THE UNIVERSITY OF ADELAIDE

SCHOOL OF ELECTRICAL & ELECTRONIC ENGINEERING

ADELAIDE, SOUTH AUSTRALIA, 5000

Cracking the Voynich manuscript code

Student name: Yaxin Hu

Student ID: a1672395

ELEC ENG MASTER PROJECT NO. 141

B.E. in Electrical and Electronic Engineering

Date submitted: 03/Jun/2016

Supervisor: Prof. Derek Abbott and Dr. Brain Ng.

Executive summary

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.

Contents

1. Introduction

1.1 Project Background

The Voynich the manuscript was created in the first half of the fifteenth century (probably between 1404 and 1438) [1]. No one today knows what it says or who wrote it. The book is in a strange alphabet. At 1912, a book collector named Wilfried Voynich found it in an Italian Jesuit college [1].

Since this book cannot be read, 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 [2].

b) **Astronomical**

There are circular diagrams such as suns, moons, and stars which suggest this part as something about astronomy or astrology [2].

c) **Biological**

Mostly naked women shows that this part should be biological section [2].

d) **Cosmological**

Circular diagrams of obscure nature make this section as cosmological section [2].

e) **Pharmaceutical**

Drawings of isolated plants parts and objects resembling apothecary jars show that this section should be something about pharmaceutical [2].

f) **Recipes**

This part are full pages of text in short paragraphs [2].

1.2 Aim

The aim of this project is to search the text and determine whether there are any possible features that can be used to decode the Voynich manuscript using statistical methods. The investigation of languages and linguistics is required to be processed

with the unknown text. But, it is not necessary to fully decode the Voynich manuscript since it is not possible to be done in a one-year project.

1.3 Significance and Motivation

With statistical methods, trying to carry out a project that is used to investigate the language and linguistics of an unknown book is an attempt that may beyond excellent. Trying to find any features of relationships and patterns of the Voynich manuscript could be used to decode the unknown text with unknown languages. It may contribute significant progress in attempting decode a part of the book. The outcomes can be used to further linguistic or language decryption, such as information decoding, search engines and data mining. They can also be used in specific applications such as Google, Turn-it-in, Google translate, Yahoo, and Grammarly.

1.4 Technical Background and Challenges

Data mining as an important part in this project, is the foundation of analysing the Voynich manuscript. It is an interdisciplinary subfield of computer science that is used to process the discovering patterns in large data sets involving methods such as artificial intelligence, machine learning, statistics and database systems [3]. In this project, data mining should be used to test and analyse the specific linguistic and language features.

As the Voynich manuscript has been transcript into an English alphabet version with several method such as the European Voynich Alphabet (EVA). There is an example of a part of text of the Voynich manuscript and EVA in appendix 1.

Since it is an unknown hand-written book for more than five centuries, there is no useful material that can be used to determine the symbols of the manuscript. The way that can be used to determine word allocation is the spacing between different sets of symbols. Also, it is believed that this manuscript has several pages missing. Also, there is strong evidence that many of the book's bifolios were reordered at various points in its history, and that the page order may be different from what it is today [4].

Due to the pre-study, no useful technique can be used to translate or determine the manuscript [5]. Therefore, what we can use is basic linguistics and languages.

1.5 Knowledge Gaps

This project is a decoding project, therefore, it will require a lot of software work with variety of statistical methods to achieve the aim. None of us have mastery so much kinds of particular knowledges in different subjects. Therefore, each of us will be required to develop software programming skill and statistics skill. Beside, since we have no evidence that any kind of particular skill can be used to solve this project, several different kinds of skill will be needed in processing the Voynich manuscript.

2. Requirements

Although it is not necessary to fully decode the Voynich manuscript, this project should present several outcomes:

a) A clear investigation of language and linguistics of the Voynich manuscript

b) Any possible results within the attempts.

c) Any hypotheses within the results.

d) Any decoded text if possible.

3. Related work

The Voynich manuscript has been investigated for almost a century by a large number of professors and specialists. They have contributed several possible hypotheses that can be used in this project through their analysis.

Stephen Bax (2014) states that the Voynich manuscript is not a hoax, and it is probably an explanatory treatise which appears to act as a type of manual for interpreting and transmitting information across cultures [6]. If it is possible, it may lead to a new direction of analysing the Voynich manuscript.

Another work that may contribute possible impact is Gbariel Landin (2001)'s "Evidence of Linguistic Structure in the Voynich Manuscript Using Spectral Analysis". He used statistical method to characterise the Voynich manuscript with natural languages. Zipf's law that he used to analyse on entropy in this book shows that there may exist some linguistic form in Voynich manuscript because the long range correlation, length modal and periodic structures in the Voynich manuscript [7].

A multiple tests of the Voynich manuscript carried out by Roush (2014) shows that there may needs several kinds of attempts such as:

a) Word length distribution
b) Word and image association
c) Word recurrence intervals
d) Zipf's law
e) N-Grams [8]

They made a brief conclusion of these attempting, however, none significant result is approached by them, which may indicated that further attempting should be taken.

Another statistical investigation token by Costa (2013) on the Voynich manuscript in related to vocabulary statistics shows that the Voynich manuscript is similar to natural languages [9].

4. Proposed Methods

4.1 Characterisation of the Voynich manuscript

Mainly, there are several task in characterisation of the manuscript.

a) Total words in the whole manuscript
b) Total characters in the whole manuscript
c) Unique words
d) Unique character
e) Frequency of words
f) Frequency of character
g) Character that only appear at the start or the end of words

Compare these statistical results with known languages may contribute significant progress in determining the features of the Voynich manuscript.

4.2 Text investigation: Digits

Digits investigation will be our first breakpoint in decoding the Voynich manuscript. This part will be taken following by several steps.

a) Find patterns in known language digits such as Roman digits and Greek digits.
b) Trying to search any words in the Voynich manuscript that may related to any patterns in known language digits and locate all of them.
c) Translate all the possible words and check whether these words conform to the images that may nearby the words.
d) Use statistical methods analyse any possible digital patterns that may conform to the Voynich manuscript.
e) Decode all the digits if step d is success.

Digital investigation may contribute significant influence in the whole investigation. If not, the follow investigation will become more important.

4.3 Illustration investigation

Illustrations investigations is associated to the digitals investigation. It will follows several steps:

a) Locate all the images that contains one thing that appears more than once in an image.
b) Number the time that things appears in each image.
c) Trying to search words nearby the image that may conform any digital patterns in known language digits.
d) Decode all the digits if step c is success.

Illustration investigation is a different way that used to investigate digits. The difference is that there may contains different kinds of encryption in the Voynich manuscript if it is encrypted, therefore, it is a way to ensure that digital investigation can solute this possibility.

4.4 Marginal symbol investigation

Marginal symbol investigation is a method that is used to investigate the last section of the Voynich manuscript. In the recipes section, there are many solid stars or hollow stars in front of each paragraph. This method will to goes in the following steps:

a) Locate all the stars in recipes section.
b) Count the number of solid stars and hollow stars separately.
c) Search the texts nearby all the stars that may contain any possible numbers.
d) Compare all the recipes sections and try to find any pattern for any possible numbers.

Marginal symbol investigation is a way that if both digital investigation and Illustration investigation cannot get significant result. It may provide another breakpoint in the whole digital analysis.

