SCHOOL OF
ELECTRICAL AND
ELECTRONIC ENGINEERING

THE UNIVERSITY of ADELAIDE

141: Cracking the Voynich manuscript code (2nd thesis draft)

Student name: Ruihang Feng
Student ID: 1674940

ELEC ENG 7076 A/B MASTERS PROJECT

M.E. in Avionics & Electronic Systems Engineering
M.E. in Computer Systems Engineering
M.E. in Electrical and Electronic Engineering
M.E. in Electrical & Sustainable Energy
M.E. in Telecommunications 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: 03 June 2016

Supervisor: Prof. Derek Abbott

Adviser: Dr. Brian Ng

Abstract

The Voynich manuscript, written by a kind of unknown language, is a famous mystery in the field of linguistic. Over the years, many researchers had tried to crack this manuscript, but no one had succeeded. In this thesis (2nd thesis draft), the project of cracking the Voynich manuscript will be introduced from five aspects:

- Introduction.
- Related work (the history of the Voynich manuscript research).
- Requirement.
- Proposed method.
- Project management.

Moreover, the methods of solving this project mainly involve two aspects:

- Matlab.
- Statistics.

Content

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.

There are six sections in the Voynich manuscript: herbal, astronomical, cosmological, biological, pharmaceutical and recipes section. 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.

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 result 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]. Therefore, major data will be extracted from EVA 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 [5] 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 [6]. In the following years, she summed up different features of the Voynich manuscript text [7].

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 [8].

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 [9].

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 [10].

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 [11].

First study group:

The first study group (FSG) was founded at 1944, dissolved at 1946 [12]. 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 [13]. The details of the transcription alphabet are as shown in the Appendix Section A.7.

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.
- Illustration research: look for laws from some illustrations.
- Marginal symbols investigation: make a thorough inquiry about laws from marginal symbols.
- 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.1, the proposed methods of this project are divided into four 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 words which look like digits from some paragraphs of the manuscript.
- The frequency of special 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.2. The word which is as shown in the Appendix Section A.3 is extracted from the Voynich manuscript, it is obvious that the form of the word in the Appendix Section A.3 is like '*##'. According to the method of comparison mentioned above, this word maybe means 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.4.

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

- The number of different elements in the illustration.
- The feature of words.
- Relations between different illustrations.

4.3 Phase 3: Marginal symbol research.

A page which contains marginal symbols is as shown in the Appendix Section A.5. This phase also requires proficiency in programming by using Matlab. During the process of this process, there are four major aspects:

- Ordering and quantitative features of the symbols at the margin of the page.
- Linguistic features of the words after marginal symbols.
- The differences between different marginal symbols in one page.
- Relations between two different pages containing marginal symbols.

4.4 Phase 4: Translation.

During the process of this phase, team will try to translate some parts of the Voynich manuscript. The major method is as shown in the Appendix Section A.6. There are four important steps:

- Tool: Matlab.
- English linguistic research.
- Structure of words in the Voynich manuscript investigation.
- Some parts of the manuscript translation.

5. Project management.

5.1 Deliverables

As shown in table 1, deliverables involve eleven parts.

Deliverable	Deadline
Proposal seminar	4st of April, 2016
Project wiki (introduction)	Semester 1, week 5.
Thesis (1st draft)	22th of April, 2016.
Thesis (2nd 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.

Table 1

5.2 Work breakdown

The details about tasks are as shown in the Appendix Section A.1. The key tasks involve two aspects:

- Text investigation (digits).
- Translation.

5.3 Timeline

Timeline of project involves five parts. The specific details are as shown in the table 2.

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 4: Translation	5-9

Table 2

5.4 Task allocation

Task allocation is divided into five parts:

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 4: Translation	Ruihang Feng, Yaxin Hu

Table 3

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 content reorganization.

- 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.

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

Table 4

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 Bling, Grammarly and other websites.

5.7.5 Health issues

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

6. Preliminary outcomes

As the introduction in the section 4.1, the phase 1 'Text investigation' is divided into three aspects:

- The total number of words in the Voynich manuscript.
- The frequency of special words.
- The words which look like digits from some paragraphs of the manuscript.

Over the past few weeks, this phase had been finished by using the Matlab and statistics. The result is shown in the Chapter 7.

7. Preliminary results

7.1. Phase 1: Text investigation

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

7.1.1. The total number of words

In this stage, Matlab is used to count the total number of words in the Voynich manuscript. The result is 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)	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.

7.1.2. 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.

