2015 Honours Project Cracking the Voynich Code

Honours Project ID: 31

Supervisors: Derek Abbott, Brian Ng and Maryam Ebrahimpour

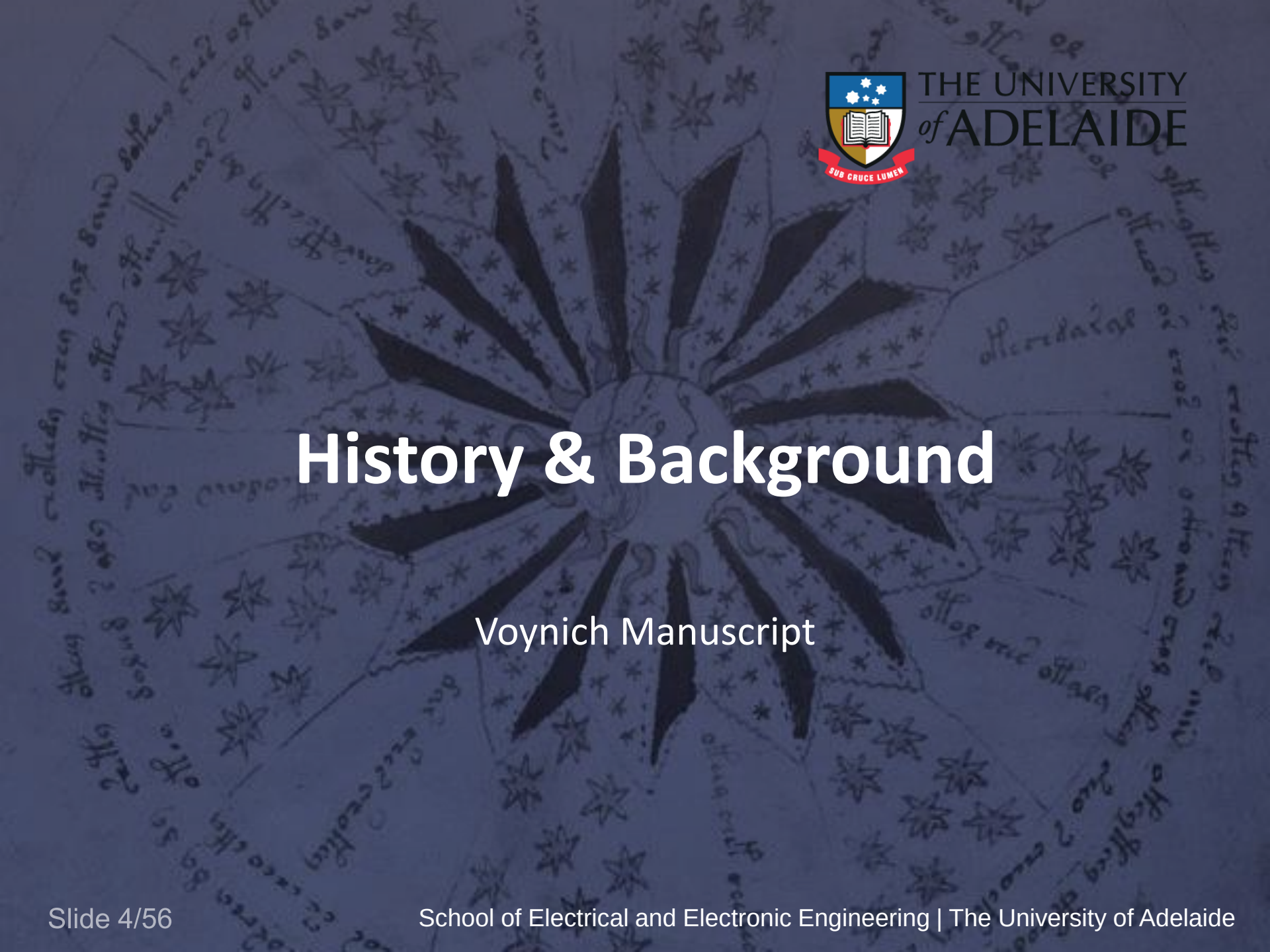
Students: Andrew McInnes and Lifei Wang

Voynich Manuscript

Reproduced from Internet Archive https://archive.org/details/TheVoynichManuscript

Outline

- History, background and objective
  - Background, nature, scopes and aims
- Detail of Voynich manuscript
  - Layout, content and Interlinear Archive
- Tests and Results
  - Characterisation of Text
  - Illustration Investigation
  - English Investigation
  - Morphology Investigation
  - Collocation Investigation
- Project management
  - Risk, work breakdown and budget
- Conclusion
  - Concluding remarks on results
  - Future Pathways

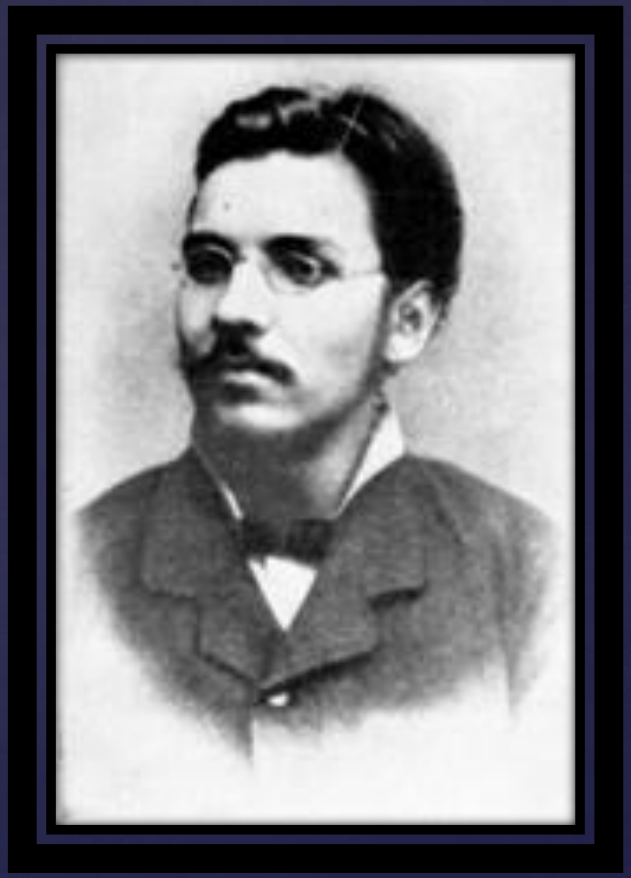


History & Background

Voynich Manuscript

Wilfred Voynich

- Born in Telsze, a town in Russian Empire on 31 October 1865
- Revolutionary, Antiquarian and Book Dealer
- He found the manuscript in an ancient castle in Southern Europe in 1912
- He took the Voynich Manuscript to London in 1912, and later in 1914 to the United States.

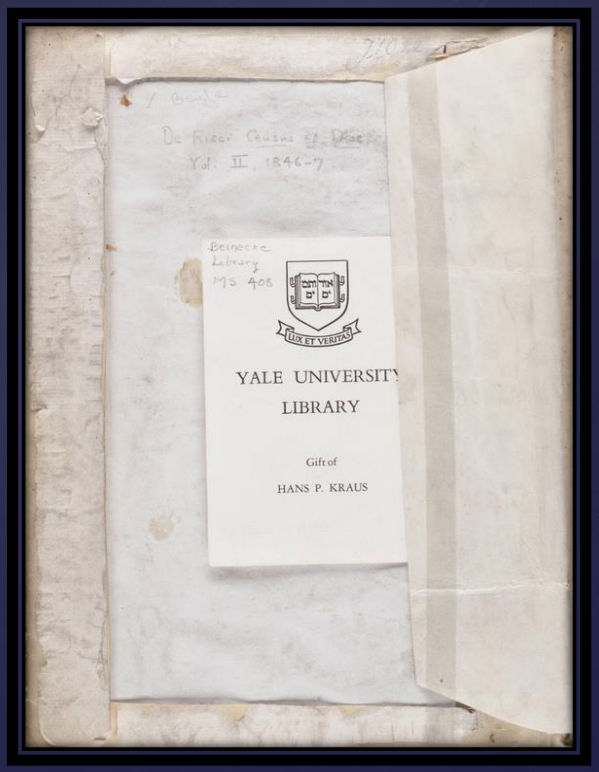


Rene Zandbergen. (2014). The Voynich MS - General Introduction. Accessed on 22/03/15 from http://www.voynich.nu/intro.html

Beinecke Rare Book and Manuscript Library

- Hans P. Kraus bought the Voynich Manuscript in 1961
- It was donated to Yale University in 1965.
- Housed at the Beinecke Rare Book and Manuscript Library
- Official register number MS 408.

Beinecke Rare Book and Manuscript Library



Label of Voynich manuscript in Yale University

Objective

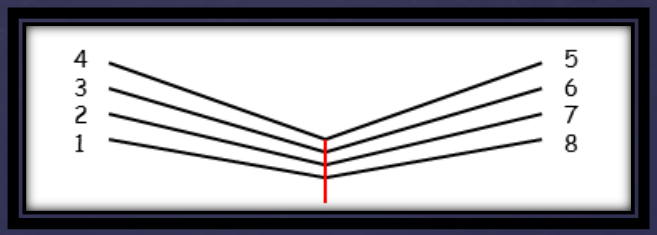
- Develop software program that search the text and perform statistical tests
- Separate the alphabet from other tokens of the Voynich manuscript.
- Compare linguistic features between the Voynich languages and other known languages.
- Investigate whether the language in the Voynich Manuscript is cipher, codes, natural language, constructed language or hoax.
- Investigate the possible relation between texts, words and illustrations.

Details

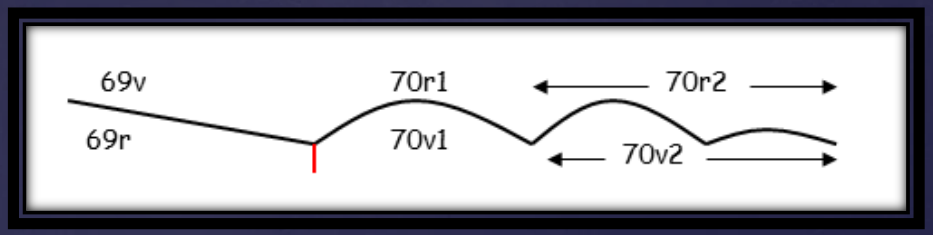
Voynich Manuscript

Layout of the manuscript

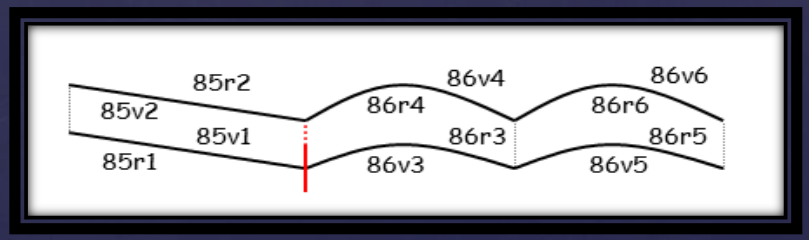
- The manuscript is made up many folios, numbered from f1 to f116
- Each folio consists two pages, labelled r and v.
- There are gaps between the folio number , which indicate missing folios.