5. Preliminary Outcomes and Further steps

Until now, there are several outcomes have been concluded.

a. By comparing the letter frequency of the Voynich manuscript and English, Latin, French, German, Greek and Spanish, there may exist relationships between the Voynich manuscript and anyone of these languages. However, there is no strong evidence shows that there is a relationship between them. Greek is the most possible language the Voynich manuscript used.
b. By comparing the word frequency of the Voynich manuscript and English, there may exist relationship between them, however, there is no strong evidence shows that there is any relationship between them.
c. By locating all possible VII words and VIII words, e, I and o can be considered as possible numerical characters.
d. By comparing the percentage of unique words/total words, word length and the percentage of words appear more than once /total unique words of the Voynich manuscript and three English books, French books or German books, there may exist relationships between them. However, there is no strong evidence shows that there is a relationship between them. German is the most possible language the Voynich manuscript used.

There are several steps will be held on for the next half year:

a. Continue find all possible numerical words in Roman numerals from I to XX, and several obvious pattern numerals such as XX, XXX, C, CC and CCC. Locate all these possible numerical words in the Voynich manuscript.
b. Compare the possible numerical words with the images around these words, try to find any possible evidence that whether there is any potential numerical words that can be used to decoded the Voynich manuscript.
c. Try other kinds of numerical languages such as Greek numerals and continue step b.

6. Middle Sections

6.1 Letter Frequency

Figure 1 shows the letter frequency in Voynich manuscript. There are 24 letters in Voynich manuscript. As the figure shows, that o, e, h, y are the four most frequency letters, and S, z, v, x are the four least frequency letters. The blue line is the tendency of all the letters.

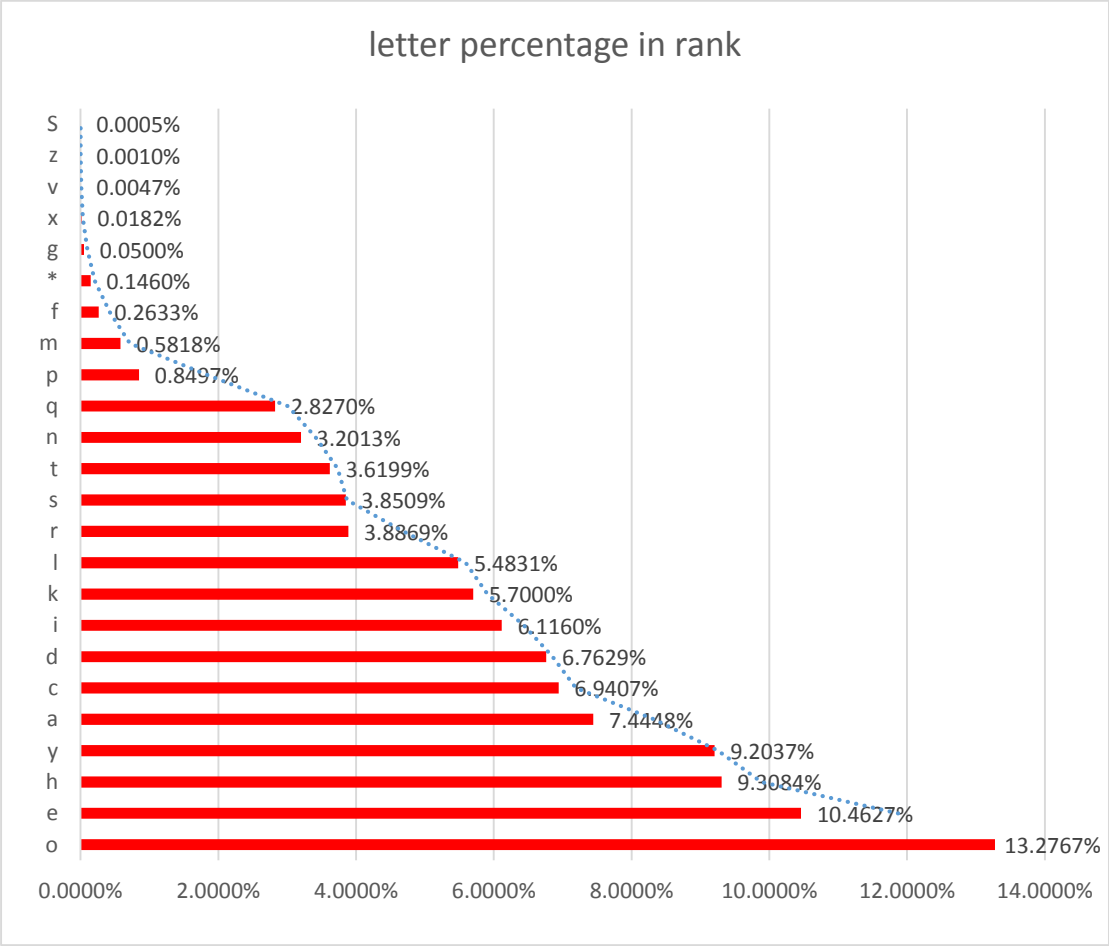


Fig 1. Frequency of the Voynich manuscript

There are six kinds of languages are used in comparing the letter frequency, those are English, Latin, French, German, Greek and Spanish.

Figure 2 shows the letter frequency of English. There are 26 words in total. The most frequency letters are e, t, a and o, and the least frequency letters are z, q, j and x.

Figure 3 shows the letter frequency of Latin. There are 23 words in total. The most frequency letters are i, e, a and u, and the least frequency letters are z, y, x and h.

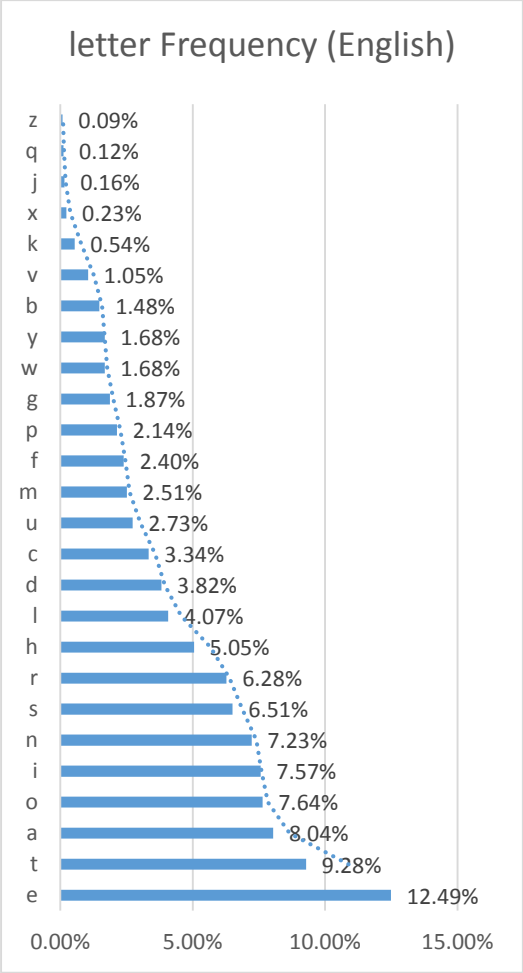


Fig 2. Letter Frequency of English [10]