7.1.2.1: The frequency and the number of simple letters

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

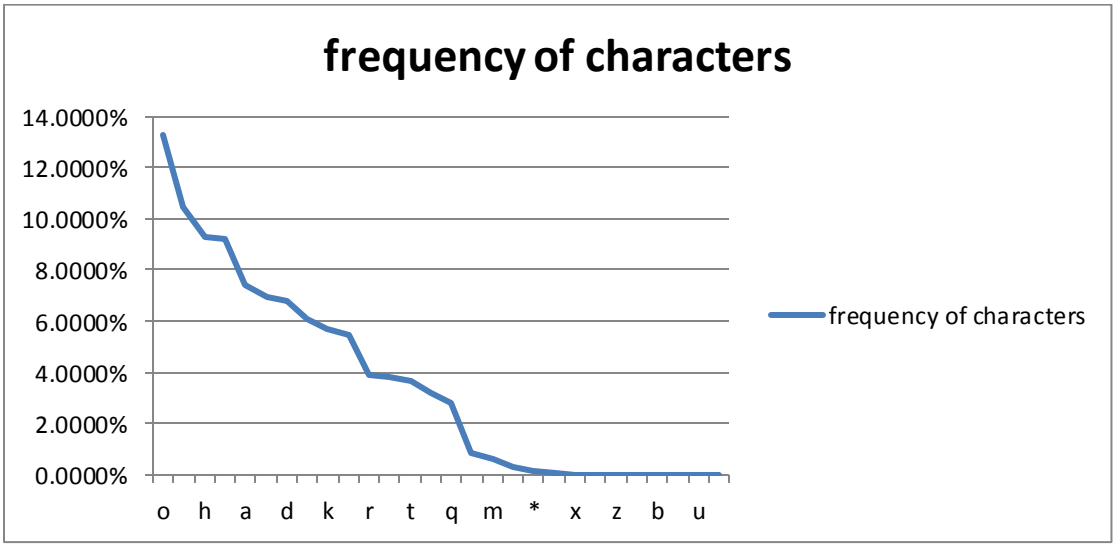


Figure 1

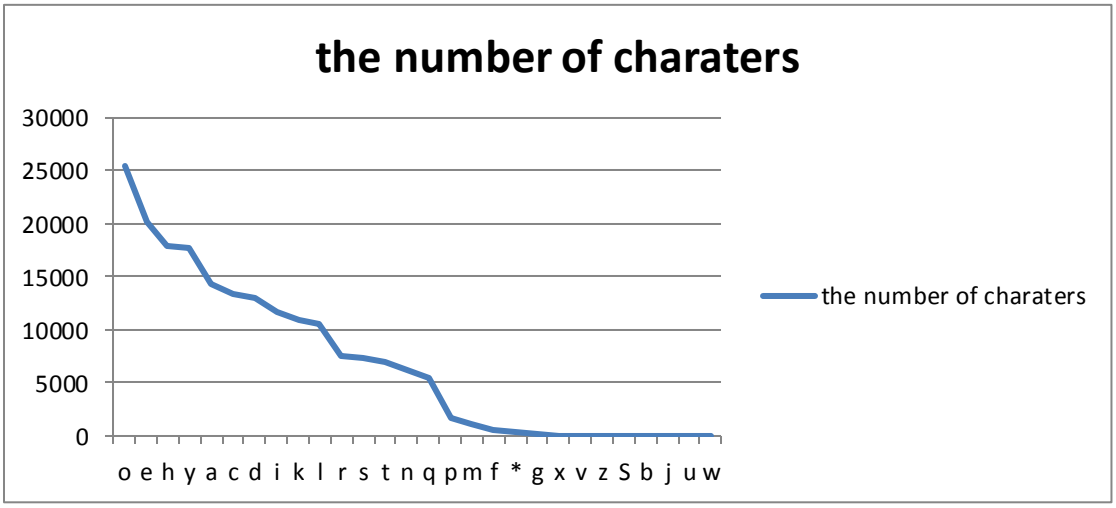


Figure 2

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 those letters have never appeared in the Voynich manuscript. In addition, the letter with the highest frequency (0.133) is ‘o’.

7.1.2.2: The frequency of words

The result is shown in the figure 4.