Folio 1-66, 73-84, 108-116



Folio 69-72, 85-90, 93-94, 99-102



Folio 85-86

Folio 67-68

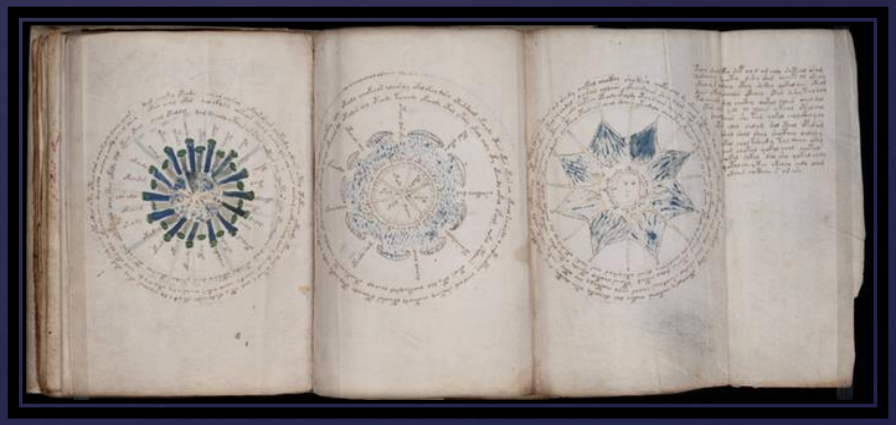
Rene Zandbergen. (2014). The Voynich MS - General Introduction. Accessed on 22/03/15 from http://www.voynich.nu/intro.html

Sections of the manuscript

- Herbal
  - Folios 1r - 66v
- Astronomical
  - Folios 67r - 73v
- Biological
  - Folios 75r - 84v
- Cosmological
  - Folios 85r – 86v
- Pharmaceutical
  - Folios 87r - 102v
- Recipes
  - Folios 103r - 116v

Herbal section

Recipes section

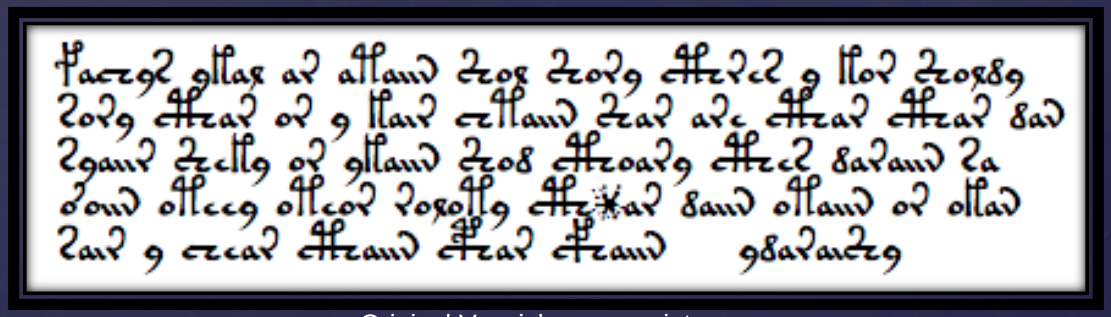


Astronomical section

Reproduced from Internet Archive https://archive.org/details/TheVoynichManuscript

Interlinear Archive

- A digital archive of transcriptions from the Voynich Manuscript
- There are 19 transcriptions from various transcribers
- This is an example to convert original Voynich manuscript into machine-readable European Voynich Alphabet (EVA)



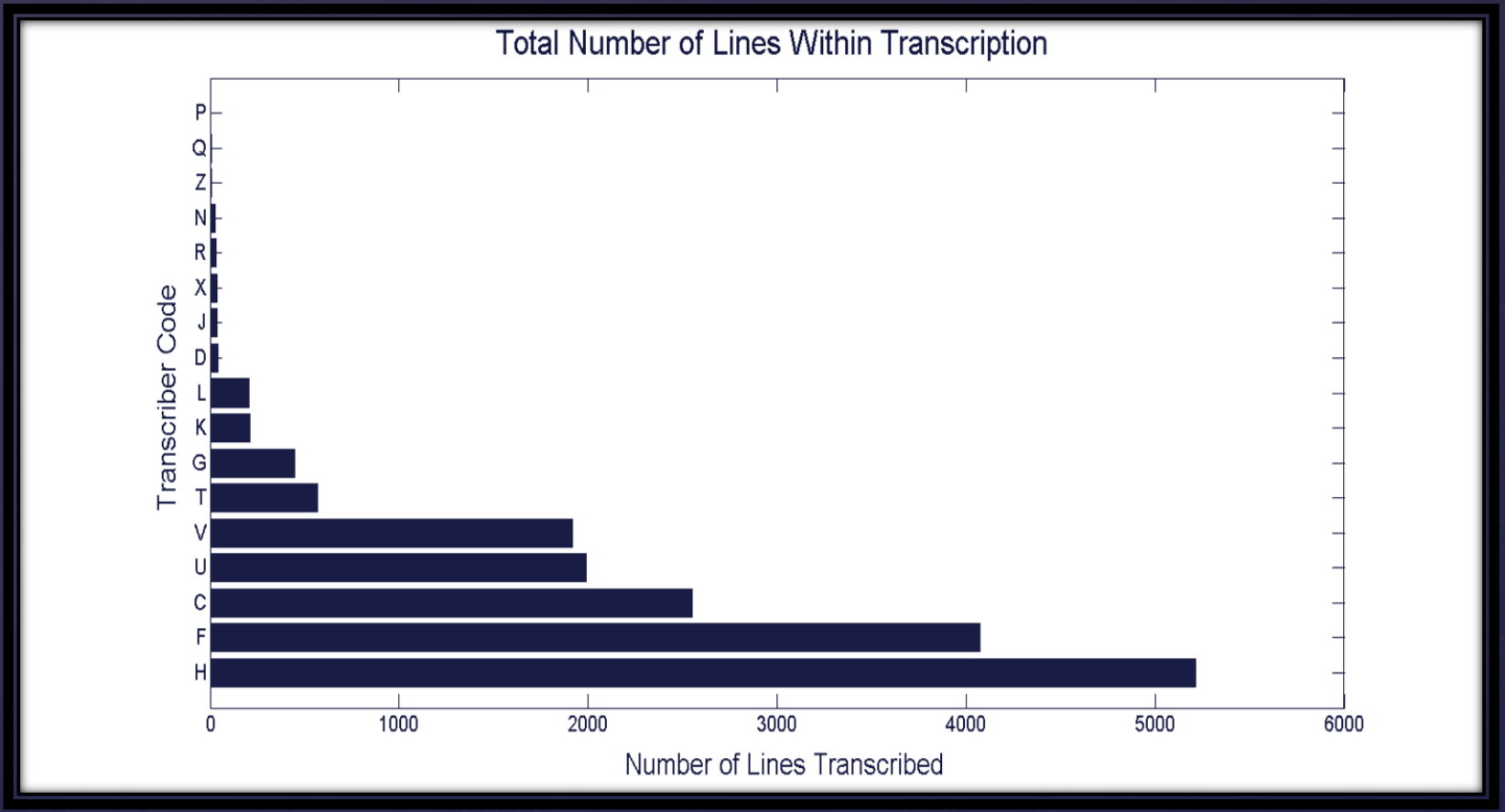
Original Voynich manuscript

```
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
```

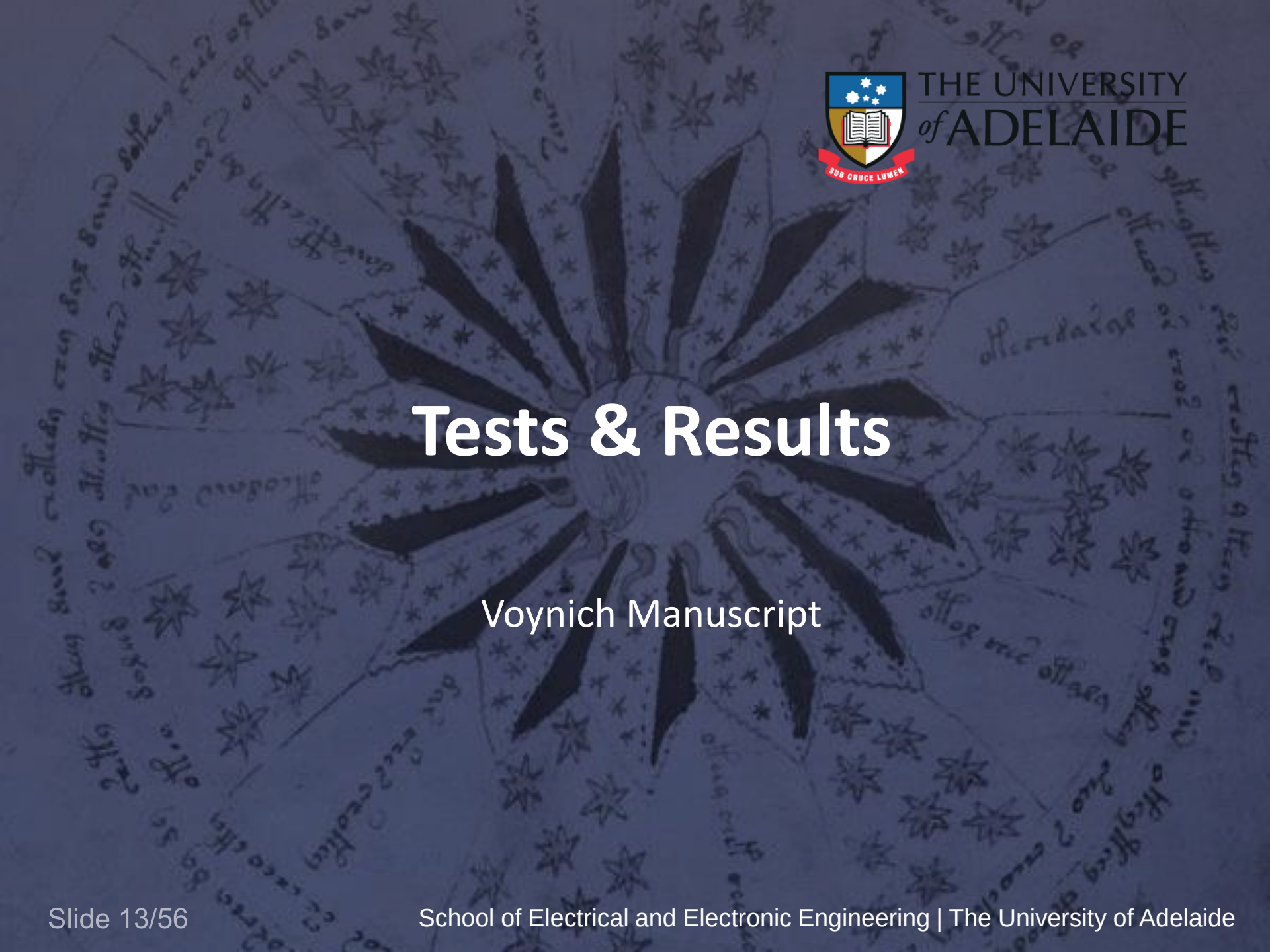
Convert using the EVA True Type font

Rene Zandbergen. (2014). The Voynich MS – Analysis of the text. Accessed on 22/03/15 http://www.voynich.nu/analysis.html

Most complete transcription



- Used Takahashi Transcription (H) for all of the test



Tests & Results

Voynich Manuscript

Characterization of text phase

- Write C++ and MATLAB codes to count the basic features of a given text as well as the Voynich:
  - Number of words
  - Number of characters
  - Frequency of specific words
  - Frequency of specific characters
  - Tokens that only appear at the start, middle or end
- Comparing the features of Voynich with known languages.

