M, *Roger Bacon's Cipher: The Right Key Found*, Rochester, 1943.

[13] D'Imperio, Mary E, *The Voynich Manuscript - an elegant enigma*, Aegean Park Press, 1978.

[14] R. Zandbergen (2016), *History of research of the Voynich MS* [Online]. Available: <http://www.voynich.nu/solvers.html#n43>

[15] Wikipidia, *Medieval Literature* [Online].

Available: [https://en.wikipedia.org/wiki/Medieval\\_literature#Languages](https://en.wikipedia.org/wiki/Medieval_literature#Languages)

[16] Wikipidia, *Letter frequency* [Online].

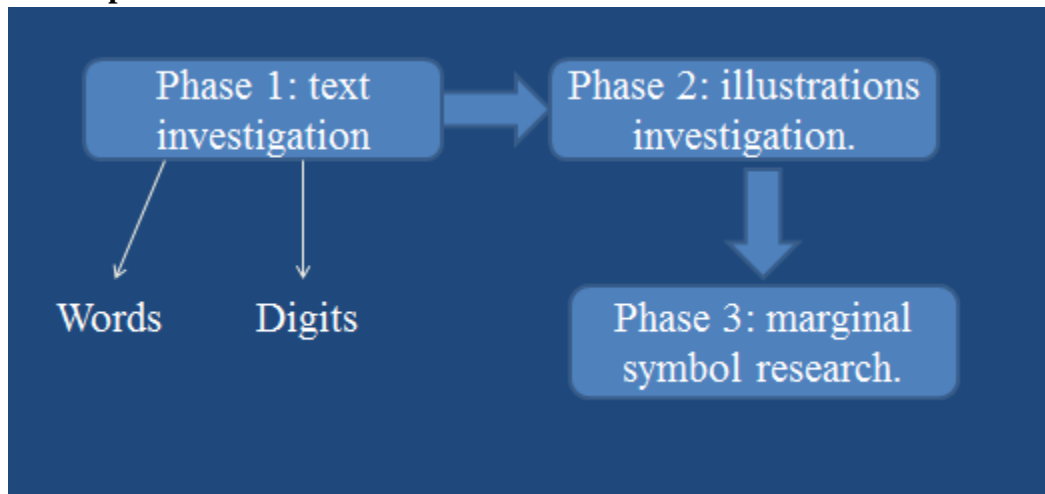
Available: [https://en.wikipedia.org/wiki/Letter\\_frequency](https://en.wikipedia.org/wiki/Letter_frequency)

## 10. Appendix

### A.1. FSG

Char	Bennett	FSG	Currier		Char	Bennett	FSG	Currier
4	D	4	4		∩	I	I	I
o	O	O	O		∩	IL	IE	G
8	S	8	8		∩∩	IIL	IIE	H
9	G	G	9		∩∩∩	IIIL	IIIE	1
2	Z	2	2		∩∩	IQ	IR	T
∩	L	E	E		∩∩∩	IIQ	IIR	U
∩	Q	R	R		∩∩∩	IIIQ	IIIR	0
CT	CT	T	S		∩	U	L	D
ET	ET	S	Z		∩	N	N(*)	N
HP	H	H	P		∩∩	M	M(*)	M

### A.2. Proposed Method



### A.3. Roman Numeral

Number	Roman Numeral
1	I
2	II
3	III
4	IV
5	V
6	VI
7	VII
8	VIII
9	IX
10	X

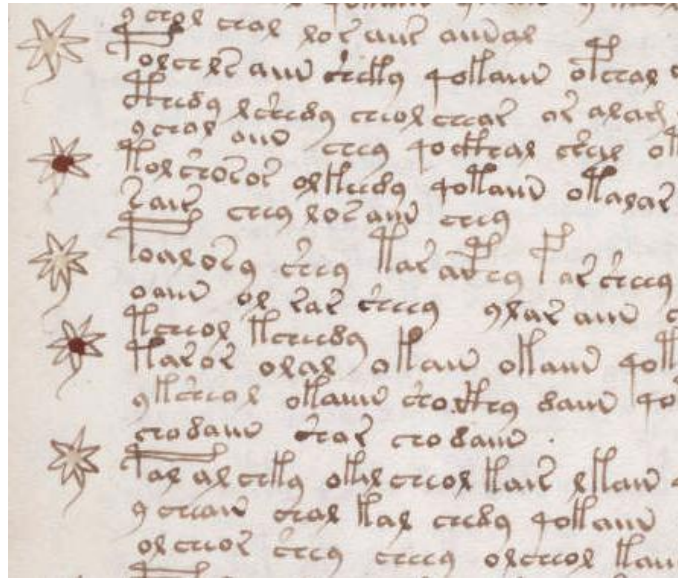
**A.4. Words from the Voynich manuscript.**