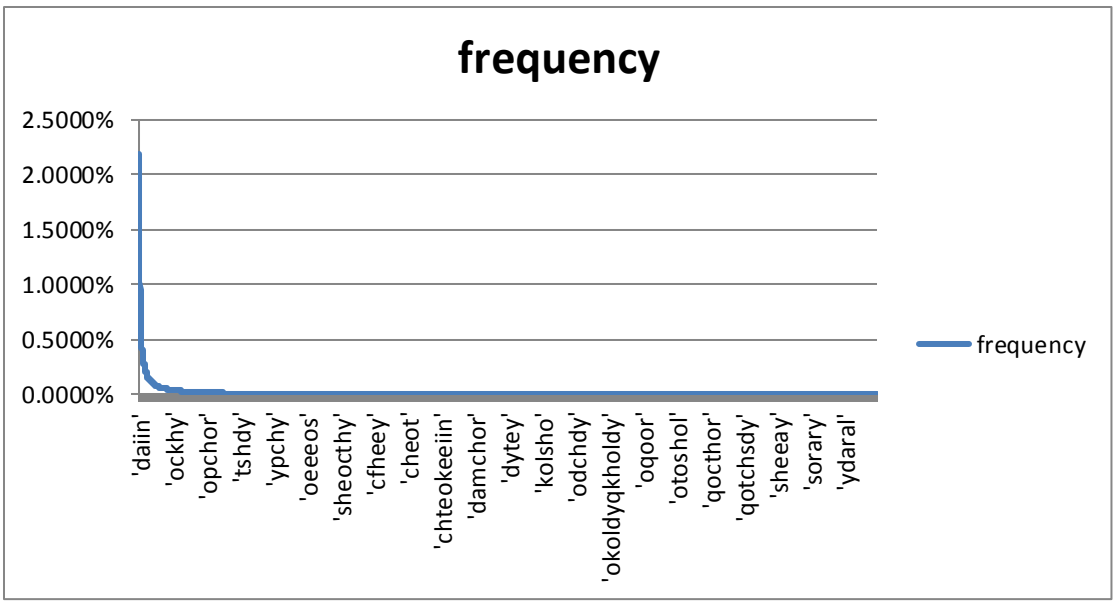


Figure 4

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

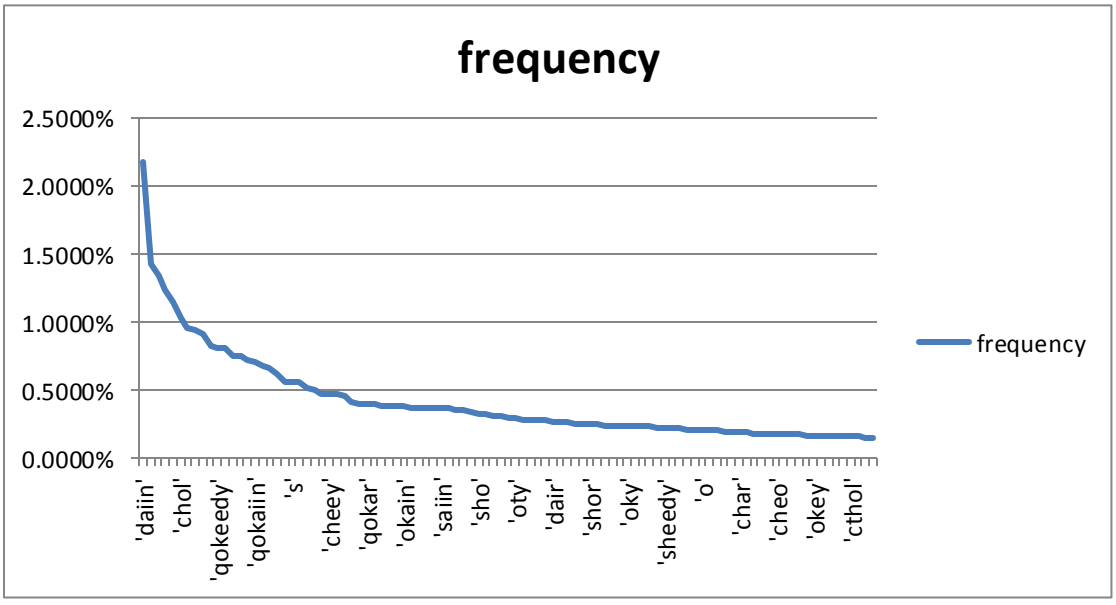


Figure 5

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 axes still

can't show every word. In that case, I extract the first 20 words. The result is shown in the figure 6 and figure 7.

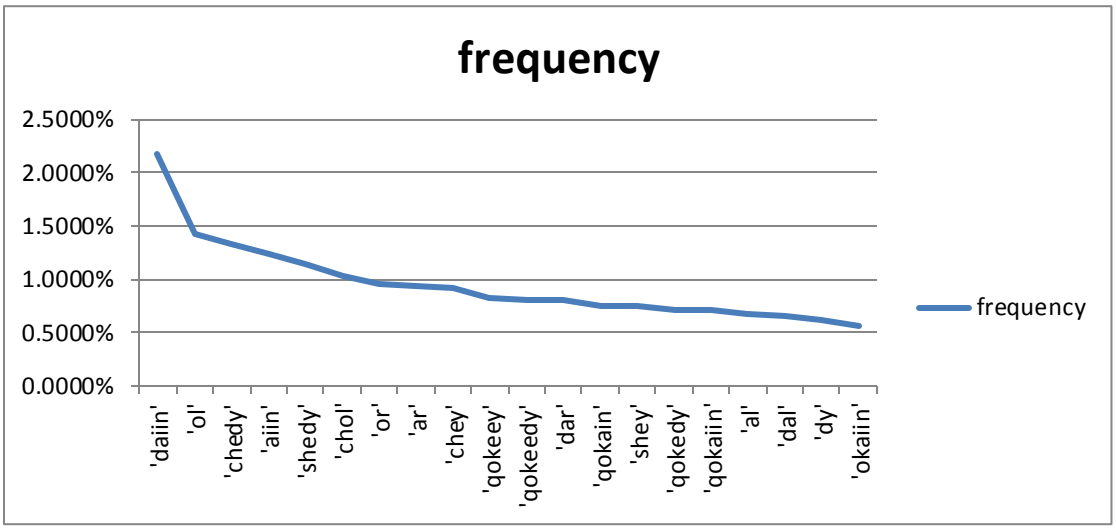


Figure 6

	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’.

7.1.2.3: 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 language is Latin,

English, French and German [14]. So in this section, I 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 [15].

Part 1: The Voynich versus Latin.

The occurrence frequency of letters in the Voynich manuscript and Latin is shown in the figure 8 and figure 9.