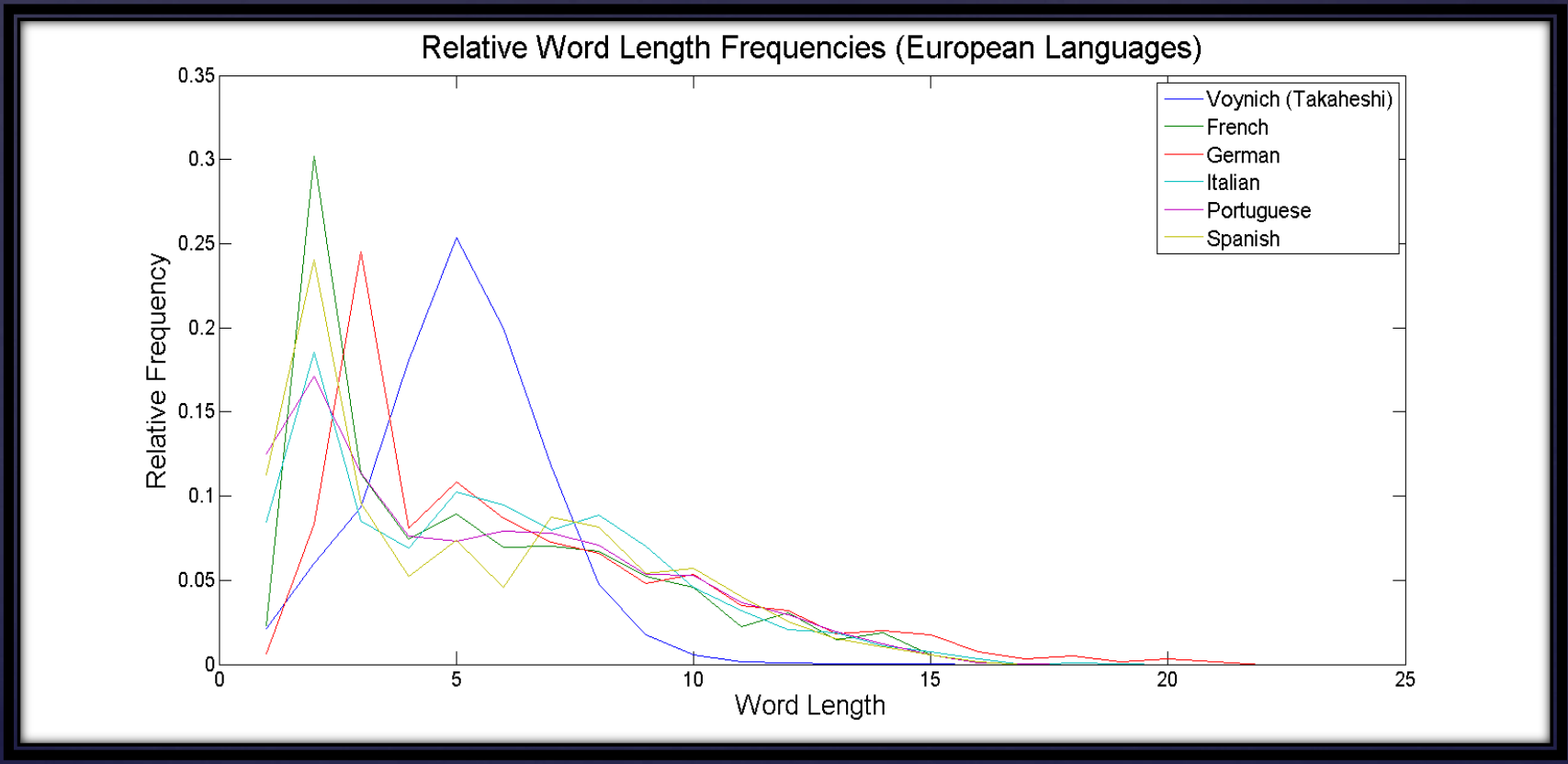
Takahashi transcription

Basic statistics of the full Takahashi transcription

Total Word Tokens	**37919**
Total Unique Word Tokens	**8151**
Total Character Tokens	**191825**
Total Unique Character Tokens	**23**
Longest Word Token	**15**
Shortest Word Token	**1**
Average Word Length	**5.0588**

Word length

- Compare word length between Voynich and other European languages



Single letter word tokens

- Some European languages use single letter to represent numbers
- Single letter word in some European languages

Language	Number of single letter word
Tamil	42
Hebrew	22
Greek	27
Voynich	20
French	3
English	2

Shortest word			
Word length	Word	Frequency	Rel. Frequency
1	*	31	0.0818%
1	a	3	0.0079%
1	c	7	0.0185%
1	d	50	0.1300%
1	e	3	0.0079%
1	f	8	0.0211%
1	g	11	0.0290%
1	k	13	0.0343%
1	l	58	0.1500%
1	m	11	0.0290%
1	n	4	0.0105%
1	o	81	0.2100%
1	p	5	0.0132%
1	q	1	0.0026%
1	r	98	0.2600%
1	s	243	0.6400%
1	t	8	0.0211%
1	v	7	0.0185%
1	x	9	0.0237%
1	y	151	0.4000%

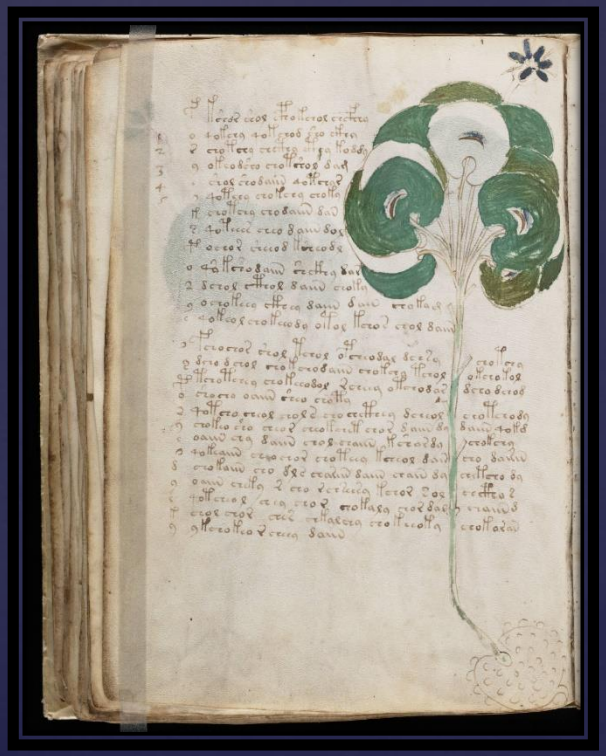
Single letter word in folios

- Left column of letters, folio 49v

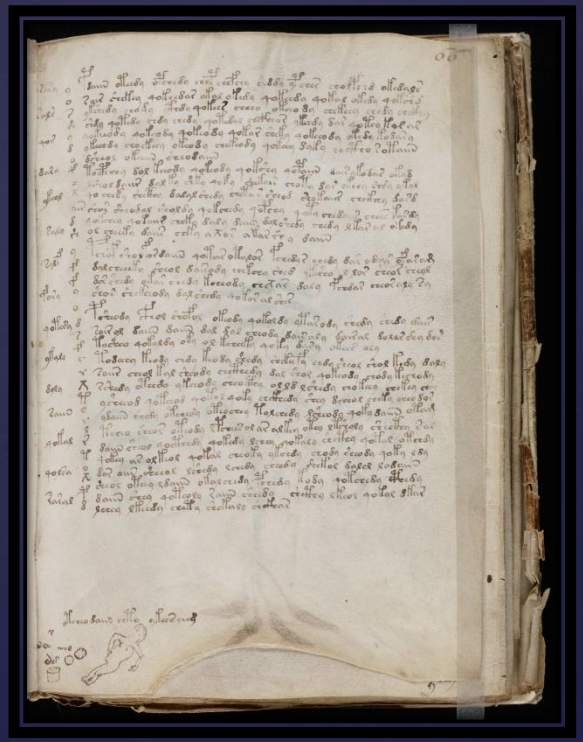
f o r y e * k s p o * y e * * p o * y e * d y s k y

- Left column of letters, folio 66r

y o s sh y d o f * x air d sh y f f y o d r f c r x t o * l r t o x p d



Folio 49v



Folio 66r

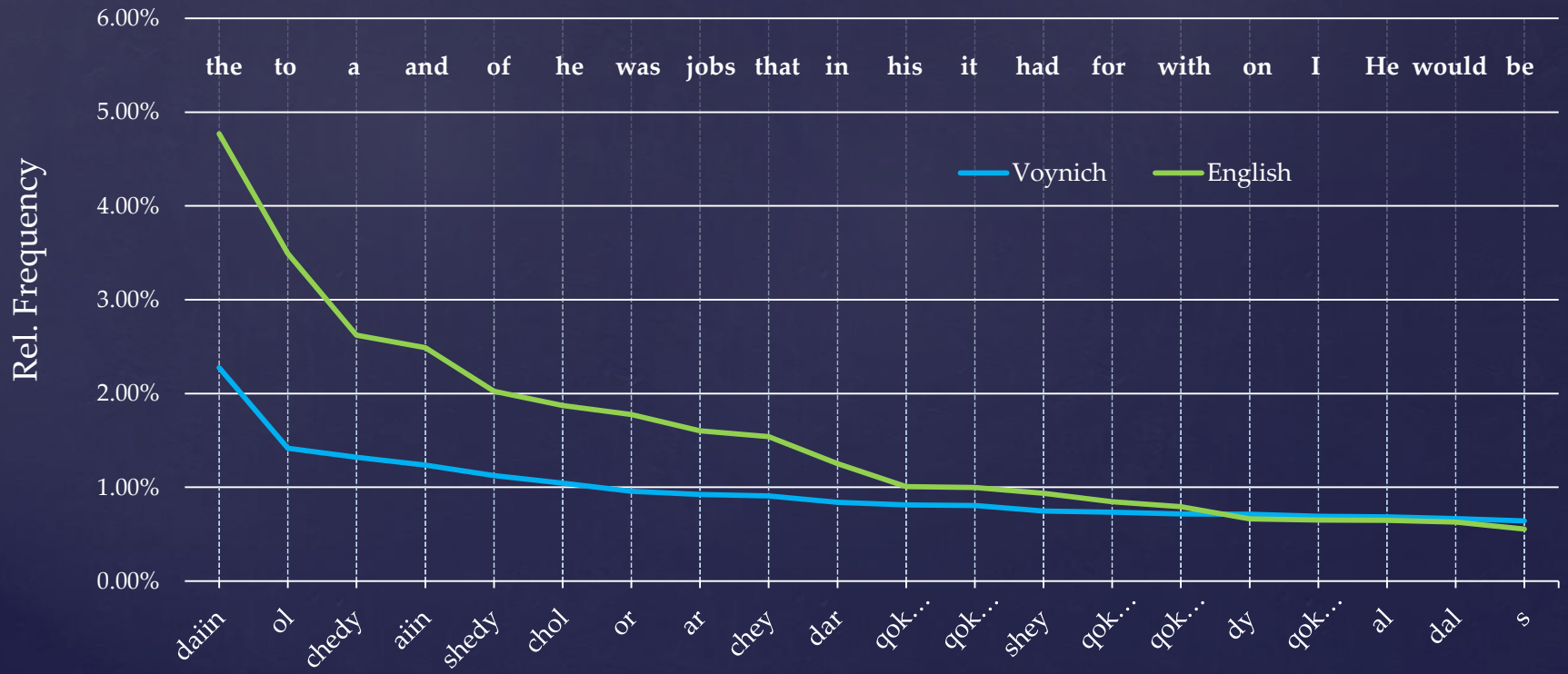
Tokens appear at the start and end

- Punctuation as symbols that occur only at word edges.
- They are only found at the ends of words.
- The table left shows that there is no tokens which only appear at the ends.
- Therefore, in traditional sense, there is likely no punctuation.