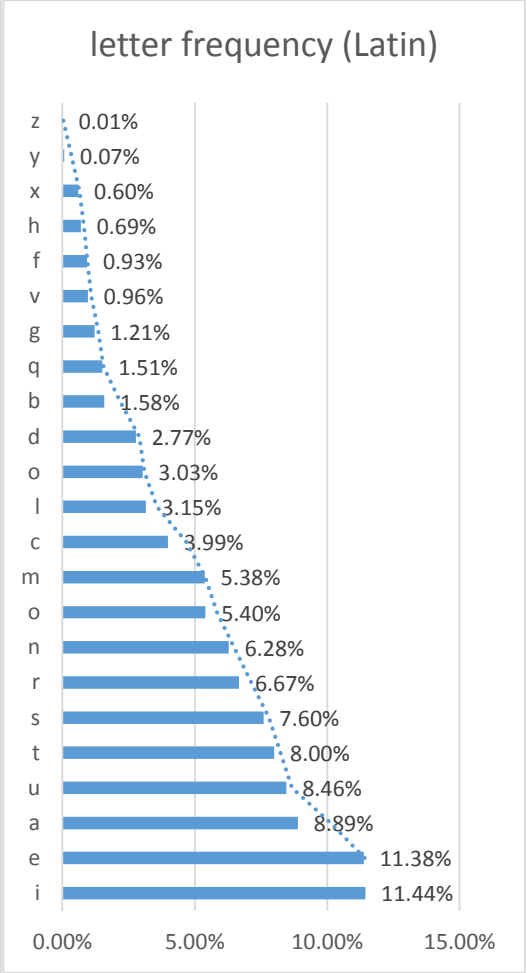


Fig.3 Frequency of Latin [11]

Figure 4 shows the letter frequency of French. There are 38 words in total. The most frequency letters are e, s, a and i, and the least frequency letters are ï, ë, œ and ô. Figure 5 shows the letter frequency of German. There are 30 words in total. The most frequency letters are e, n, s and r, and the least frequency letters are q, x, y and j.

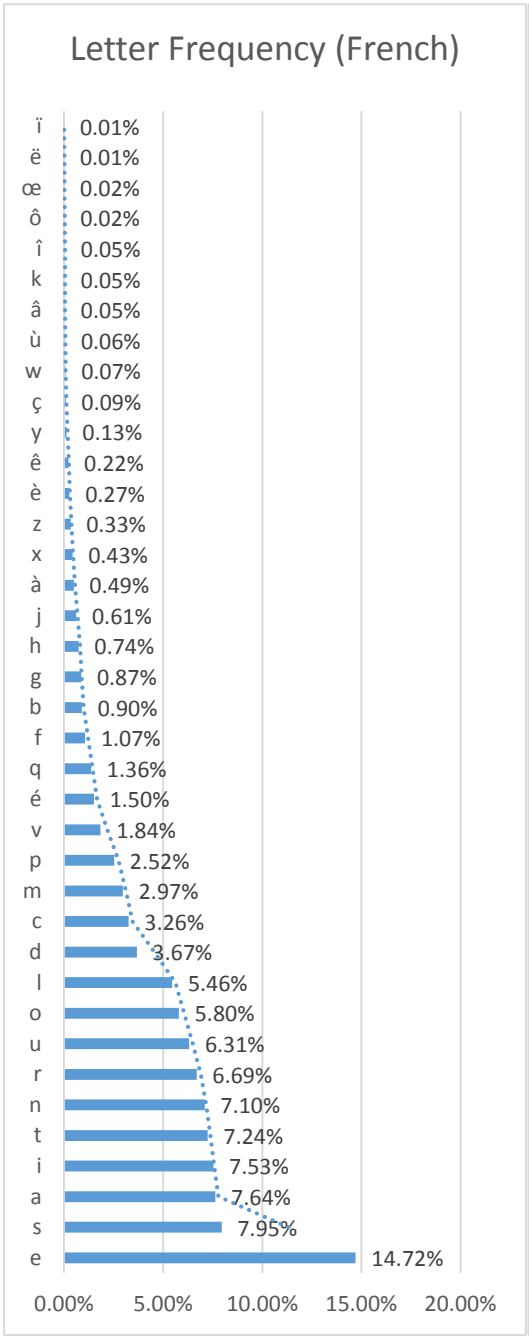


Fig 4. Letter Frequency of French [12]

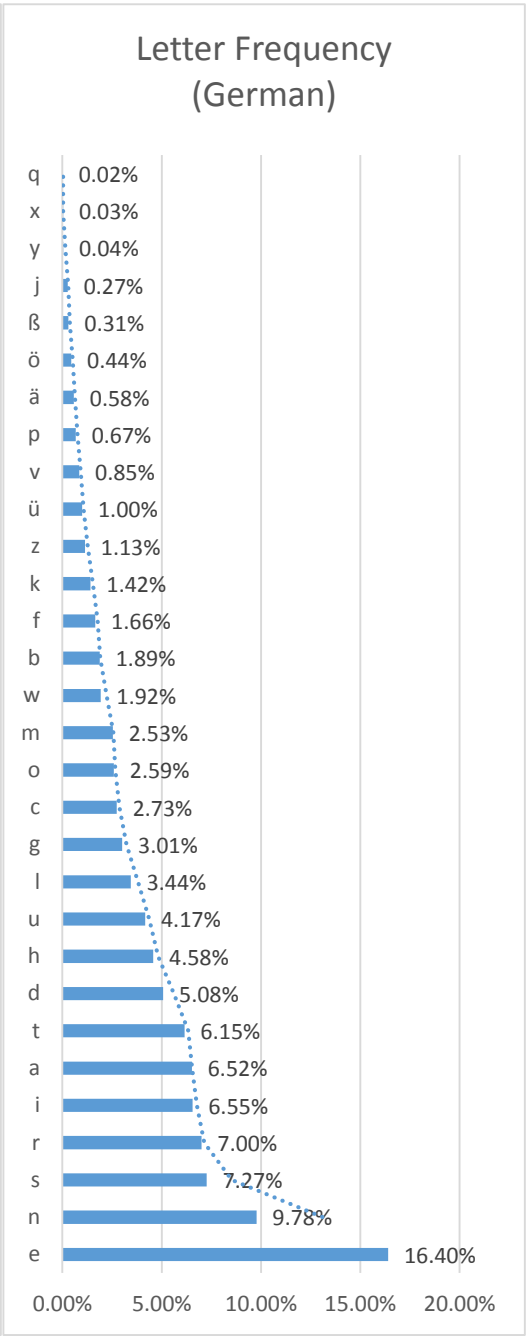


Fig 5. Letter Frequency of German [13]

Figure 6 shows the letter frequency of Greek. There are 24 words in total. The most frequency letters are A, E, O and I, and the least frequency letters are Ψ, Z, Ξ and B. Figure 7 shows the letter frequency of Spanish. There are 33 words in total. The most frequency letters are e, a, o and s, and the least frequency letters are k, ü, w and ú.