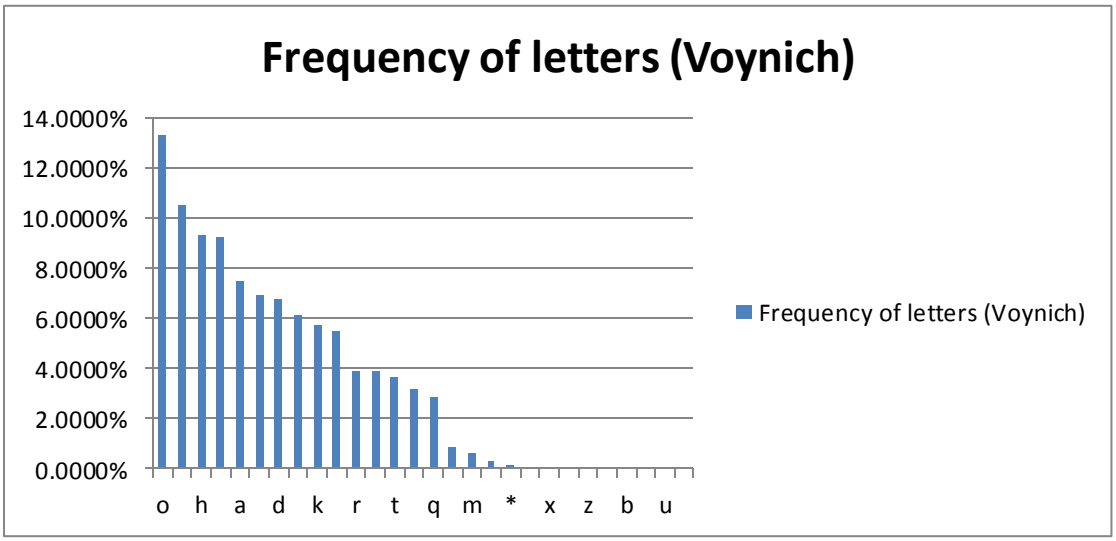


Figure 8

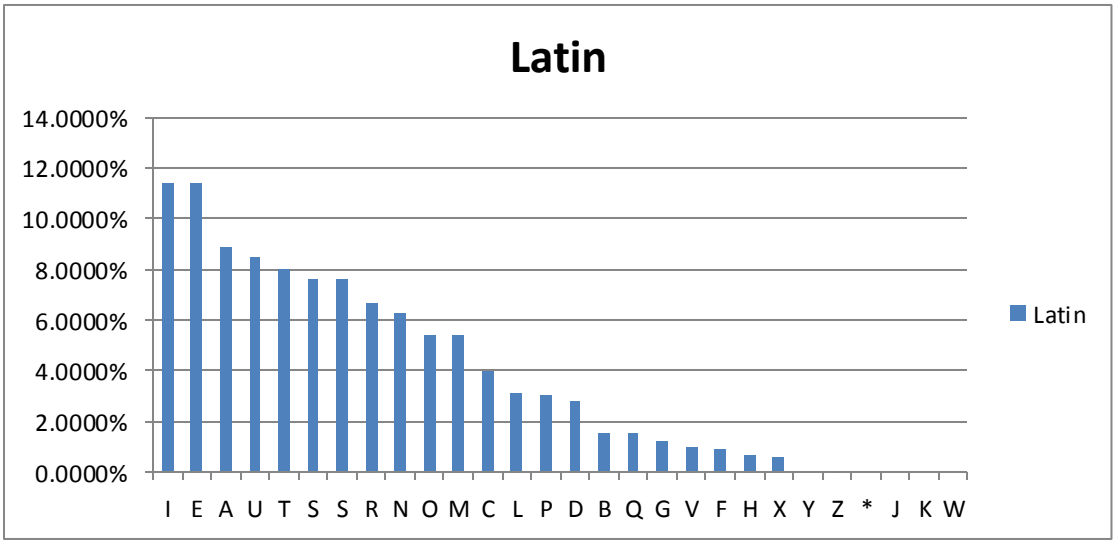


Figure 9

In order to analyse conveniently, I change the Figure 8 to the form of proportion, which is shown in the figure 10.

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

Figure 10

According to the Figure 10, we can find 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, I calculated the correlation between the Voynich and Latin, the result is 98.60%, which means the ‘o’ in the Takahashi edition may stand for ‘I’ in Latin.

Part 2: The Voynich versus English.

The occurrence frequency of letters in the Voynich manuscript and English is shown in the figure 11 and figure 12.

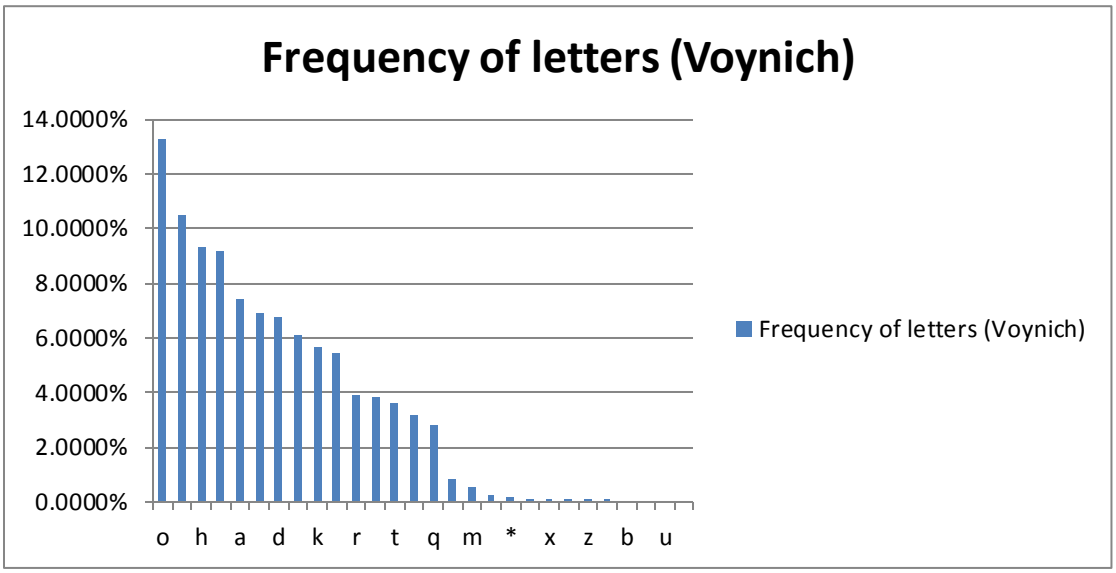


Figure 11

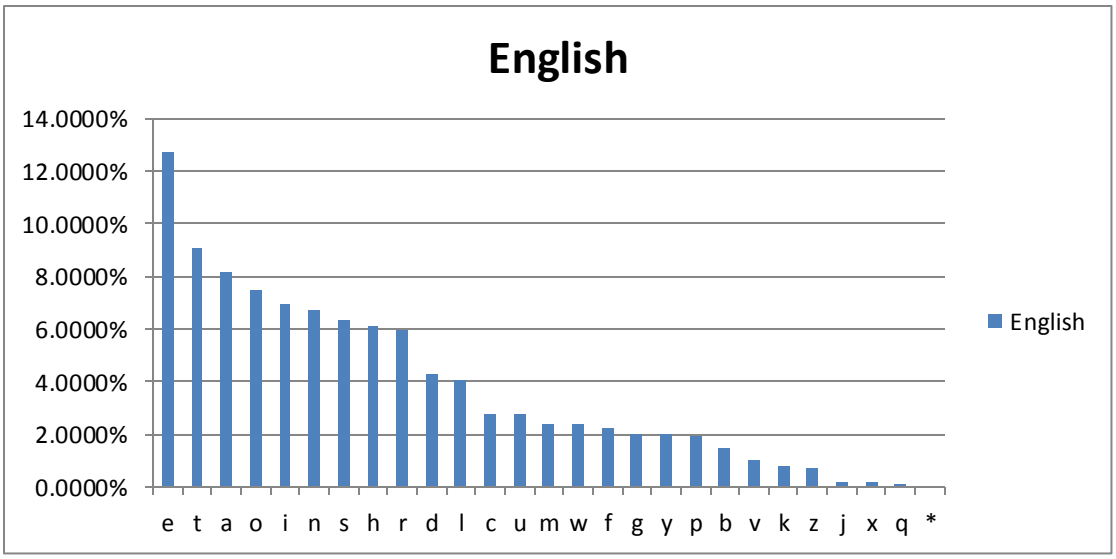


Figure 12

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, I calculated the correlation between the Voynich and Latin, 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 result is shown in the part 5.