Takeshi Takahashi		
Token	Start	End
*	75	76
a	1959	91
c	6921	9
d	3671	635
e	143	92
f	126	25
g	16	84
h	2	69
i	15	6
k	1166	76
l	1371	5909
m	13	1061
n	4	6064
o	8530	1173
p	547	27
q	5389	1
r	507	5689
s	4552	1348
t	980	53
v	9	7
x	16	15
y	1906	15409
z	1	0

Most frequent word in Voynich & English

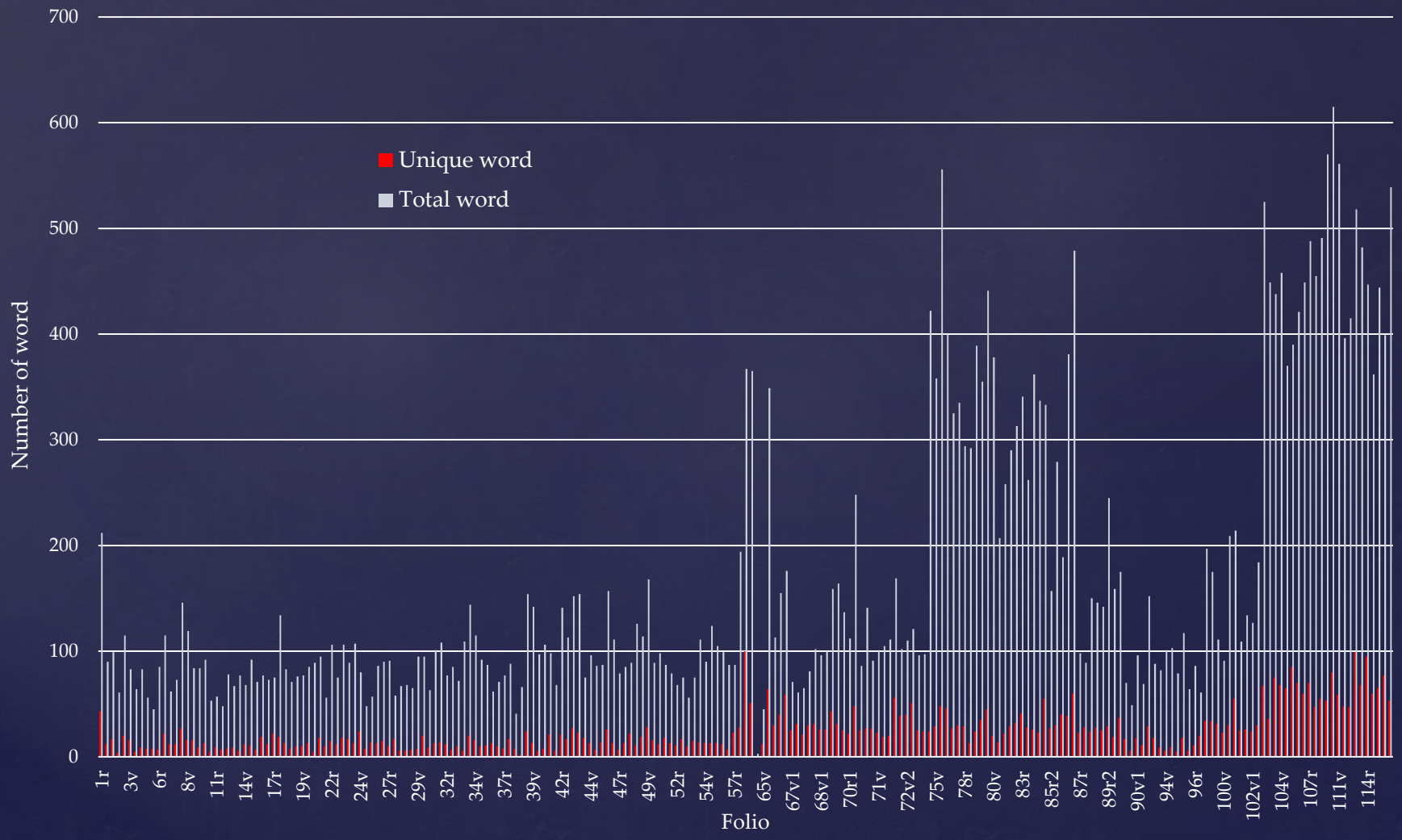
- Zipf's law can be used to determine if a given language is a natural language.
- Number of the most frequency word is as twice as the second most frequency word, and as 3 times as the third most frequency word.
- The total word of the English corpus is 40786, which is close to 37919 of Voynich.



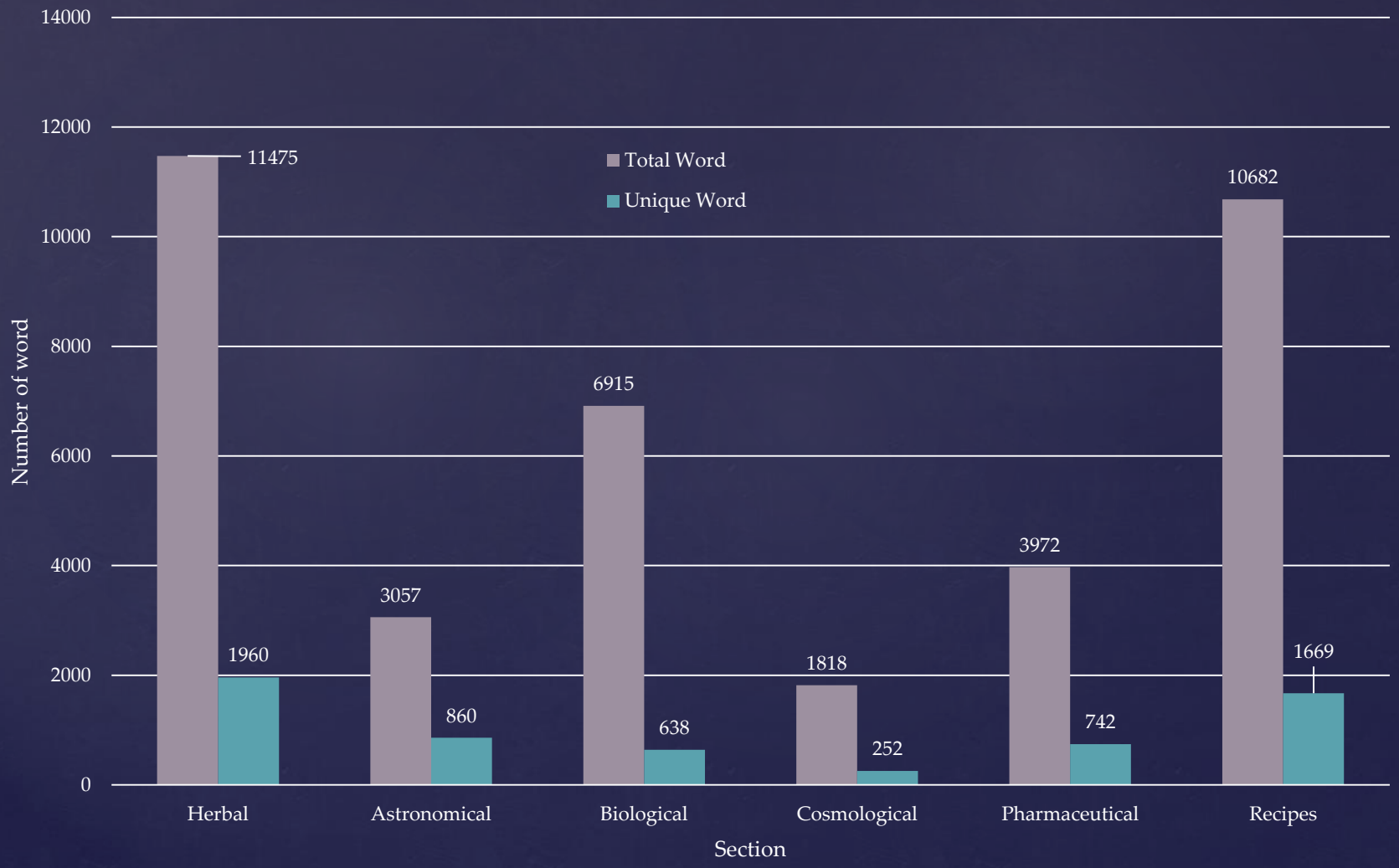
Illustrations investigation phase

- Write Matlab code to achieve the following functions:
  - Find out unique word tokens in each pages and sections
  - Determine the location of a given word token
  - Determine the frequency of a given word token
- Investigate the possible relation between texts, words and illustrations.

Total Word VS. Unique Word



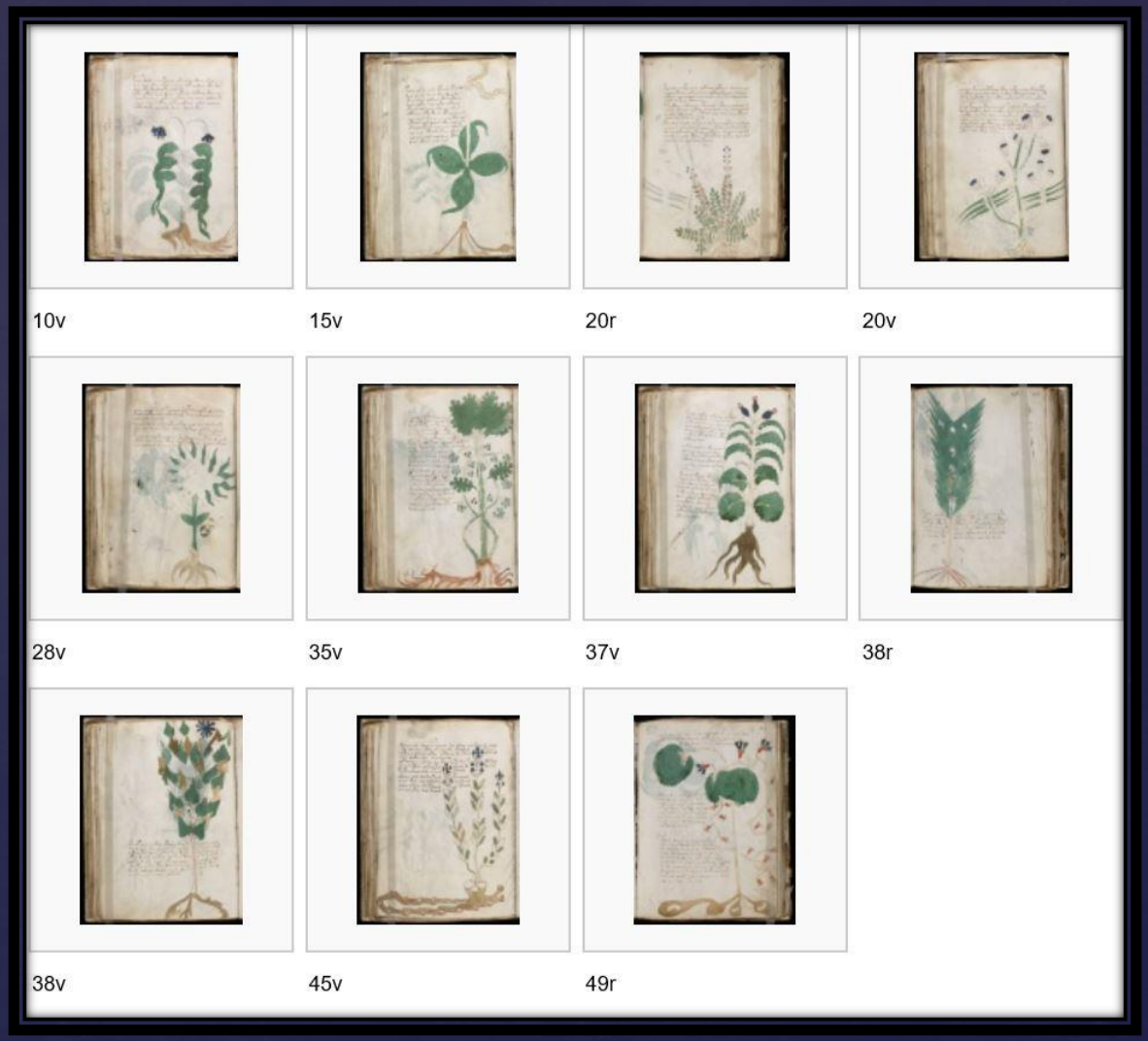
Total Word VS. Unique Word by section



Unique Word in herbal section

- Folios that contain the word “choiin”