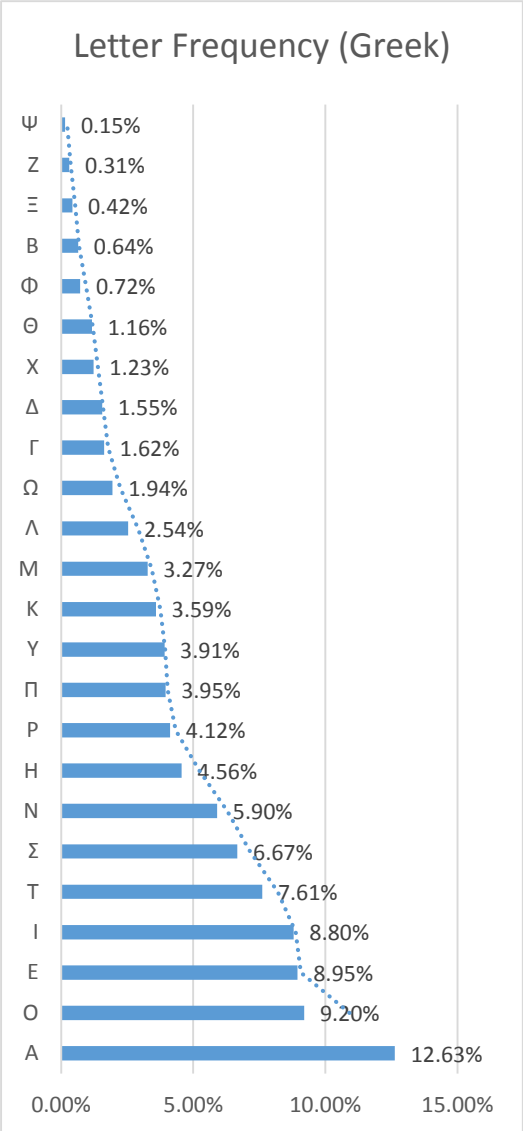


Fig 6. Letter Frequency of Greek [14]

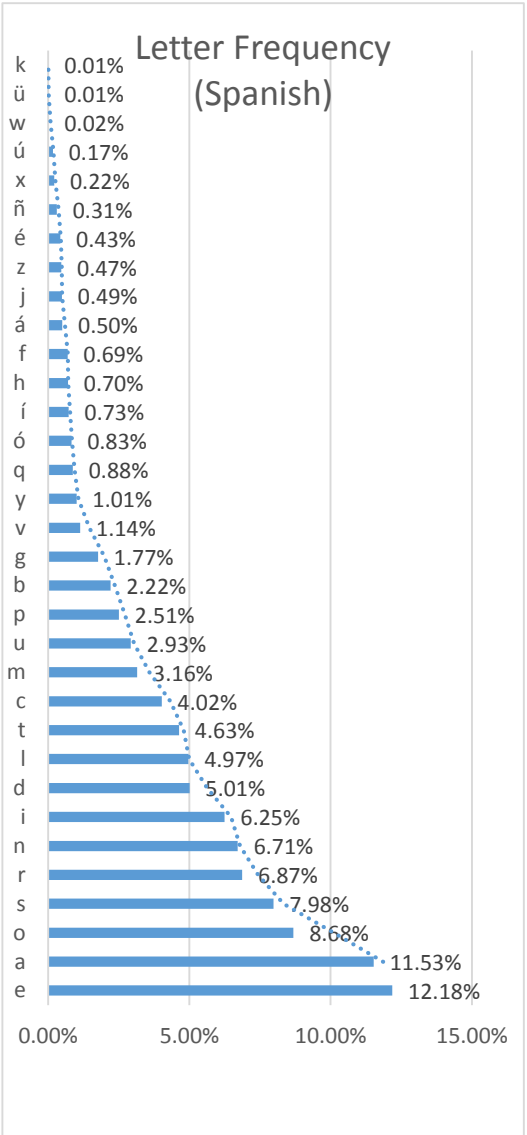


Fig 7. Letter Frequency of Spanish [15]

With the Matlab, correlations between the tendency of letter frequency of the Voynich manuscript and English, Latin, French, German, Greek and Spanish. The correlation between the Voynich manuscript and English is 98.04%. The correlation between the Voynich manuscript and Latin is 98.66%. The correlation between the Voynich manuscript and French is 94.55%. The correlation between the Voynich manuscript and German is 94.81%. The correlation between the Voynich manuscript

and Greek is 98.34%. The correlation between the Voynich manuscript and Spanish is 96.09%.

Comparing the Voynich manuscript with English, Latin, French, German, Greek and Spanish, the letter number of these languages shows that the most possible language is Greek, because they both have 24 letters. Furthermore, the letter frequency is also similar for the Voynich manuscript and Greek. In addition, the correlation between the Voynich manuscript and Greek is high. Therefore, Greek can be considered as a possible language that the Voynich manuscript used. However, this is not a strong evidence that can prove the Voynich manuscript is written in Greek. In conclusion, there is no specific evidence can prove that Voynich manuscript is one of these six kind of language, Greek is one of the possible language that the Voynich manuscript used.

6.2 Word Frequency

Figure 8 shows the word frequency in the Voynich manuscript. There are 37104 words in the whole manuscript, and the total unique words are 8486. Furthermore, there are 2472 words that appears more than once, and 6014 words appears only once. 515 words appears more than 10 times and these words counts 65.66% of the total words in the Voynich manuscript.

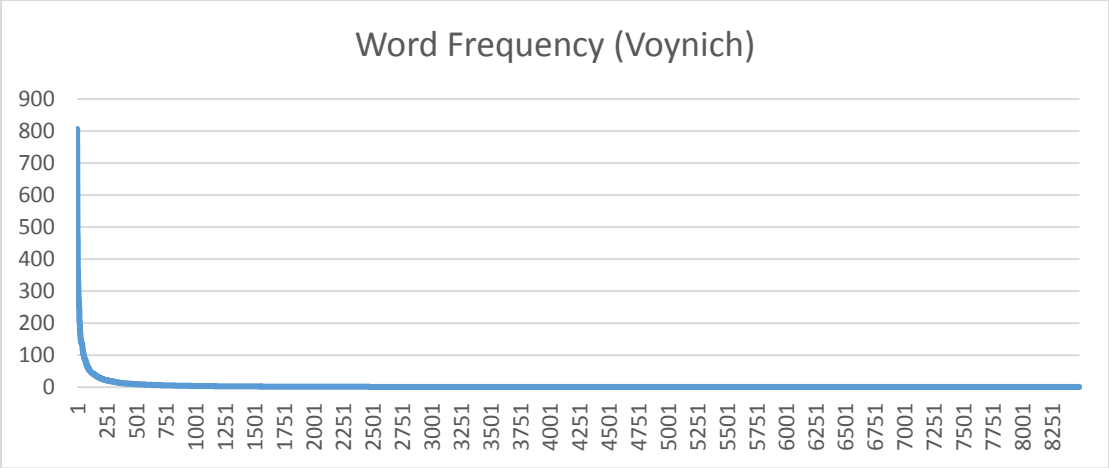


Fig 8. Word frequency in the Voynich manuscript

In figure 9, 50 most frequency words are token in order to make a comparison with English.

Comparing word frequency in the Voynich manuscript and in English, the correlation between the tendencies of both curve is 93.65%, which shows that there may exist relationship between the Voynich manuscript and English. In conclusion, there is no strong evidence shows that there is any relationship between the Voynich manuscript and English.