Part 3: The Voynich versus French.

The occurrence frequency of letters in the Voynich manuscript and French is shown in the figure 14 and figure 15.

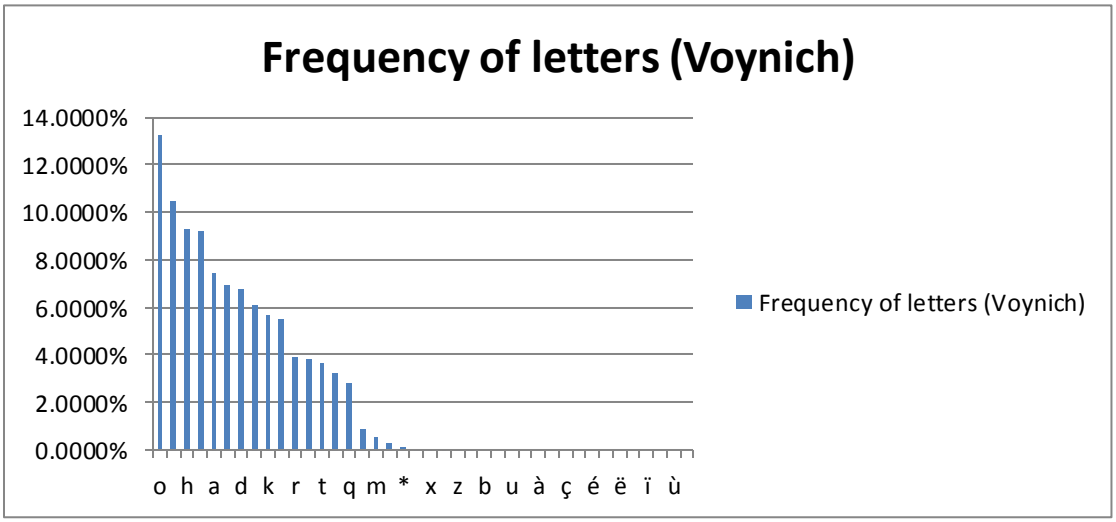


Figure 14

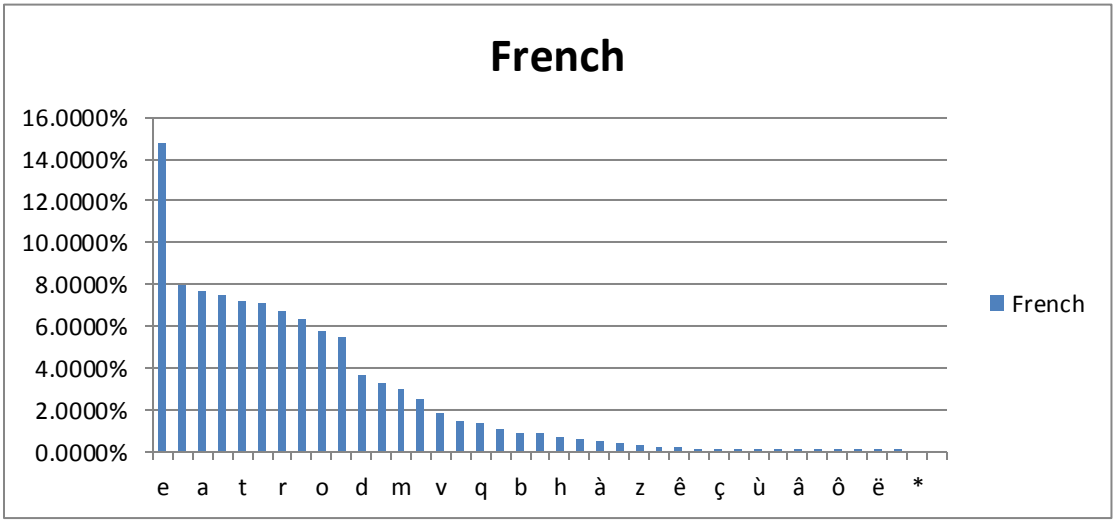


Figure 15

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%		ô	0.0200%
î	0.0000%		œ	0.0200%
ï	0.0000%		ë	0.0100%
ô	0.0000%		ï	0.0100%
ù	0.0000%		*	0.0000%
œ	0.0000%		S	0.0000%

Figure 16

According to the figure 16, I calculated the correlation between the Voynich and Latin, the result is 98.11%.

Through 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 sixteen letters in the first column have never appeared in the Voynich manuscript, but they appear in French. So there still many differences between the Voynich and French.

Part 4: The Voynich versus German.

The occurrence frequency of letters in the Voynich manuscript and German is shown in the figure 17 and figure 18.