Words	Frequency
choiin	13
cthaiin	13
x	9
f	8
ksho	8
v	7
cfhy	6
chkchy	6
kcho	5
kshy	5



Similar illustrations in folios

- Find out the common words in folios which have similar illustration
- Table below are the common word and frequency of the word

Common Word	Frequency
chor	220
s	244
aiin	470
sar	85
daiin	864

Folio 99r

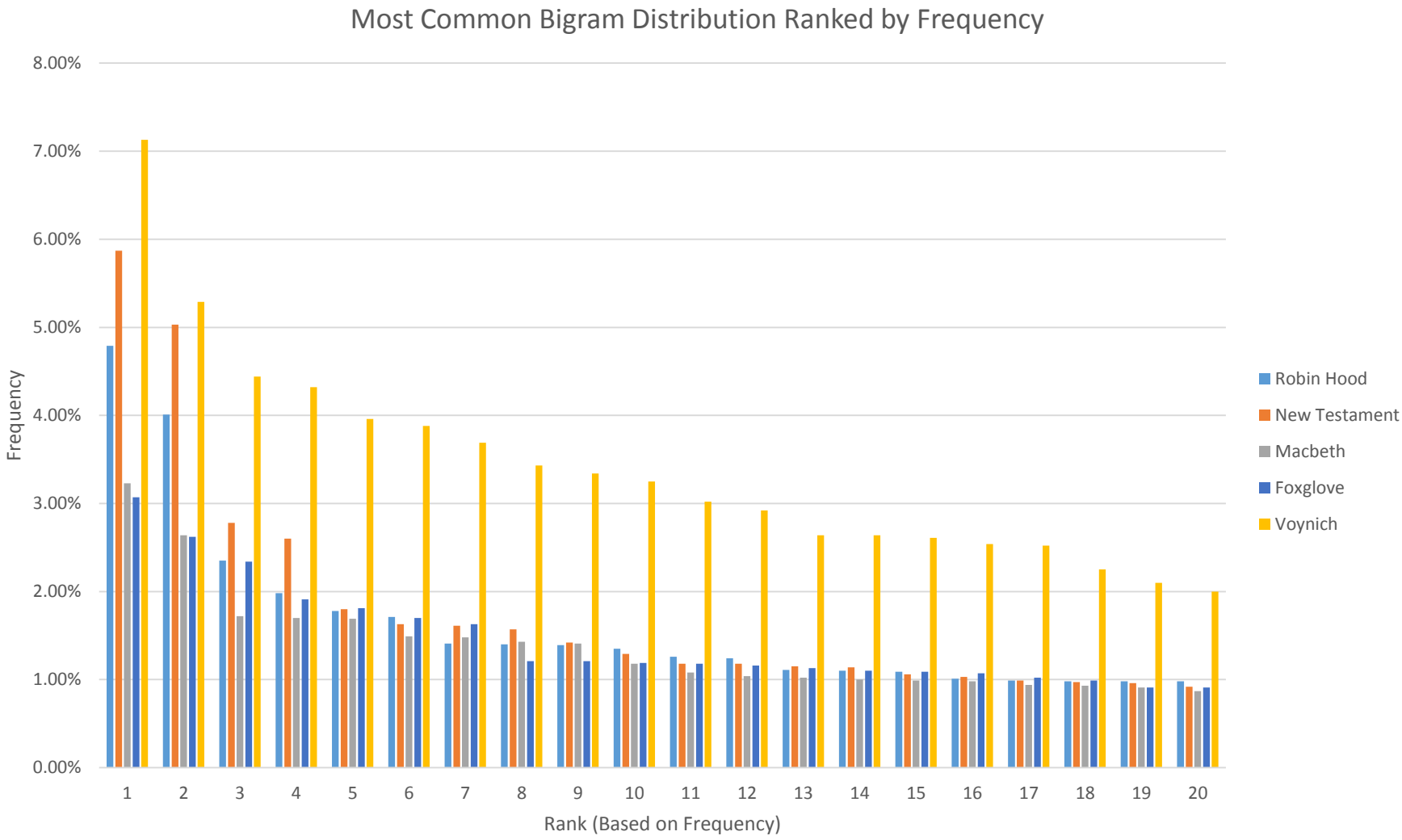
Folio 17v

Folio 96v

English Alphabet Investigation

- Investigate how basic statistics can be used to extract the alphabet.
- Use character bigrams.
- Bigrams give the frequency distribution of two adjacent elements.
- Sort character tokens into alphabet and non-alphabet characters.
- Write an extraction algorithm that could be run to determine any punctuation characters within the Voynich?

English Alphabet Investigation



English Alphabet Investigation

- Tabulate all possible bigrams based on the total unique character tokens.
- Example bigram data table subsection.

aa	783	0.03%	ba	2399	0.09%	ca	8303	0.31%	da	5748	0.22%	ea	20349	0.76%	fa	4310	0.16%
ab	4695	0.18%	bb	369	0.01%	cb	0	0.00%	db	8	0.00%	eb	1194	0.04%	fb	0	0.00%
ac	6151	0.23%	bc	0	0.00%	cc	1150	0.04%	dc	6	0.00%	ec	4383	0.16%	fc	0	0.00%
ad	7434	0.28%	bd	56	0.00%	cd	0	0.00%	dd	381	0.01%	ed	22641	0.85%	fd	3	0.00%
ae	2852	0.11%	be	17247	0.65%	ce	8817	0.33%	de	12647	0.47%	ee	11167	0.42%	fe	4680	0.18%
af	1882	0.07%	bf	0	0.00%	cf	0	0.00%	df	71	0.00%	ef	4713	0.18%	ff	2719	0.10%
ag	3523	0.13%	bg	0	0.00%	cg	0	0.00%	dg	1157	0.04%	eg	1340	0.05%	fg	0	0.00%
ah	4506	0.17%	bh	50	0.00%	ch	12276	0.46%	dh	6	0.00%	eh	2519	0.09%	fh	1	0.00%
ai	13608	0.51%	bi	1663	0.06%	ci	2499	0.09%	di	5072	0.19%	ei	7530	0.28%	fi	4545	0.17%
aj	34	0.00%	bj	34	0.00%	cj	0	0.00%	dj	7	0.00%	ej	323	0.01%	fj	0	0.00%
ak	5154	0.19%	bk	0	0.00%	ck	2265	0.09%	dk	3	0.00%	ek	925	0.03%	fk	0	0.00%
al	26608	1.00%	bl	2927	0.11%	cl	1856	0.07%	dl	547	0.02%	el	12622	0.47%	fl	1559	0.06%

English Alphabet Investigation

- Table can be used to determine:
  - ‘Valid’ bigrams
  - Character tokens that only appear at the start or end of a word token
- Non-alphabet characters can significantly differ statistically.
  - Can vary depending on the type of literature, writing, or even time.

1:	4028	0.15%	2:	3793	0.14%	3:	3396	0.13%	4:	3305	0.12%

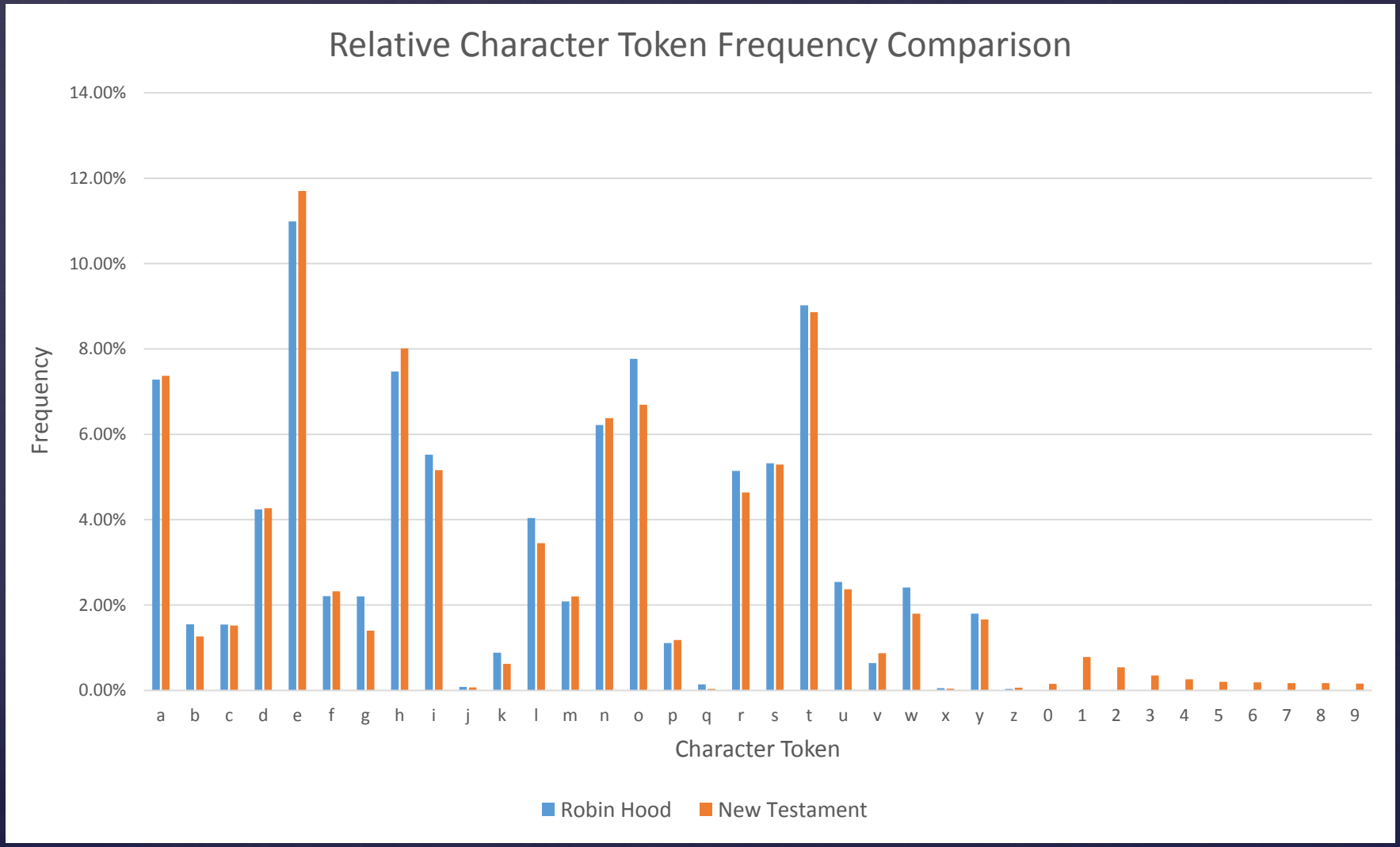
- ASCII representation may also represent the tokens differently.
  - Quotation marks.