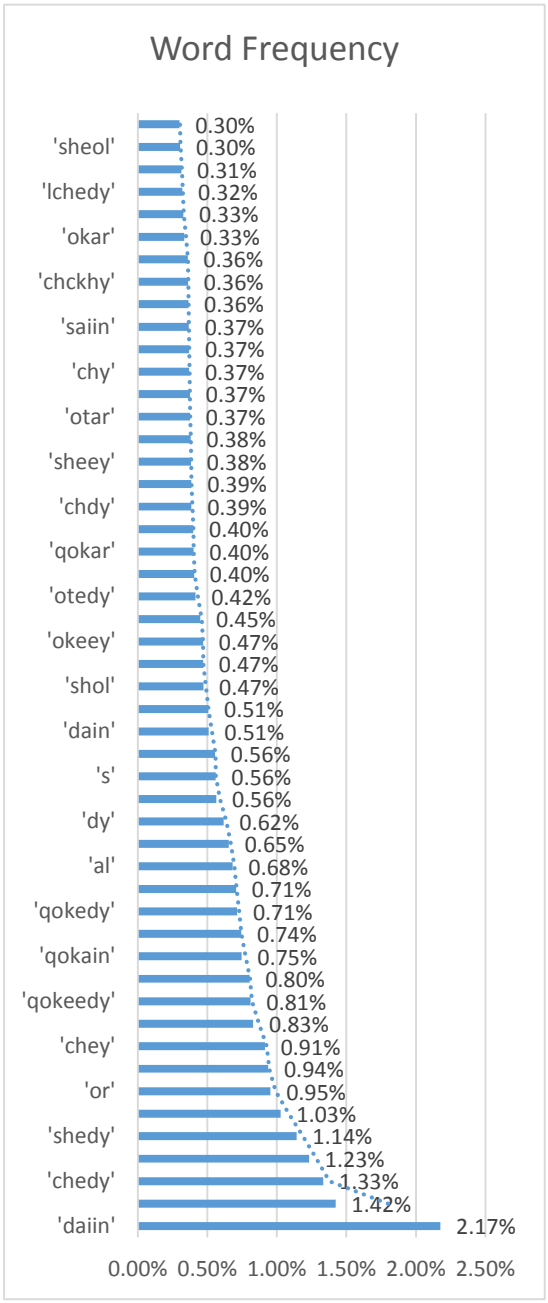


Fig 9. Word frequency in Voynich

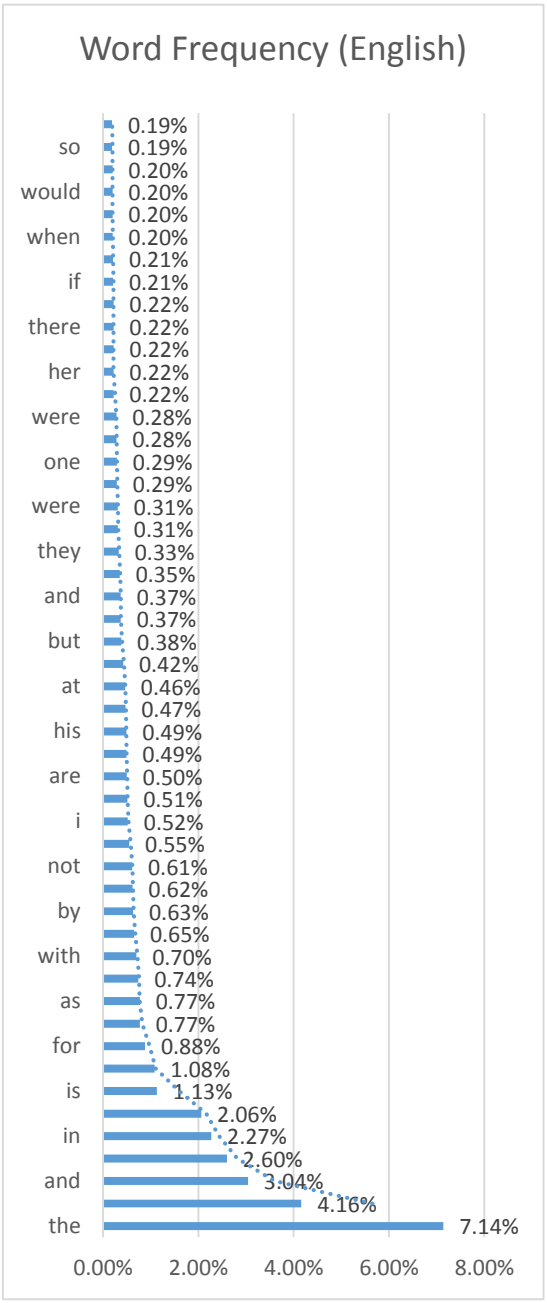


Fig 10. Word frequency in English [10]

6.3 Specific Pattern Words

This section shows gives a brief analysis of current results of the specific pattern words. The first numeral language is Roman numerals. In Roman numeral, VII stands for 7 and VIII stands for 8. These two numerals have obvious patterns that are easy to search in the Voynich manuscript.

Words follow VII pattern and VIII pattern have been found, and next step will continue finding all possible numerical words in Roman numerals from I to XX, and several obvious pattern numerals such as XX, XXX, C, CC and CCC.

aii	dee	kee	lee	oee	oii	qee	qoo	ree	see	tee	yee
534	1087	203	26847	949	101	14777	725	13401	555	1234	56
1158	1138	204	29041	994	1837	15515	1054	31974	5212	2886	285
1459	2242	745	31389	1015	2183	15653	4176		10905	3327	1183
1583	3163	810	31760	1064	3198	15995	4264		10983	3969	
2145	3818	1090		1132	3495	15996	5460		11561	4270	
2170	3929	1166		1812	3529	16394	7334		11744	6855	
2243	3937	2431		1984	5777	19498	9135		13298	7967	
2465	4172	2551		2241	9108	21304	11560		23093	10843	
2979	4408	3058		3036	10942	29882	13100		25972	11708	
3889	4597	3146		3137	11091	30017	14725		26713	11859	
4218	9336	4011		3565	11343	31171	14940			12878	
4240	10908	4040		3604	13215	31476	21881			12899	
4526	10975	4386		3616	13262	31880	22243			13280	
4534	15292	4390		3704	13665	32402	22770			13376	
4564	15313	4399		4715	14489	33196	23232			15383	
4623	16286	4434		8306	14694	33217	23583			15512	
4667	18941	4459		9127	15520	33412	25348			15803	
4691	21329	4523		10110	15775	34419	25663			15935	
4705	23172	4552		11165	16078		25772			16739	
4717	23964	4599		11275	16794		25849			16844	
4920	25239	4642		11373	17392		25944			17319	
4924	25745	4682		11441	17417		25970			17343	
4928	26151	5008		12163	17539		26284			17475	
4945	28473	5038		13245	18785		26347			19179	
4947	29896	5224		13291	18826		26361			20302	
4958	31523	5246		13448	18834		26637			20356	
5015	34861	5786		13508	23251		35440			20481	
5019		5804		13509	23634		35556			22366	
5026		6170		13520	24759		35782			22420	
5069		6385		13557	24994		36091			23886	

Fig 11. All locations of VII pattern words

Figure 11 shows part of Vii pattern words. All the numbers below the words are locations of these words, such as 534 means that 534th word in the Voynich manuscript contains aii. Since a large number of locations of several words were found, this figure could not show all the locations. From the locations, there are 562 aii, 201 kee, 77 oee, 72 tee, 51 oii, 30 qoo, 27 dee, 18 qee, 10 see, 4 lee, 3 yee and 2

ree were found. All along with the result, it is obviously that *ii, *ee and *oo are three patterns that may be numerical words for VII.