**A.5. Illustration from the Voynich manuscript.**



**A.6. Marginal symbols from the manuscript.**



g ceo ceo roe and andae  
foceed and ecclg qolland otece  
oecdg lobece qool cece ad aece  
oecce and cece qolland ecce o  
focecece oeccece qolland otece  
cece cece cece and cece  
locecece cece cece cece cece  
oand cece cece cece cece and cece  
cecece cecece cecece cecece qolland  
cecece cecece cecece cecece cecece  
cecece cecece cecece cecece  
cecece cecece cecece cecece cecece  
cecece cecece cecece cecece cecece  
cecece cecece cecece cecece cecece



### A.7. Digits ‘\*###’

aiii	deee	keee	leee	oeee	oiii	qeee	reee	seee	teee	yeee
5137	1087	3146	26847	1064	9108	14777	13401	555	3969	568
5795	1138	5804	31760	1812	15520	19498	31974	11561	12899	
6266	2242	6170		1984	15775	29882		25972	32933	
15471	4408	6625		2241	16794	31171			33658	
16034	10975	7130		3137	17417	31476			35650	
16785	18941	7985		3565	18834	34419				
17706	28473	17468		3616	23251					
19970		18993		3704	28329					
21518		19013		4715	28698					
21591		22443		8306	34001					
21605		23923		9127	36436					
21616		23965		11275						
21778		28217		11441						
21884		30737		13245						
22007		30757		13291						
22030		32374		13625						
22032		32418		18465						
22400		33670		21269						
23963		33973		22163						
25099		34870		25085						
26180		35895		26113						
27114		36929		28856						
28102				31346						
28302				33468						
28466				33490						
28710										
28746										
28835										
29267										
29554										
30070										
30134										
30375										
33461										
33476										
33498										
33594										
33848										
34045										
34670										
34695										
35925										
36443										
36523										

### A.8. Mapping list for letters and numbers

letter	number	frequency
m	1	5
	2	5
	3	3
	8	3
	4	2
	5	2
	9	2
	6	1
7	1	

letter	number	frequency
n	2	2
	4	2
	1	1
	3	1
	5	1
	6	1
	8	1
	9	1
	7	0

letter	number	frequency
o	1	24
	2	20
	3	16
	4	14
	5	12
	6	8
	9	8
	8	6
	7	4

word	number	frequency
oc	10	1

word	number	frequency
od	12	1
	14	1
	16	1

word	number	frequency
ok	24	2
	10	1
	13	1
	20	1

word	number	frequency
ol	10	14
	13	12
	12	11
	11	9
	14	7
	15	7

20	6
18	5
29	5
24	4
19	3
16	2
17	2
22	2
27	2
21	1
25	1
26	1
32	1
33	1
35	1
37	1
39	1
40	1
46	1

word	number	frequency
om	12	3
	10	2
	13	2
	11	1
	14	1
	17	1
	24	1
	29	1
	40	1

word	number	frequency
or	10	19
	12	13
	13	12
	14	6
	15	5
	19	5
	20	5
	11	4
	24	4
	29	4
	16	3

	18	3
	21	3
	17	2
	22	2
	25	1
	26	1
	27	1
	32	1
	33	1
	37	1
	40	1
	46	1

word	number	frequency
os	19	11
	12	4
	13	4
	10	3
	11	2
	18	2
	16	1
	17	1
	20	1
	25	1
	30	1
	35	1
	37	1
40	1	