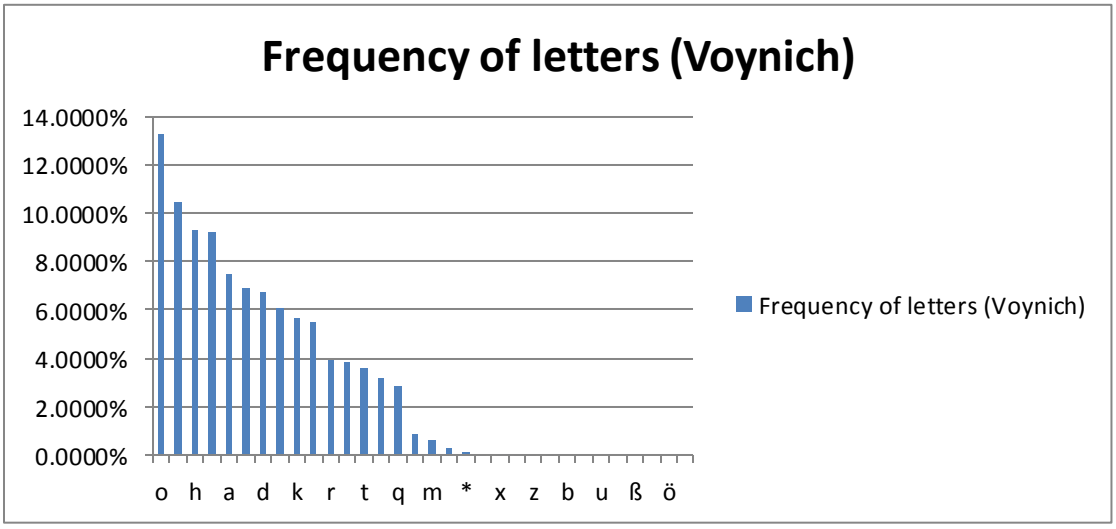


Figure 17

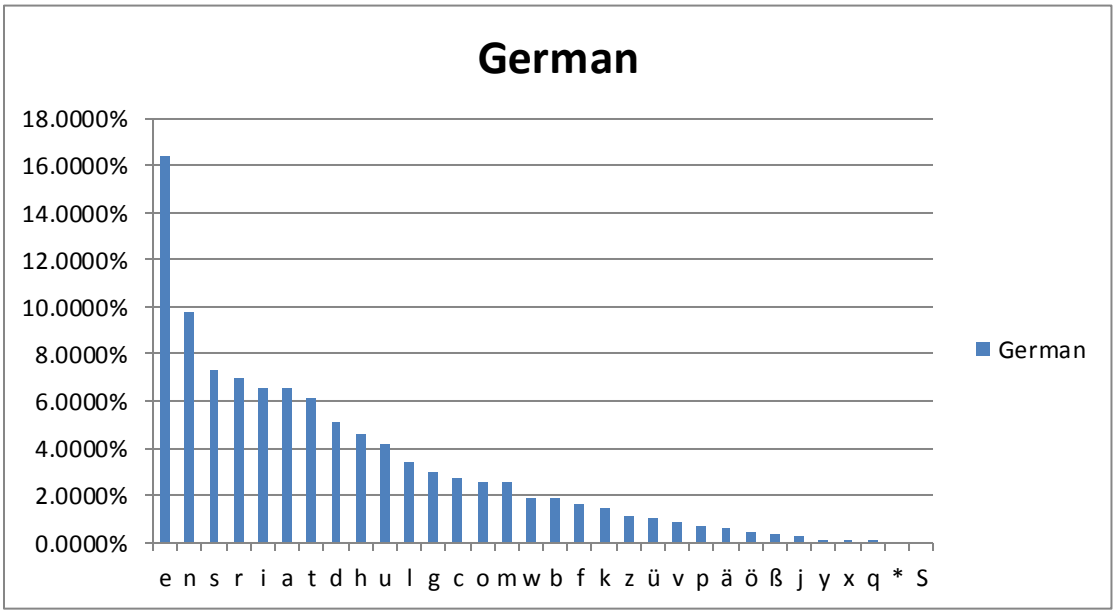


Figure 18

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%	j	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%
ü	0.0000%	S	0.0000%

Figure 19

According to the figure 19, I calculated the correlation between the Voynich and Latin, the result is 95.86%.

Through 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 eight letters have never appeared in the Voynich manuscript, but they appear in

German. In addition, the frequencies of those eight letters in German are very low, so I think maybe those eight letters are not belong to the range of commonly used letters. 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, I also 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, I extract the same number of words from every book. The results are shown in the Figure 16.

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_sourc	235942	37457	6859
karr_sous_les_tilleuls_source	213019	37484	5371
Faust	195835	30654	6079
FAUST: EINE TRAG	229801	37082	7242
Carlos Ruiz Zafòn	238584	37270	7423

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

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 Zafòn	4797	64.62%

Figure 16

As shown in the figure above, we can find that the characteristics of the books which were written by German have the highest degree of similarity with the characteristics

of the Voynich manuscript, especially in the aspects of 'Unique words number', 'characters per word' and 'Ratio (words appears once/Uwn)'. In addition, as the analysis in the part 4 section 7.1.2.3, I think maybe the language which was used in the Voynich manuscript is a branch of German.

7.1.3. 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.2. In addition, the method which I use is introduced in the section 4.1.

The result of searching the words with the form '*##' in the Voynich manuscript is shown in the Figure 17.

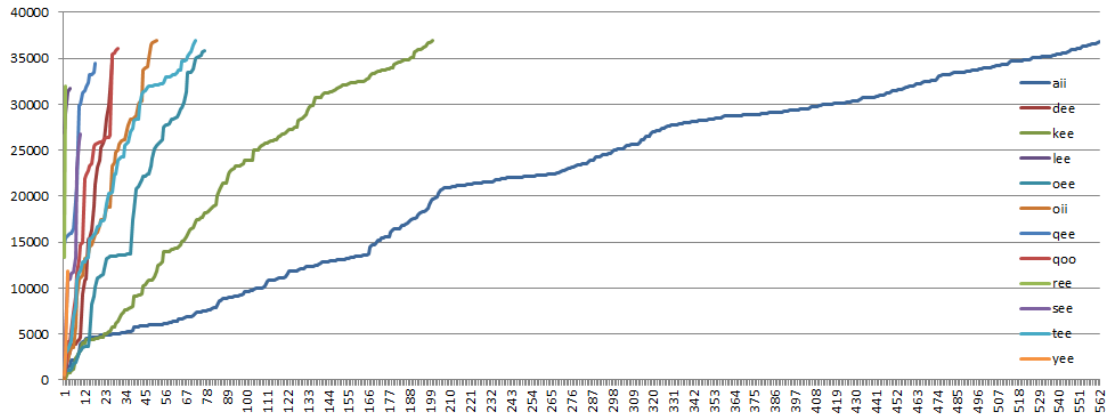


Figure 17

As shown in the Figure 17, the words with the form '*##' in the Voynich manuscript involve: 'aii', 'dee', 'kee', 'lee', 'oee', 'oii', 'qee', 'qoo', 'ree', 'see', 'tee' and 'yee'. X axes means the occurrence number of each word. As shown in the Figure , the most commonly used word is 'aii' and the occurrence number of 'aii' is 563. In addition, the occurrence number of 'ree' is the smallest, which are 2. Y axes means the positions of each word. For example, the position of 563rd 'aii' is 36821, which means this word is the 36821st word in the Voynich manuscript.