Figure 12 shows part of Vii pattern words. From the locations, there are 44 aiii, 25 oeee, 22 keee, 72 tee, 11 oiii, 7 deee, 6 qeee, 5 teee, 3 seee, 2 leee , 2 reee and 1 yeee were found. All along with the result, it is obviously that *iii and *eee are two patterns that may be numerical words for VIII.

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										

Fig 12. All locations of VIII pattern words

All along with the result, it is obviously that *ii (*iii) and *ee (*eee) are two patterns that may be numerical words for VIII.

Comparing all possible VII words and VIII words, e, I and o can be considered as possible numerical characters.

6.4 Statistical Comparison of Letters and Words

This section gives a brief statistical comparisons between the Voynich manuscript and three book in English, French and German. Among these languages, the percentage of unique words/total words, word length and the percentage of words appear more than once /total unique words were compared.

Figure 13 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%).

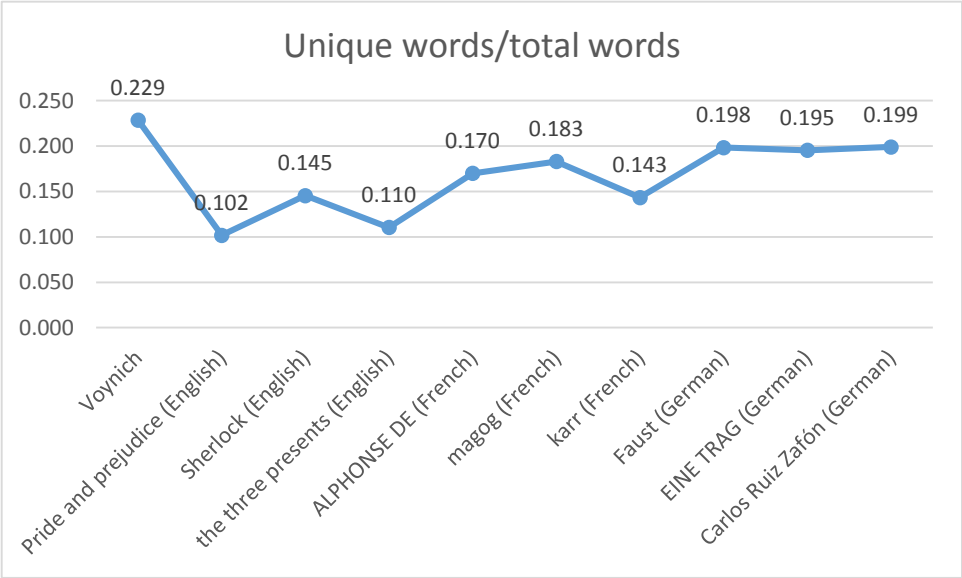


Fig 13. Unique words/total words

Figure 14 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%).

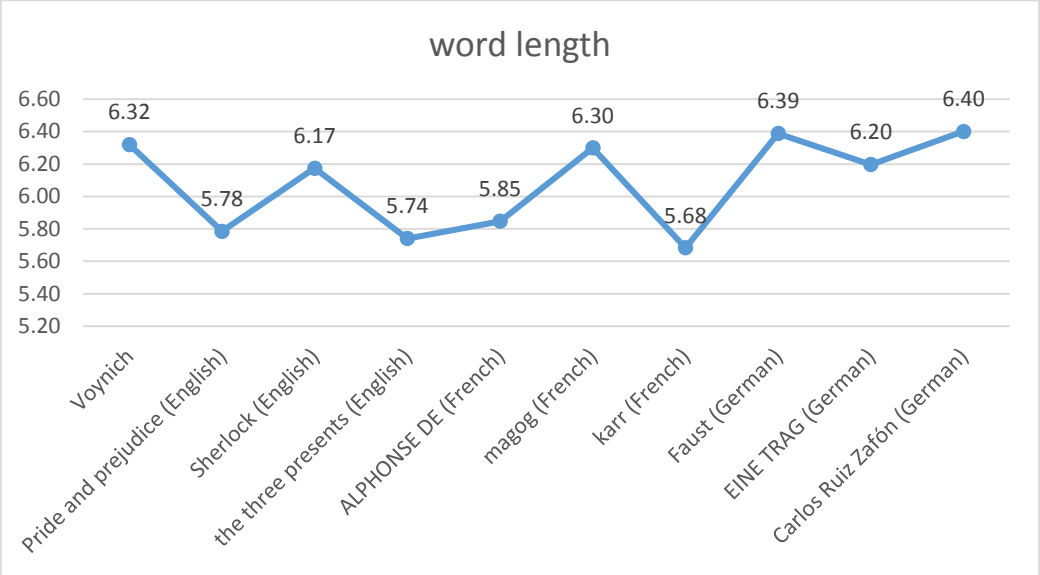


Fig 14. Word length

Figure 15 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.

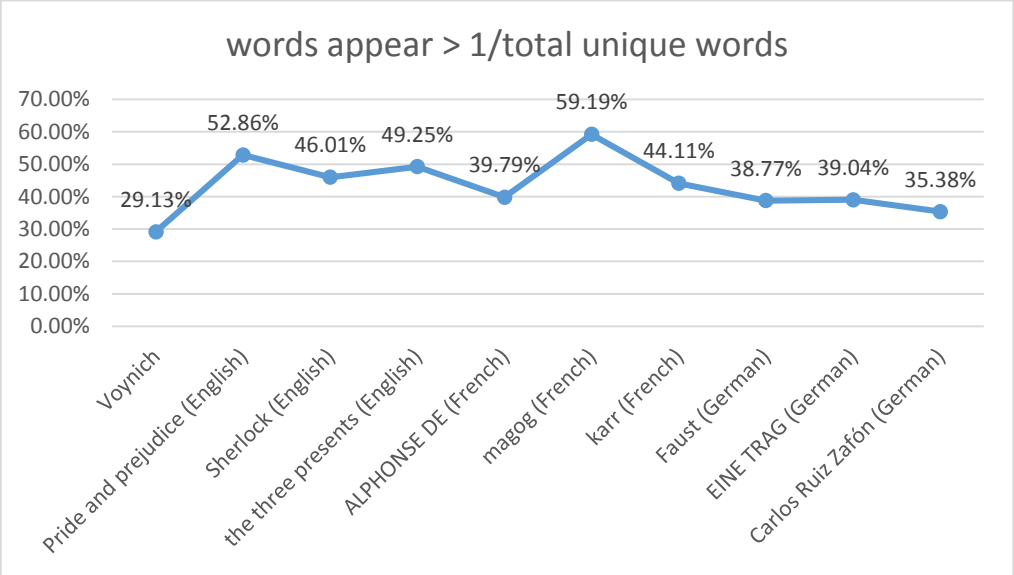


Fig 15. Words appear more than once/total unique words

Among these statistical comparisons, German can be considered as a possible language that the Voynich manuscript used.

7. Project Management

7.1 Time Management

As shown in figure 16, time management is divided into five parts. Background research should be taken from week 1 to week 5 in semester 1. After that, text analysis will be taken between week6 and week 7. Then illustration investigation should begin from week 8. Following should be marginal symbol research which will begin from week 10. The last task will be translation of any possible test from week 5 to week 9 in semester 2.