word	number	frequency
ot	12	4
	10	1
	13	1
	16	1
	17	1
	40	1

word	number	frequency
oy	11	1
	14	1

letter	number	frequency
p	1	1

	3	1
	6	1
	2	0
	4	0
	5	0
	7	0
	8	0
	9	0

letter	number	frequency
q	1	1
	2	1
	5	1
	6	1
	8	1
	9	1
	3	0
	4	0
	7	0

word	number	frequency
qo	14	4
	10	2
	11	2
	13	2
	12	1
	15	1
	26	1
	39	1

word	number	frequency
qy	11	1
	13	1
	14	1

letter	number	frequency
r	1	48
	2	26
	3	21
	5	19
	4	14
	6	9
	8	6

	7	5
	9	4

word	number	frequency
ra	13	1
	16	1

word	number	frequency
rk	18	1

word	number	frequency
ro	11	1
	13	1
	14	1

word	number	frequency
ry	15	2
	11	1
	13	1
	14	1
	19	1
	26	1

letter	number	frequency
s	2	54
	1	46
	3	41
	5	32
	4	26
	6	22
	8	19
	9	15
	7	10

word	number	frequency
sh	13	2
	10	1
	12	1
	14	1
	18	1
	20	1
	46	1

word	number	frequency
------	--------	-----------

so	11	2
	10	1
	19	1

word	number	frequency
ss	46	1

word	number	frequency
sy	11	4
	14	3
	13	2
	16	2
	10	1
	26	1
	30	1

letter	number	frequency
t	1	1
	4	1
	2	0
	3	0
	5	0
	6	0
	7	0
	8	0
	9	0

word	number	frequency
tl	10	1
	16	1

word	number	frequency
to	12	1
	24	1

word	number	frequency
ty	10	5
	13	3
	16	2
	18	2
	21	2
	11	1
	14	1
	19	1

	24	1
	29	1

letter	number	frequency
v	4	1
	1	0
	2	0
	3	0
	5	0
	6	0
	7	0
	8	0
	9	0

letter	number	frequency
x	4	1
	1	0
	2	0
	3	0
	5	0
	6	0
	7	0
	8	0
	9	0

letter	number	frequency
y	2	36
	1	30
	3	29
	5	20
	4	15
	7	13
	8	12
	6	11
	9	7

word	number	frequency
ya	12	1
	18	1

word	number	frequency
yd	10	1



word	number	frequency
yk	10	1

word	number	frequency
yl	11	1

word	number	frequency
yy	10	1
	13	1

letter	number	frequency
*	1	1
	2	1
	4	1

### A9. Mapping list for most frequency letters and numbers

letter	number	frequency
m	1	5
	2	5
	3	3
	8	3
n	2	2
	4	2
o	1	24
	2	20
	3	16
	4	14
	5	12
oc	10	1
od	12	1
	14	1
	16	1
ok	24	2
	10	1
	13	1
	20	1
ol	10	14
	13	12
	12	11
	11	9
om	12	3
	10	2

	13	2
or	10	19
	12	13
	13	12
os	19	11
	12	4
	13	4
ot	12	4
oy	11	1
	14	1
p	1	1
	3	1
	6	1
q	1	1
	2	1
	5	1
	6	1
	8	1
qo	9	1
	14	4
	10	2
	11	2
qy	13	2
	11	1
	13	1
r	14	1
	1	48
	2	26
	3	21
	5	19
ra	4	14
	13	1
rk	16	1
	18	1
ro	11	1
	13	1
	14	1
ry	15	2
	11	1
	13	1
	14	1
	19	1
	26	1

s	2	54
	1	46
	3	41
	5	32
sh	13	2
	10	1
	12	1
so	11	2
	10	1
	19	1
ss	46	1
sy	11	4
	14	3
	13	2
	16	2
	10	1
	26	1
	30	1
t	1	1
	4	1
tl	10	1
	16	1
to	12	1
	24	1
ty	10	5
	13	3
	16	2
	18	2
	21	2
v	4	1
x	4	1
y	2	36
	1	30
	3	29
	5	20
ya	12	1
	18	1
yd	10	1
yk	10	1
yl	11	1
yy	10	1
	13	1
*	1	1

	2	1
	4	1

**A.10. Statistics for marginal stars of each page**

Page number	The colour of stars	The number of stars	Arrangement	Location in text (White stars)	Location in text (coloured stars)
f58r	white	3	WWW	1 16 26	No coloured stars
f58v	white	1	W	1	No coloured stars
f103r	White and coloured	19 12 white 7 coloured	BW <b>W</b> BWBWW BWBWBWW BWW	5 7 12 18 21 27 30 36 41 43 48 52	1 9 16 24 33 38 45
f103v	White and coloured	14 7 white 7 coloured	BWBW	5 11 17 21 29 37 42	1 9 14 20 27 34 39
f104r	White and coloured	13 6 white 7 coloured	BW <b>B</b> BWB WBWBWBW	5 16 22 31 36 43	1 10 12 19 27 34 39
f104v	White and coloured	13 6 white 7 coloured	BWBW	6 12 19 27 32 37	1 8 15 22 29 35 40
f105r	White and coloured	10 4 white 6 coloured	<b>B</b> BWBWBWBW	10 16 25 30	1 6 13 23 27 33
f105v	White and coloured	10 5 white 5 coloured	BWBW	5 14 20 25 32	1 8 18 23 29
f106r	White and coloured	15 7 white 8 coloured	BWBW	4 10 15 22 27 34 38	1 8 13 18 24 32 36 42
f106v	White and coloured	14 7 white 7	WBWB	1 9 15 20 28 34 40	5 11 18 26 32 37 42