As the analysis above, I infer that 'aii' may means seven in Roman (VII).

Then, I extract the words with the form '*###' by using the same method. The result is shown in the Figure 18.

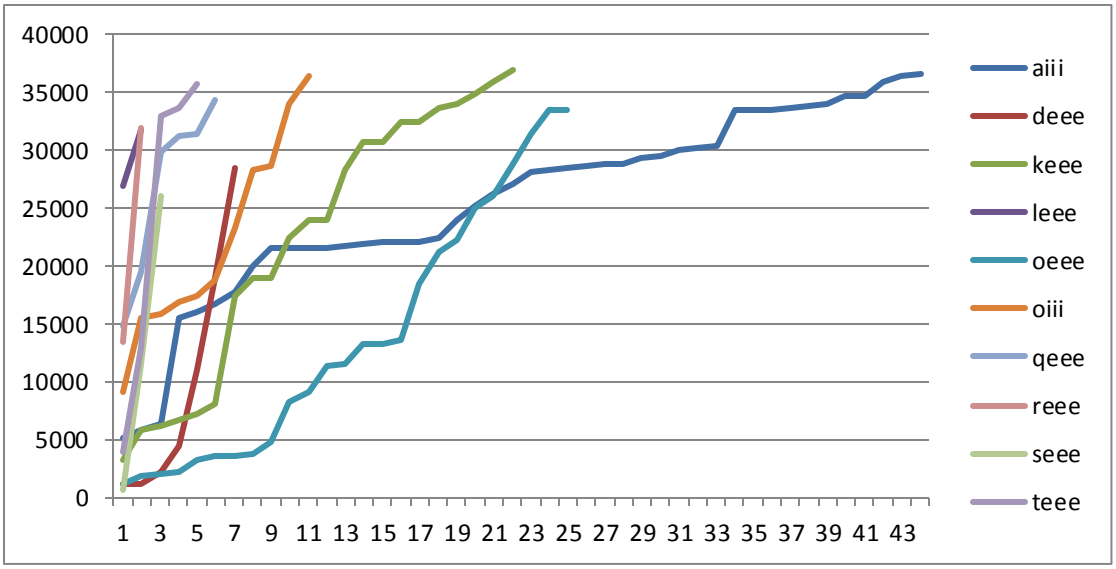


Figure 18

From the Figure 18, we can see 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.8.

7.2 Preliminary conclusion

As the analysis in the section 7.1, we can infer that the language which is used in the Voynich manuscript may be a branch of German. In addition, the method of expressing digits in the Voynich manuscript may be using Roman.

8. Comment on progress

Over the past few weeks, the ‘Phase 1: Text investigation’ has been finished.

Comparing with the initial plan which is shown in the section 5.3, the status of our progress is almost normal. When we made our plan at the beginning of Semester 1, we didn’t find that the first phase is the most difficult. So we spent most time of the semester 1 to analyse this phase. Through we have not finished ‘Phase 2: Illustration investigation’ and ‘Phase 3: Marginal symbol research’, we have finished the most difficult phase.

Next, we will continue to search phase 2 and 3. Then we will search ‘Phase 4: Translation’.

9. Conclusion

This project is divided into four phases: Text investigation, illustration research, marginal symbol investigation and translation.

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

- Use statistical method Matlab to search the language rules in the Voynich manuscript.
- Search laws from illustration.
- Investigate laws from marginal symbols.

On the other hand, the major works of this project can be achieved by using computer.

Over the past few weeks, the first phase has been finished. As the analysis in the section 7.1, we can infer that the language which is used in the Voynich manuscript may be a branch of German. In addition, the method of expressing digits in the Voynich manuscript may be using Roman.

10. 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] 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

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

[7] 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.

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

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

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

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

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

[13] R. Zandbergen (2016), *History of research of the Voynich MS* [Online]. Available:

http://www.voynich.nu/solvers.html#n43

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

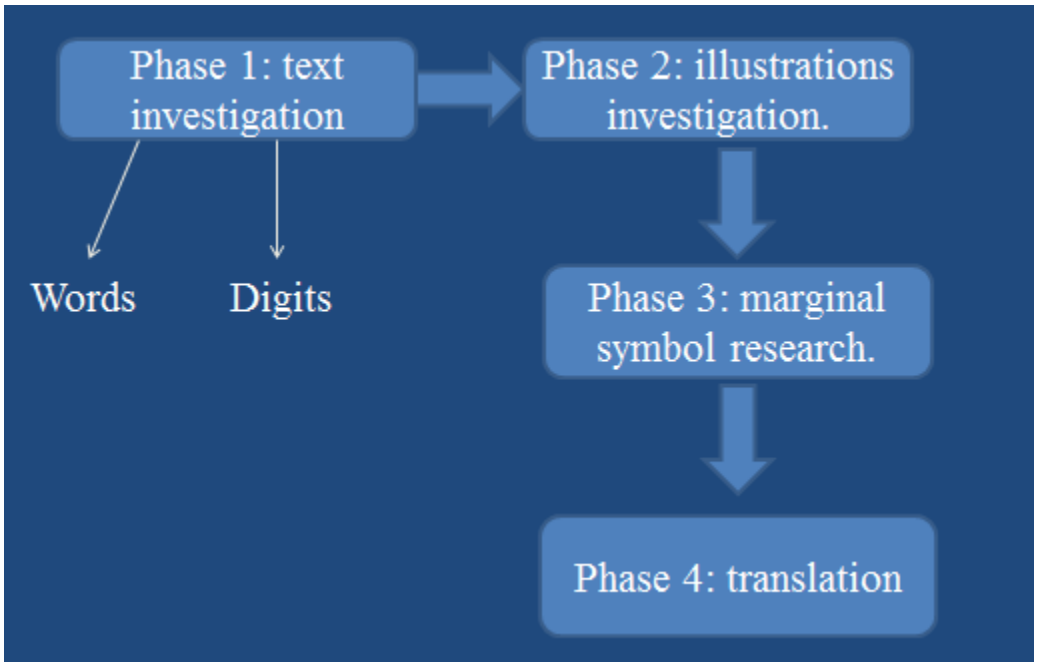
Available: https://en.wikipedia.org/wiki/Medieval_literature#Languages