No.	Task	Week
	Semester 1	
1	Background research	1
2	Text analysis	6
3	Illustration investigation	8
4	Marginal symbol research	10
	Semester 2	
5	Translation	5-9

Figure 16. Time management

7.2 Risk Management

As shown in figure 17, six risks have been identified. There important risk that may cause important impact also may occur in a significant probability should be mismanagement of time, lack of references and health issues. Each of them may cause significant impact to the final result.

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
6	Supervisor unavailable	Low	Moderate

Figure 17. Risk management

7.3 Task Allocation

As shown in figure 18, the task management is associated to the time management. Except the illustration investigation will be carried out by Yaxin Hu, marginal symbol research will be carried out by Ruihang Feng, other tasks will be done by both of us.

No.	Task	Student
	Semester 1	
1	Background research	Cooperation
2	Text analysis	Cooperation
3	Illustration investigation	Yaxin Hu
4	Marginal symbol research	Ruihang Feng
	Semester 2	
5	Translation	Cooperation

Figure 18. Task allocation

7.4 Budget

a) 500 AUS dollars for each member.

b) Research need to be carried out further research.

c) All program that need to be used are available on University system.

d) Major work are based on computer.

All the software that be used in this projects are Matlab, Visual Studio, Office and Python. Therefore, there is no expenditure.

7.5 Management Strategy

Meetings will be the main way to exchange project progress between project members and supervisors. At least one meeting should be held between project members each week and a minimum of one meeting should be held between project members and supervisors each three weeks.

These meetings should involve phase progress, issues occurs, further ideas that could help processing the project and any possible results.

8. Conclusion

With a literature review of the Voynich manuscript, the background and proposal methods have been settled down. As discussed in section 4, digital investigation will be the main breakpoint in the whole project. All possible methods will be carried out to determine any possible features of the Voynich manuscript.

In section 6, there are four parts about the present outcomes. In section 6.1, by comparing letter frequency of the Voynich manuscript with English, Latin, French, German, Greek and Spanish, Greek is more possible to be the language the Voynich manuscript used. In section 6.2, by comparing word frequency of the Voynich manuscript with English, there may exist possible relationship between the Voynich manuscript and English. In section 6.3, by locating all possible VII words and VIII words, e, I and o can be considered as possible numerical characters.

There are three statistical comparisons between the Voynich manuscript and three book in English, French and German in section 6.4. Among these languages, the percentage of unique words/total words, word length and the percentage of words appear more than once /total unique words were compared. By these comparisons, German is a possible language that the Voynich manuscript used.

As the plan shows that the text analysis and the illustration investigation should be done during this semester, however, the steps taken by project members is a little bit slower than the time management. There is more work need to be done than expected, therefore, time planning will be delay two or three weeks. All the things that have been done would contribute significant influence in the whole project.

9. Reference

[1] Schmeh, Klaus (January–February 2011). "The Voynich Manuscript: The Book Nobody Can Read". *Skeptical Inquirer.* Retrieved 2013-09-05.

[2] Shailor, Barbara A.,Beinecke MS 408, Yale University, Beinecke Rare Book and Manuscript Library, General Collection of Rare Books and Manuscripts, Medieval and Renaissance Manuscripts, accessed 24 June 2013.

[3] "*Data Mining Curriculum*". ACM SIGKDD. 2006-04-30. Retrieved 2014-01-27.

[4] Barabe, Joseph G. (McCrone Associates) (April 1, 2009). "Materials analysis of the Voynich Manuscript". Beinecke Library.

[5] G. Landini, "Evidence of Linguistic Structure in the Voynich Manuscript Using Spectral Analysis," *Cryptologia*, pp. 275-295, 2001.

[6] B. Stephen, *"A proposed partial decoding of the Voynich script",* www.stephenbax.net, Version 1, January 2014.

[7] G. Landini, "Evidence of Linguistic Structure in the Voynich Manuscript Using Spectral Analysis," *Cryptologia*, pp. 275-295, 2001.

[8] B. Shi and P. Roush, "Semester B Final Report 2014 - Cracking the Voynich code," University of Adelaide, Adelaide, 2014.

[9] D. R. Amancio, E. G. Altmann, D. Rybski, O. N. Oliveira Jr. and L. d. F. Costa, "Probing the Statistical Properties of Unknown Texts: Application to the Voynich Manuscript," PLoS ONE 8(7), vol. 8, no. 7, pp. 1-10, 2013.

[10] Mayzner. M, "English Letter Frequency Counts", Retrieved: 17 December 2012, available at http://norvig.com/mayzner.html

[11] Stefan Trost Media, "Character Frequency: Latin (Latina)", available at http://www.sttmedia.com/characterfrequency-latin

[12] Beutelspacher, Albrecht (2005). *Kryptologie* (7th Ed.). Wiesbaden: Vieweg. p. 10. ISBN 3-8348-0014-7.

[13] Pratt, Fletcher (1942). *Secret and Urgent: the Story of Codes and Ciphers.* Garden City, N.Y.: Blue Ribbon Books. pp. 254–5. OCLC 795065.

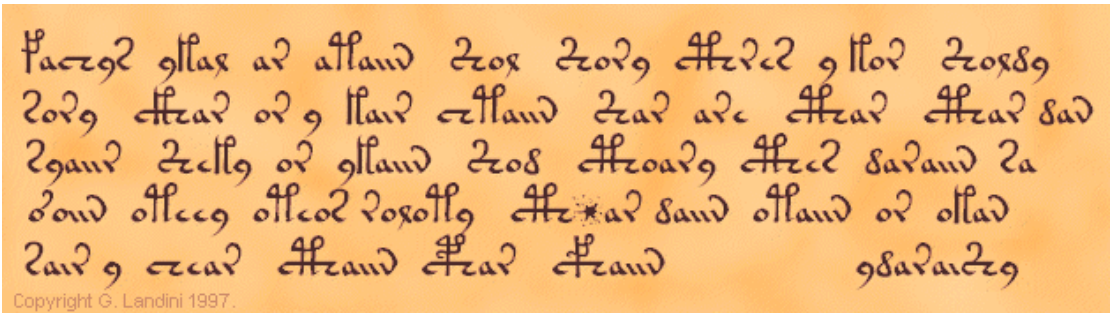
[14] Stefan Trost Media, “Character Frequency: Greek”, available at http://www.sttmedia.com/characterfrequency-greek

[15] Pratt, Fletcher (1942). *Secret and Urgent: the Story of Codes and Ciphers*. Garden City, N.Y.: Blue Ribbon Books. pp. 254–5. OCLC 795065.

Appendix

Appendix 1. Takahashi transcription

An original part of the Voynich manuscript:



Takahashi transcription of this part of the Voynich manuscript:

```
fachys ykal ar ataiin Shol Shory cThres y kor Sholdy
sory cThar or y kair chtaiin Shar are cThar cThar dan
syaiir Sheky or ykaiin Shod cThoary cThes daraiin sa
o'oiin oteey oteor roloty cTh*ar daiin otaiin or okan
sair y chear cThaiin cPhar cFhaiin    ydaraiShy
```

Appendix 2. Excel of Letter Frequency in Section 6.1

Here is the Excel of letter frequency for the Voynich manuscript, English, Latin, Greek, French, German and Spanish.