"	362	0.04%

“	19	0.01%
”	19	0.01%

:1	11164	0.42%
:2	7257	0.27%
:3	3884	0.15%
:4	2292	0.09%
:5	1617	0.06%
:6	1345	0.05%
:7	1223	0.05%
:8	1175	0.04%
:9	1145	0.04%

English Alphabet Investigation



English Alphabet Investigation

- Used a basic algorithm in MATLAB:
  1. Does the character token only appear at the end of a word token?
     - Majority appear at the end of a word token?
  2. Does the character token only appear at the start of a word token?
     - Does this character have a high relative frequency when compared to others only appearing at the start of a word token?
  3. Does the character token have a high relative frequency?
     - Does this character token appear frequently before specific tokens?
  4. Does the character token have a high bigram ‘validity’?

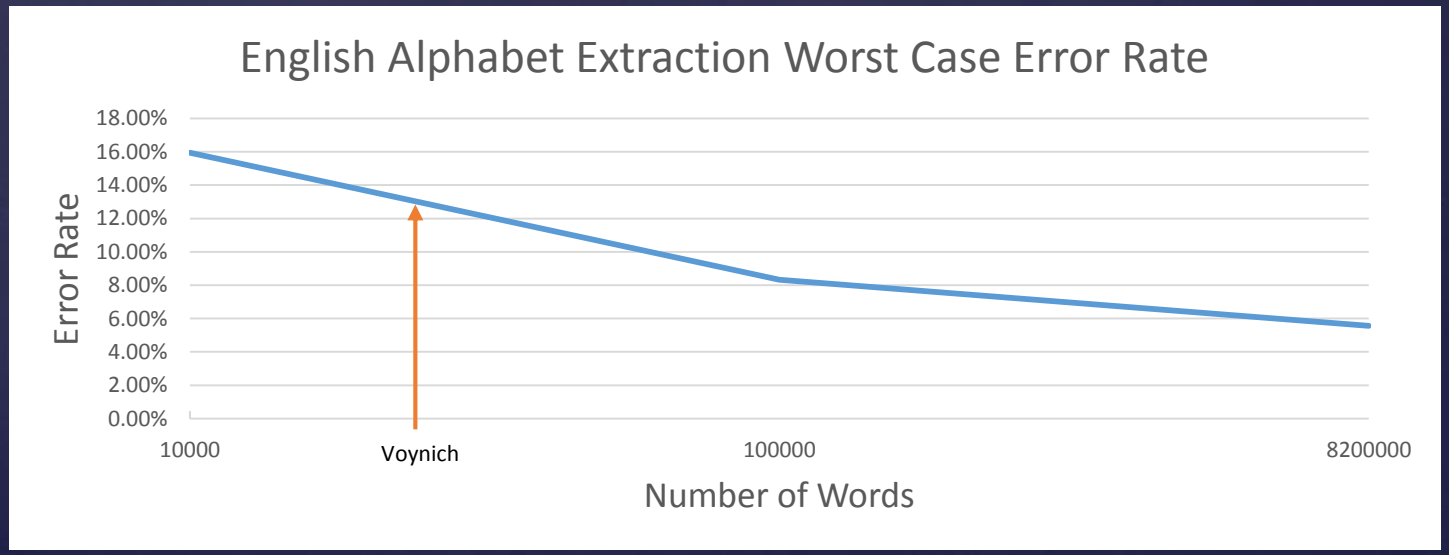
- Errors expected using small sample sizes

English Alphabet Investigation Results

Word Count	10000
Possible Alphabet Characters	
' (0 1 2 3 4 5 6 7 8 9 A B C D E F G H I J K L M N O P R S T U W Y Z a b c d e f g h i j k l m n o p q r s t u v w x y z	
Possible Non-Alphabet Characters	
!) , . : ; ?	

Word Count	100000
Possible Alphabet Characters	
(A B C D E F G H I J K L M N O P R S T U W Y Z a b c d e f g h i k l m n o p q r s t u w y	
Possible Non-Alphabet Characters	
! ') , . : ; ? 0 1 2 3 4 5 6 7 8 9 V j v x z	

Word Count	8200000
Possible Alphabet Characters	
(A B C D E F G H I J K L M N O P R S T U W Y Z a b c d e f g h i k l m n o p q r s t u w x y z	
Possible Non-Alphabet Characters	
!) , . : ; ? ' 0 1 2 3 4 5 6 7 8 9 V j v	



English Alphabet Investigation Results

- Voynich results show that characters ‘v’ and ‘z’ were possible non-alphabet characters.
- Common errors occurred from low occurring alphabet symbols
  - j, v, x, etc.
- High occurrences of numerical symbols
- Or punctuation symbols appearing at the start of word tokens

Possible Alphabet Characters
* a c d e f g h i k l m n o p q r s t x y
Possible Non-Alphabet Characters
v z

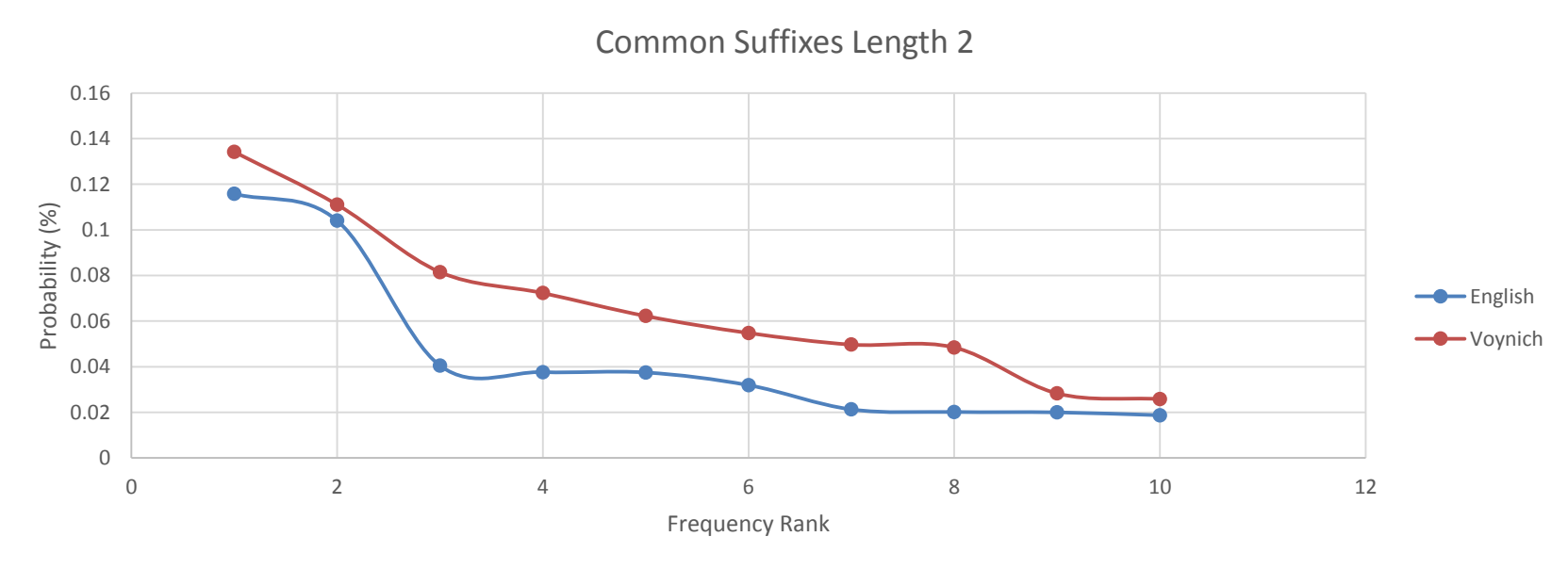
- Both reported characters had the lowest frequencies of the Voynich Alphabet (> 0.02% each).
- Not enough data on reported non-alphabet tokens.

Morphology Investigation

- Morphology deals with the structure of morphemes.
  - The smallest grammatical unit in a language.
- Check if word tokens appear ‘within’ other word tokens.
- Look at the most common prefix and suffix character token combinations at different token lengths.
  - Rank according to frequency.
- Interesting results when comparing with the most common suffix combinations of English

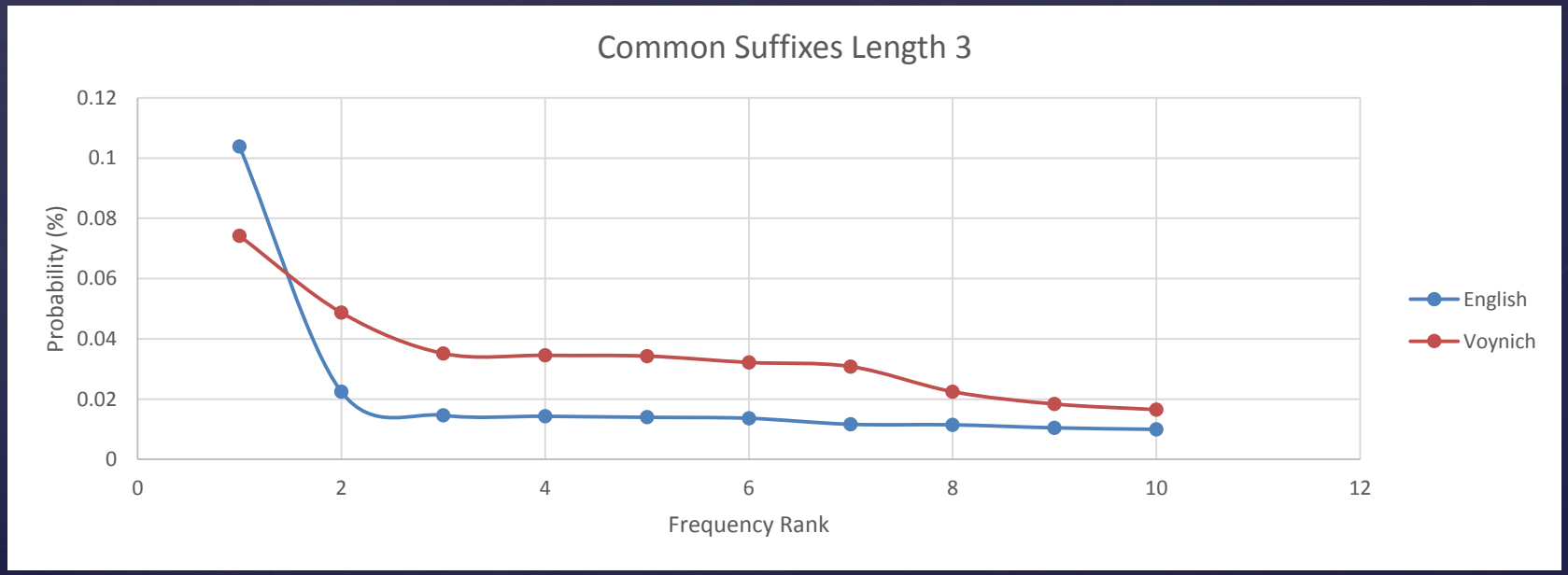
Morphology Investigation Results

- Similar curves albeit with lower probabilities in English.
  - Second ranked suffix have a similar probability.