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

Available: https://en.wikipedia.org/wiki/Letter_frequency

11. Appendix

A.1.Proposed Method



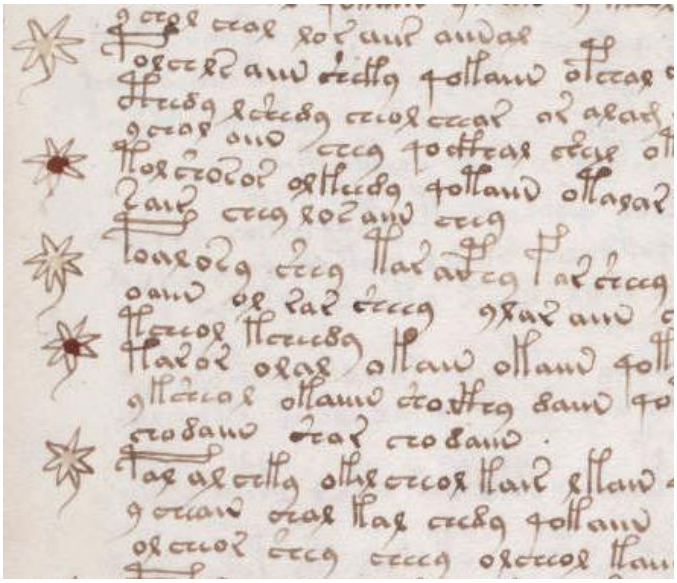
A.2. Roman Numeral

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

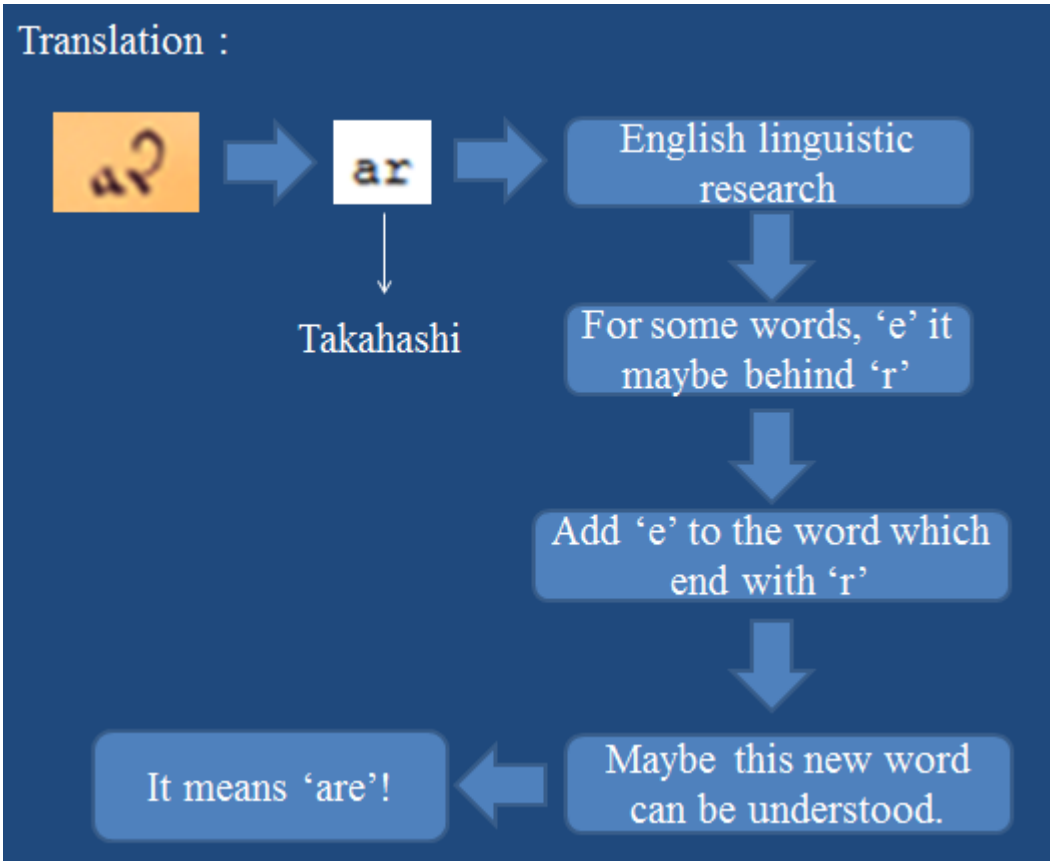
A.3. Words from the Voynich manuscript.

A.4. Illustration from the Voynich manuscript.

A.5. Marginal symbols from the manuscript.



A.6. Translation.



A.7. 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
ɾ	L	E	E		˙	IIQ	IIR	U
2	Q	R	R		˙	IIIQ	IIIR	0
ɾ	CT	T	S		˙	U	L	D
2	ET	S	Z		˙	N	N(*)	N
4	H	H	P		˙	M	M(*)	M

A.8. 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										