Voynich		Latin		English		Greek	
Letter	Frequency	Letter	Frequency	Letter	Frequency	Letter	Frequency
o	13.2766%	i	11.44%	e	12.49%	A	12.63%
e	10.4626%	e	11.38%	t	9.28%	O	9.20%
h	9.3084%	a	8.89%	a	8.04%	E	8.95%
y	9.2037%	u	8.46%	o	7.64%	I	8.80%
a	7.4448%	t	8.00%	i	7.57%	T	7.61%
c	6.9407%	s	7.60%	n	7.23%	Σ	6.67%
d	6.7629%	r	6.67%	s	6.51%	N	5.90%
i	6.1160%	n	6.28%	r	6.28%	H	4.56%

k	5.7000%	o	5.40%	h	5.05%	Ρ	4.12%
l	5.4831%	m	5.38%	l	4.07%	Π	3.95%
r	3.8869%	c	3.99%	d	3.82%	Υ	3.91%
s	3.8509%	l	3.15%	c	3.34%	Κ	3.59%
t	3.6199%	o	3.03%	u	2.73%	Μ	3.27%
n	3.2013%	d	2.77%	m	2.51%	Λ	2.54%
q	2.8270%	b	1.58%	f	2.40%	Ω	1.94%
p	0.8497%	q	1.51%	p	2.14%	Γ	1.62%
m	0.5818%	g	1.21%	g	1.87%	Δ	1.55%
f	0.2633%	v	0.96%	w	1.68%	Χ	1.23%
*	0.1460%	f	0.93%	y	1.68%	Θ	1.16%
g	0.0500%	h	0.69%	b	1.48%	Φ	0.72%
x	0.0182%	x	0.60%	v	1.05%	Β	0.64%
v	0.0047%	y	0.07%	k	0.54%	Ξ	0.42%
z	0.0010%	z	0.01%	x	0.23%	Ζ	0.31%
S	0.0005%			j	0.16%	Ψ	0.15%
				q	0.12%		
				z	0.09%		

French		German		Spanish	
Letter	Frequency	Letter	Frequency	Letter	Frequency
e	14.72%	**e**	16.40%	**e**	12.18%
s	7.95%	**n**	9.78%	**a**	11.53%
a	7.64%	**s**	7.27%	**o**	8.68%
i	7.53%	**r**	7.00%	**s**	7.98%
t	7.24%	**i**	6.55%	**r**	6.87%
n	7.10%	**a**	6.52%	**n**	6.71%
r	6.69%	**t**	6.15%	**i**	6.25%
u	6.31%	**d**	5.08%	**d**	5.01%
o	5.80%	**h**	4.58%	**l**	4.97%
l	5.46%	**u**	4.17%	**t**	4.63%
d	3.67%	**l**	3.44%	**c**	4.02%
c	3.26%	**g**	3.01%	**m**	3.16%
m	2.97%	**c**	2.73%	**u**	2.93%
p	2.52%	**o**	2.59%	**p**	2.51%
v	1.84%	**m**	2.53%	**b**	2.22%
é	1.50%	**w**	1.92%	**g**	1.77%
q	1.36%	**b**	1.89%	**v**	1.14%
f	1.07%	**f**	1.66%	**y**	1.01%
b	0.90%	**k**	1.42%	**q**	0.88%
g	0.87%	**z**	1.13%	**ó**	0.83%
h	0.74%	**ü**	1.00%	**í**	0.73%
j	0.61%	**v**	0.85%	**h**	0.70%
à	0.49%	**p**	0.67%	**f**	0.69%
x	0.43%	**ä**	0.58%	**á**	0.50%
z	0.33%	**ö**	0.44%	**j**	0.49%
è	0.27%	**ß**	0.31%	**z**	0.47%

ê	0.22%	**j**	0.27%	**é**	0.43%
y	0.13%	**y**	0.04%	**ñ**	0.31%
ç	0.09%	**x**	0.03%	**x**	0.22%
w	0.07%	**q**	0.02%	**ú**	0.17%
ù	0.06%			**w**	0.02%
â	0.05%			**ü**	0.01%
k	0.05%			**k**	0.01%
î	0.05%				
ô	0.02%				
œ	0.02%				
ë	0.01%				
ï	0.01%				

Appendix 3. Excel of Word Frequency in Section 6.2

Here is the Excel of word frequency for the Voynich manuscript (top 50 words).

words	frequency	percent
'daiin'	807	2.1750%
'ol'	528	1.4230%
'chedy'	495	1.3341%
'aiin'	457	1.2317%
'shedy'	424	1.1427%
'chol'	381	1.0268%
'or'	354	0.9541%
'ar'	348	0.9379%
'chey'	339	0.9136%
'qokeey'	308	0.8301%
'qokeedy'	301	0.8112%
'dar'	298	0.8031%

'qokain'	277	0.7466%
'shey'	276	0.7439%
'qokedy'	265	0.7142%
'qokaiin'	262	0.7061%
'al'	253	0.6819%
'dal'	243	0.6549%
'dy'	229	0.6172%
'okaiin'	209	0.5633%
's'	208	0.5606%
'chor'	206	0.5552%
'dain'	189	0.5094%
'qokal'	188	0.5067%
'shol'	175	0.4716%
'cheey'	174	0.4690%
'okeey'	174	0.4690%
'cheol'	167	0.4501%
'otedy'	154	0.4150%
'otaiin'	150	0.4043%
'qokar'	149	0.4016%
'qol'	148	0.3989%
'chdy'	143	0.3854%
'y'	143	0.3854%
'sheey'	142	0.3827%
'okain'	141	0.3800%
'otar'	139	0.3746%
'qoky'	139	0.3746%
'chy'	137	0.3692%
'otal'	137	0.3692%
'saiin'	136	0.3665%

'oteey'	135	0.3638%
'chckhy'	134	0.3611%
'okal'	133	0.3585%
'okar'	124	0.3342%
'sho'	121	0.3261%
'lchedy'	119	0.3207%
'okedy'	116	0.3126%
'sheol'	113	0.3045%
'dol'	111	0.2992%

Appendix 4. Statistical Comparison for Section 6.4

Here is the Excel of statistical comparison for the Voynich manuscript and three books in English, French and German.

	A	B	C	D	E	F	G	H	I
1	book name	total characters number	total words number	unique words number	ratio(UWN/ TWN)	characters per word	Language	words appear > once	percent words appear > once/total unique words
2	Voynich	234507	37104	8486	0.229	6.32		2472	29.13%
3	Pride and prejudice (English)	210748	36433	3706	0.102	5.78	English	1959	52.86%
4	Sherlock (English)	229037	37095	5397	0.145	6.17	English	2483	46.01%
5	the three presents (English)	211675	36874	4073	0.110	5.74	English	2006	49.25%
9	ALPHONSE DE (French)	219073	37467	6366	0.170	5.85	French	2533	39.79%
10	magog (French)	235942	37457	6859	0.183	6.30	French	4060	59.19%
11	karr (French)	213019	37484	5371	0.143	5.68	French	2369	44.11%
12	Faust (German)	195835	30654	6079	0.198	6.39	German	2357	38.77%
13	EINE TRAG (German)	229801	37082	7242	0.195	6.20	German	2827	39.04%
14	Carlos Ruiz Zafón (German)	238584	37270	7423	0.199	6.40	German	2626	35.38%

Appendix 5. Thesis Plan