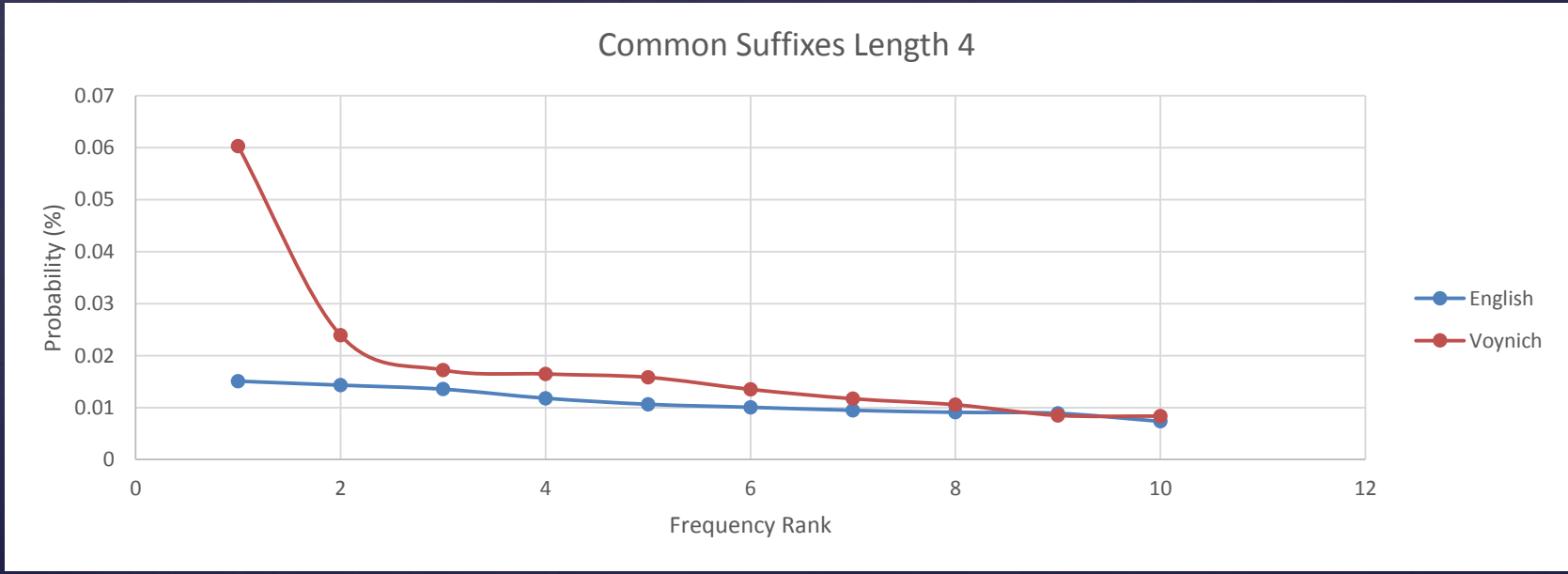
Morphology Investigation Results

- Significant drop in highest ranked suffix of the Voynich.
  - Now lower than that of English.
- Both English and Voynich approaching almost linear curves after second ranked suffix.

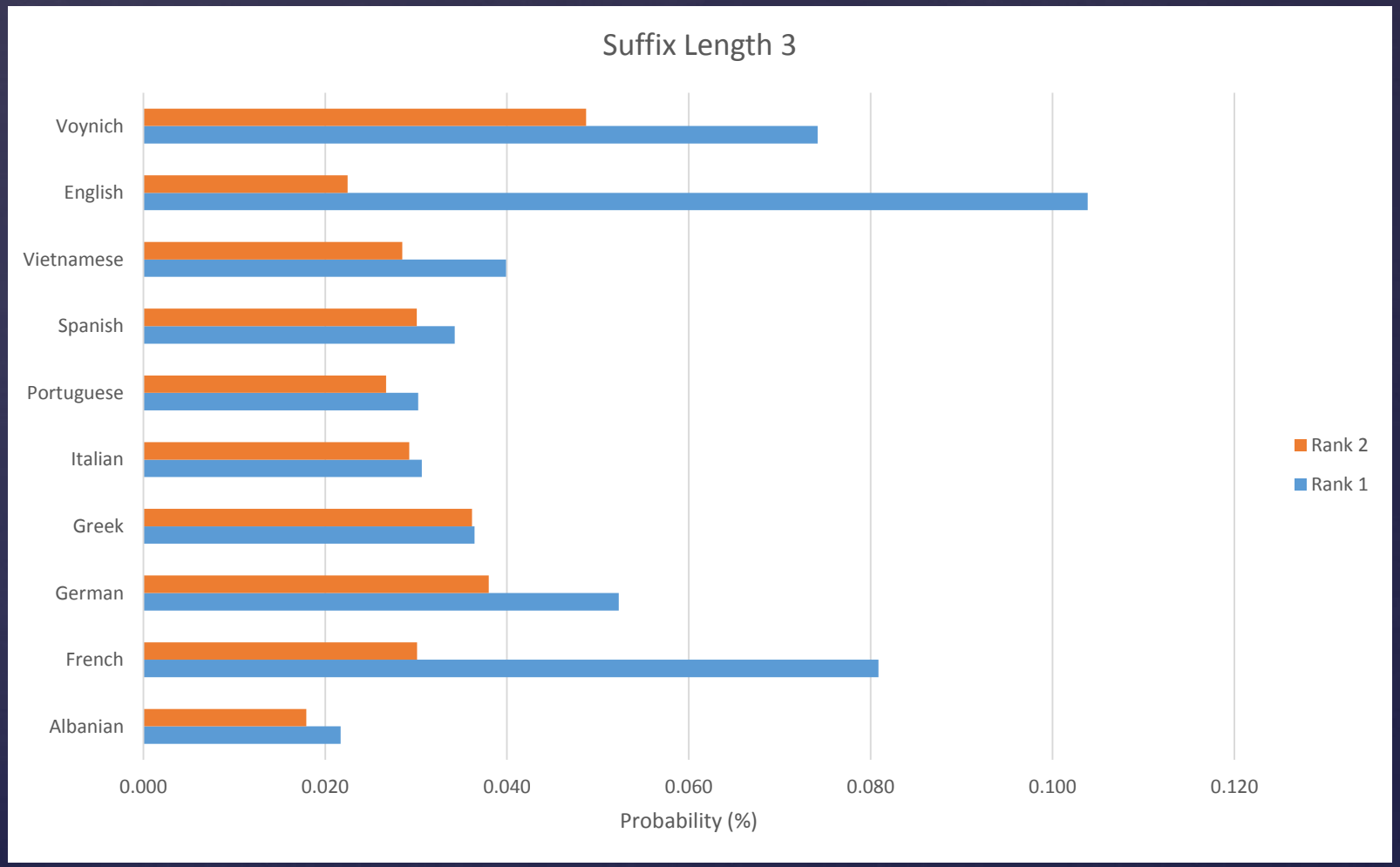


Morphology Investigation Results

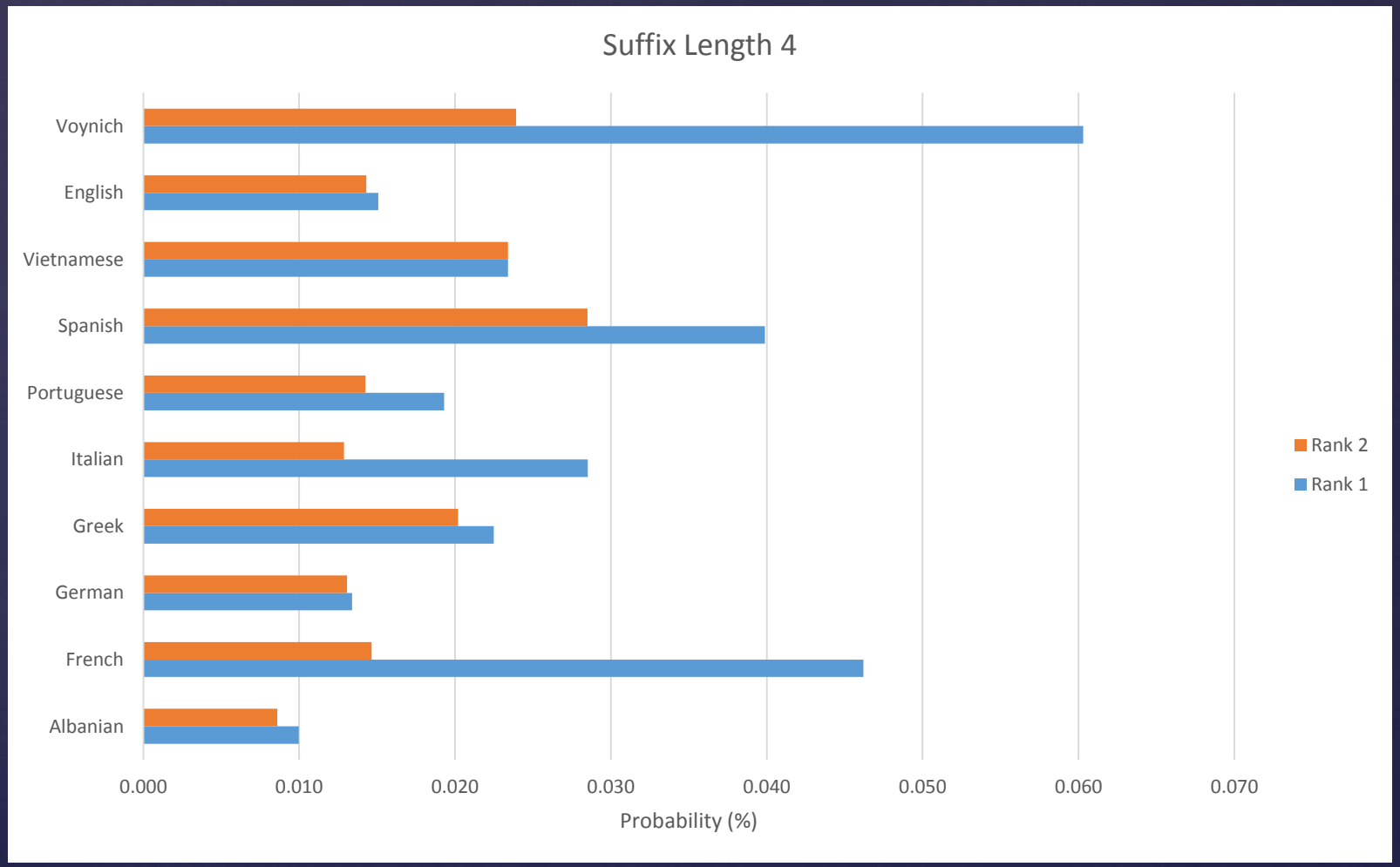
- Voynich curve stays relatively similar.
- English now has a very flat curve.
- Compare the ratios between top two ranked suffixes of length 3 and 4.