		coloured			
f107r	White and coloured	15 7 white 8 coloured	BWBW	4 13 18 23 29 37 44	1 8 16 20 26 33 41 47
f107v	White and coloured	15 7 white 8 coloured	BWBW	5 11 20 25 32 37 43	1 8 16 23 29 35 41 45
f108r	White and coloured	16 11 white 5 coloured	WWBWWWB WWBWBWWB	1 5 10 14 17 24 29 35 39 41 45	8 21 31 37 48
f108v	White and coloured	16 8 white 8 coloured	BWBW	5 12 20 26 34 37 43 48	1 7 15 23 30 40 45 50
f109r f109v f110r f110v	Missing folio				
f111r	White and coloured	17 8 white 9 coloured	BWBW	6 10 17 23 29 34 39 48	1 8 13 20 26 32 36 44 51
f111v	White and coloured	19 9 white 10 coloured	BWBW	4 8 13 18 23 29 32 37 45	1 6 10 16 21 26 31 34 39 48
f112r	White and coloured	12 6 white 6 coloured	BWBW	5 15 23 31 37 42	1 11 19 27 34 39
f112v	White and coloured	13 6 white 7 coloured	BWBW	7 15 23 30 36 42	1 11 20 27 32 39 45
f113r	White and coloured	16 8 white 8 coloured	WBWB	1 8 13 20 27 34 39 45	4 10 16 22 30 37 42 49
f113v	White and coloured	15 7 white 8	BWBW	4 10 18 25 33 38 45	1 7 13 21 28 36 42 47

		coloured			
f114r	White and coloured	13 6 white 7 coloured	BWBW	4 11 19 24 32 39	1 8 14 22 28 35 42
f114v	White and coloured	12 6 white 6 coloured	WBWB	1 8 14 20 26 33	6 11 18 23 29 36
f115r	White and coloured	13 6 white 7 coloured	BWBW <b>B</b> W BWBWBW	4 10 22 29 37 42	1 8 13 19 26 35 40
f115v	White and coloured	13 6 white 7 coloured	BWBW	6 11 18 24 31 37	1 8 13 21 29 34 41
f116r	White and coloured	10 5 white 5 coloured	WBWB	1 7 12 18 24	4 10 15 21 26
f116v	No stars				

**A.11. Digits mining**

1	f58v	y s			6 (9pages)	f104r	l y s
3	f58r	s o r m				f104v	l o y s r
4	f105r	r l			y7 s4 o4		
5 (3pages)	f105v	r l y g s				f105r	r l
y3 l2 r2	f108r	y				f112r	o n y s
	f116r	y l s q r				f112v	y
						f114r	y
						f114v	o s
						f115r	y o
						f115v	y

7 (13pages)	f103r	l s	f106v	l	8 (7pages)	f106r	l y
y10 l7 s3	f103v	y l	f107r	l y	y5 l4 o2	f107r	l y
	f104r	l y s	f107v	y		f107v	y
	f104v	l o y s r	f112v	y		f108v	o r
	f106r	l y	f113v	t		f111r	l o s y
			f114r	y		f113r	l y
			f115r	y o		f113v	t
			f115v	y			

9 (2pages)	f111r	l	10 (4pages)	f105r	al	f111v	ar
		o			ar		al
		s			or		am
		y			ol		or
					il		gm
12		ar 4					
y2		al 4					
o2	f111v	l			am		qo
		o			ry		dy
		y			ot		
		m			yl		
		r			oy		
				f105v	ar		
					al		
					or	f116r	ar
					ol		ol
					am		al
					dl		qo
					il		lm
					ky		os

12 (3pages)	f103r	ol	f114v	ol
		am		dm
		ot		ar
		ar 3		op
		ar 3		
		am 2		
	f112r	al		
		ar		
		ol		
		am		
		or		
		yk		
		sy		



13 (6pages)	f104r	ar	f112v	al	f115v	al	
		ol		ar		ar	
		or		ol		ol	
		al 6		am		lr	
		al 6		or		ld	
		ol 5		qo		or	
		or 5					
				ly			
				os		f114r	al
				lo		ol	ol
		ar	ar				
	f104v	or	do				
	ol	oy					
	ar		f115r	ar			
	or		al	or			
	am		ro	om			
	ls		od	qo			
	ll						
	al						

14 (2pages)	f103v	ol	15 (4pages)	f106r	al	f113v	al		
		ar			ar		or		
		or			ol		ol		
		ol 2			ro		dy		
		ar 2			al		ro		
		al 2			lr		lr		
		dy 2			dy				
					f106v		al	f107r	al
					ar		ar	ol	om
	ol	am	or	ky					
	ko	an							
	dy		f107v	al					
	am		ol	ar					
	ry		or	ky					
			dy						

16 (2pages)	f108r	al
		ar
		or
al 2		ol
ar 2		om
or 2		
ol 2	f108v	al
om 2		ol
		am
		ar
		or
		om
		rl
		lo

17	f111r	al
		ar
		ol
		am
		or
		om

19 (2pages)	f103r	ol
		am
ar 2		ot
dy 2		ar
an 2		lr
		ok
		lo
		dy
	f111v	ar
		al
		am
		or
		gm
		qo
		dy