Morphology Investigation Results



Morphology Investigation Results



Morphology Investigation Results

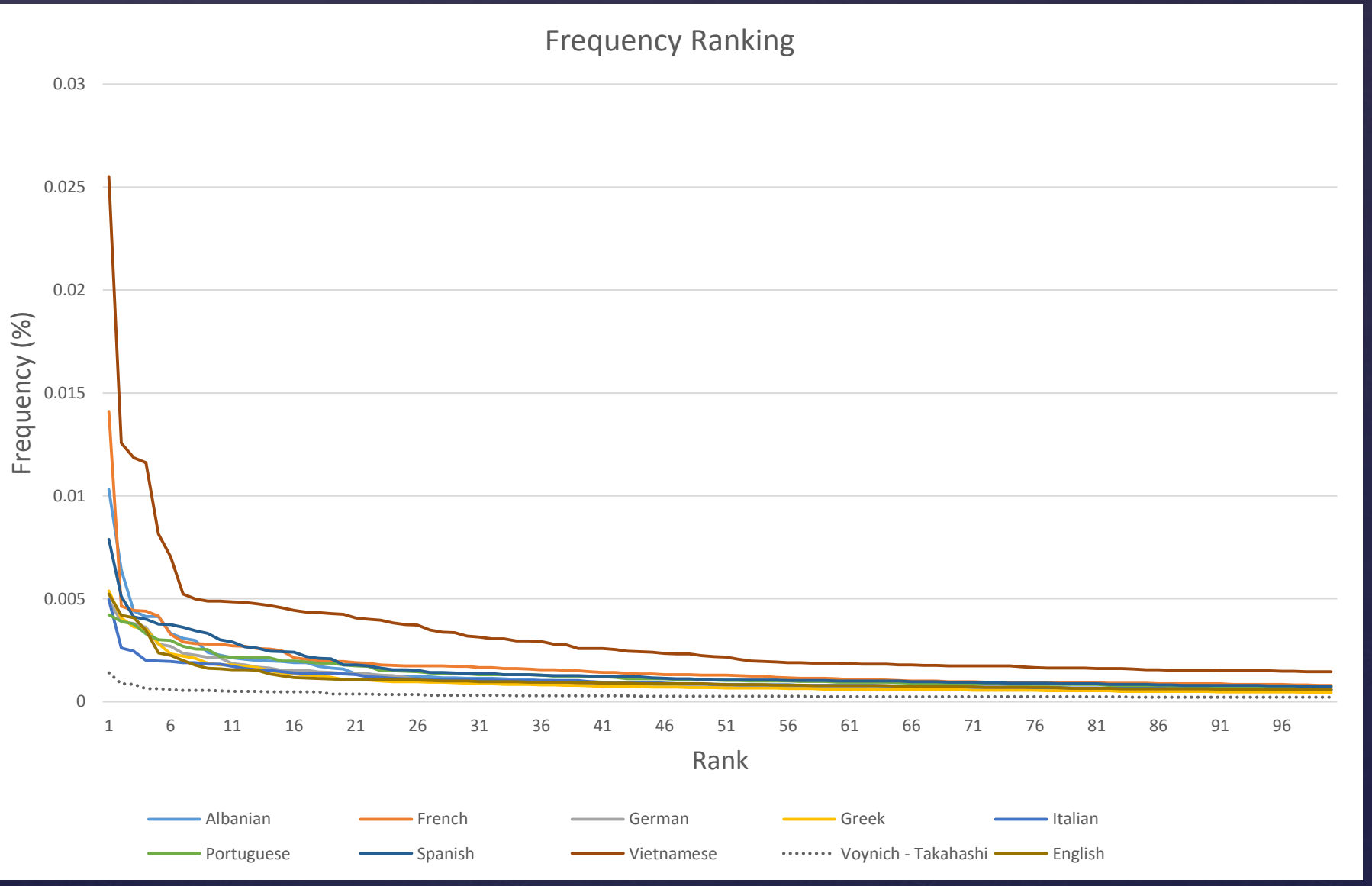
- Tabulated the difference ratios between the top two ranked suffixes.
- French has the most similar difference ratio of length 4.
- No tested language exceptionally close to the difference ratio of the Voynich at length 3.

	Difference	
Language	Length 3	Length 4
Albanian	0.00377	0.00139
French	0.05075	0.03155
German	0.01429	0.00033
Greek	0.00028	0.00028
Italian	0.00138	0.01564
Portuguese	0.00353	0.00503
Spanish	0.00418	0.00503
Vietnamese	0.01140	0.00000
English	0.08139	0.00077
Voynich	0.02547	0.03638

Word Collocation Investigation

- Word Pairs that co-occur more often than would be expected by chance.
  - ‘deep thought’, ‘heavy rain’, etc.
- Can give the association between words as a quantitative measure.
- Initially rank collocations by frequency.
- Significant difference between tested languages and the Voynich.

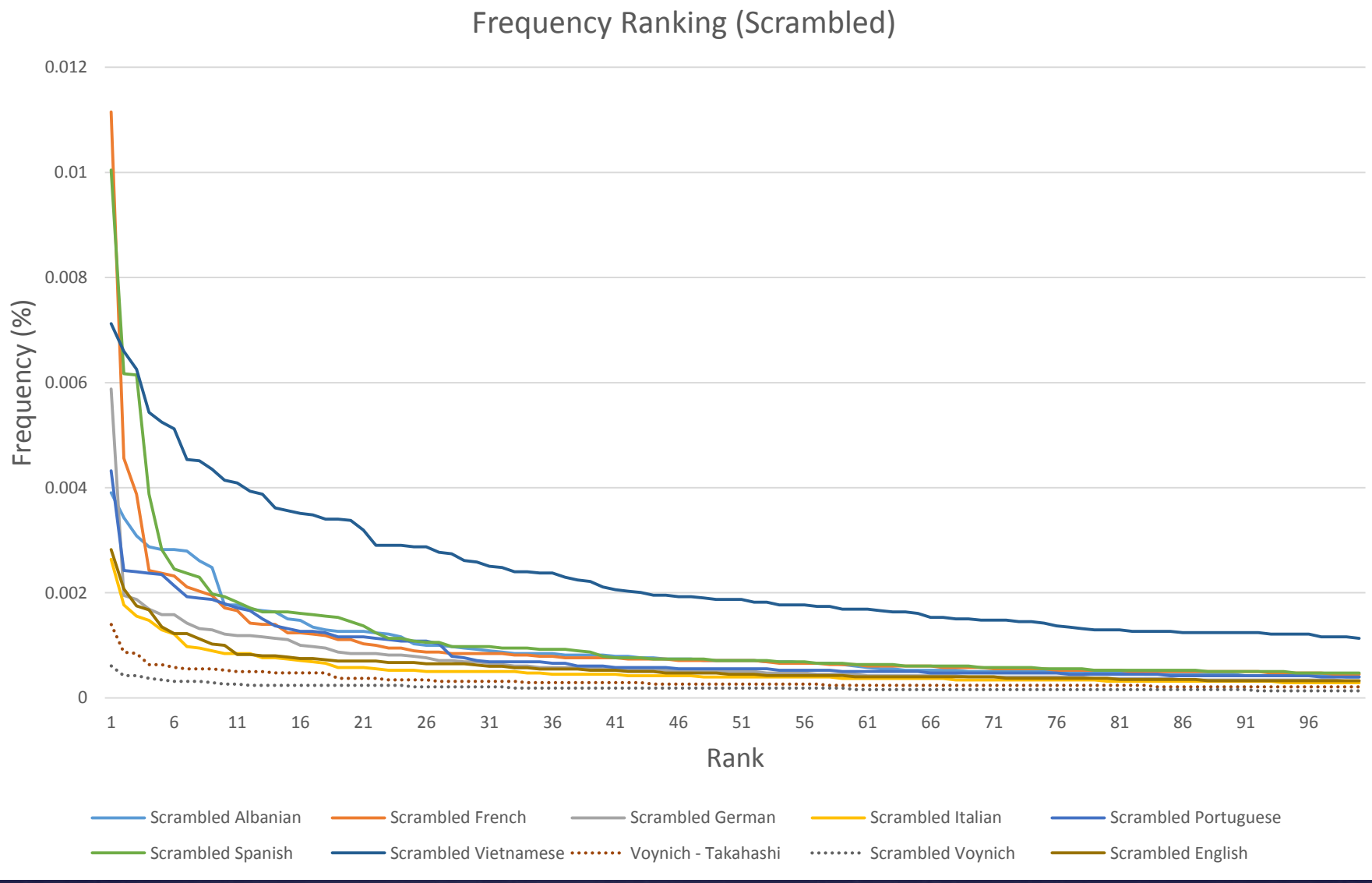
Word Collocation Results



Word Collocation Results

- Based on an association measure of frequency, Voynich has a weak measure of word association.
- What would happen if we ‘scrambled’ the texts and repeated the collocation test?
  - Testing if Voynich was created by randomly placing words.
- Also scramble the Voynich.

Word Collocation Results



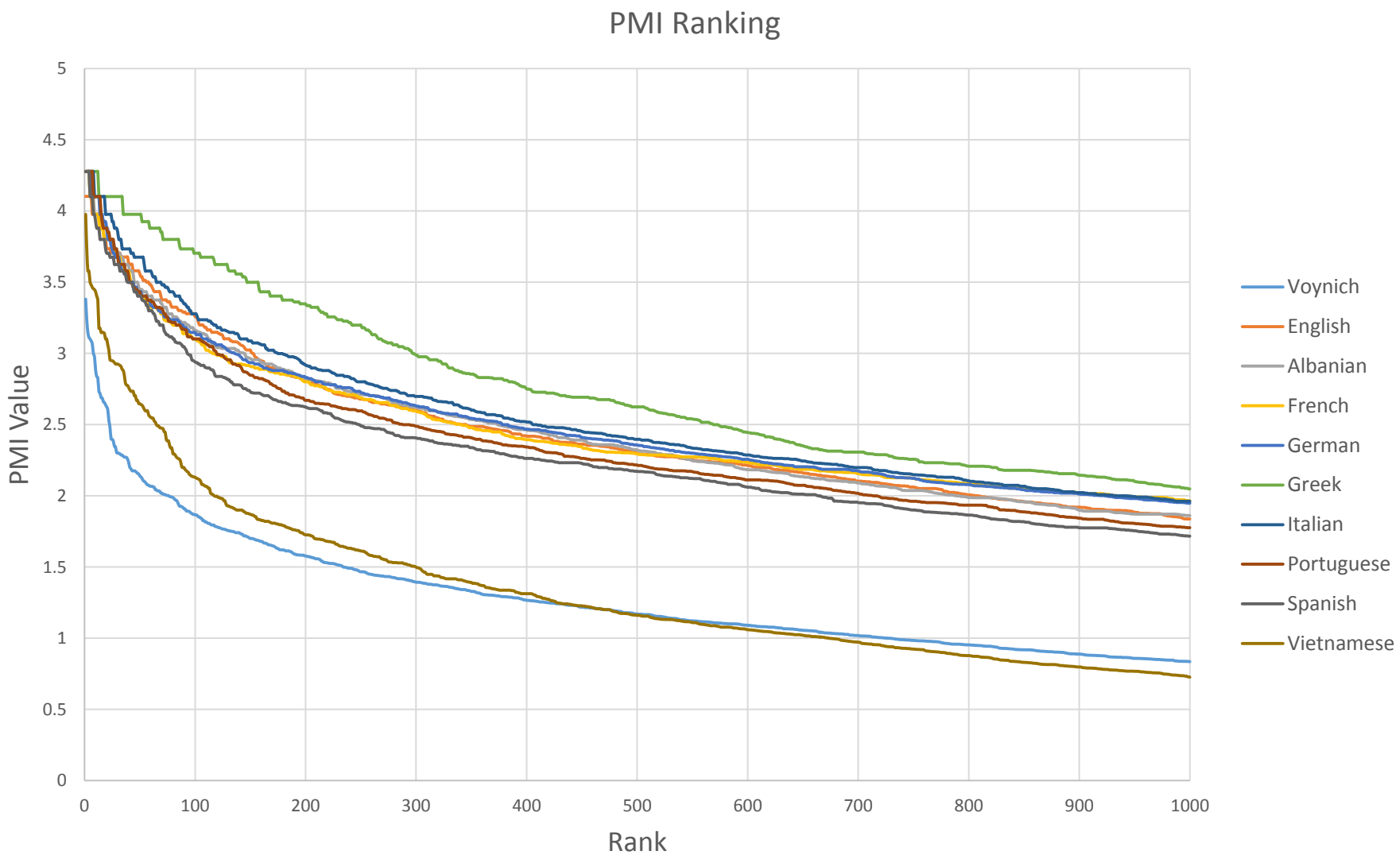
Word Collocation PMI Ranking

- Ranking by frequency does not take into account the probability of the separate words.
  - Many common English collocations involved function words.
- Use a different measure of association.
  - Pointwise Mutual Ranking (PMI)
- Mathematically:

$$PMI(x;y) = \log(\frac{P(x,y)}{P(x)P(y)})$$

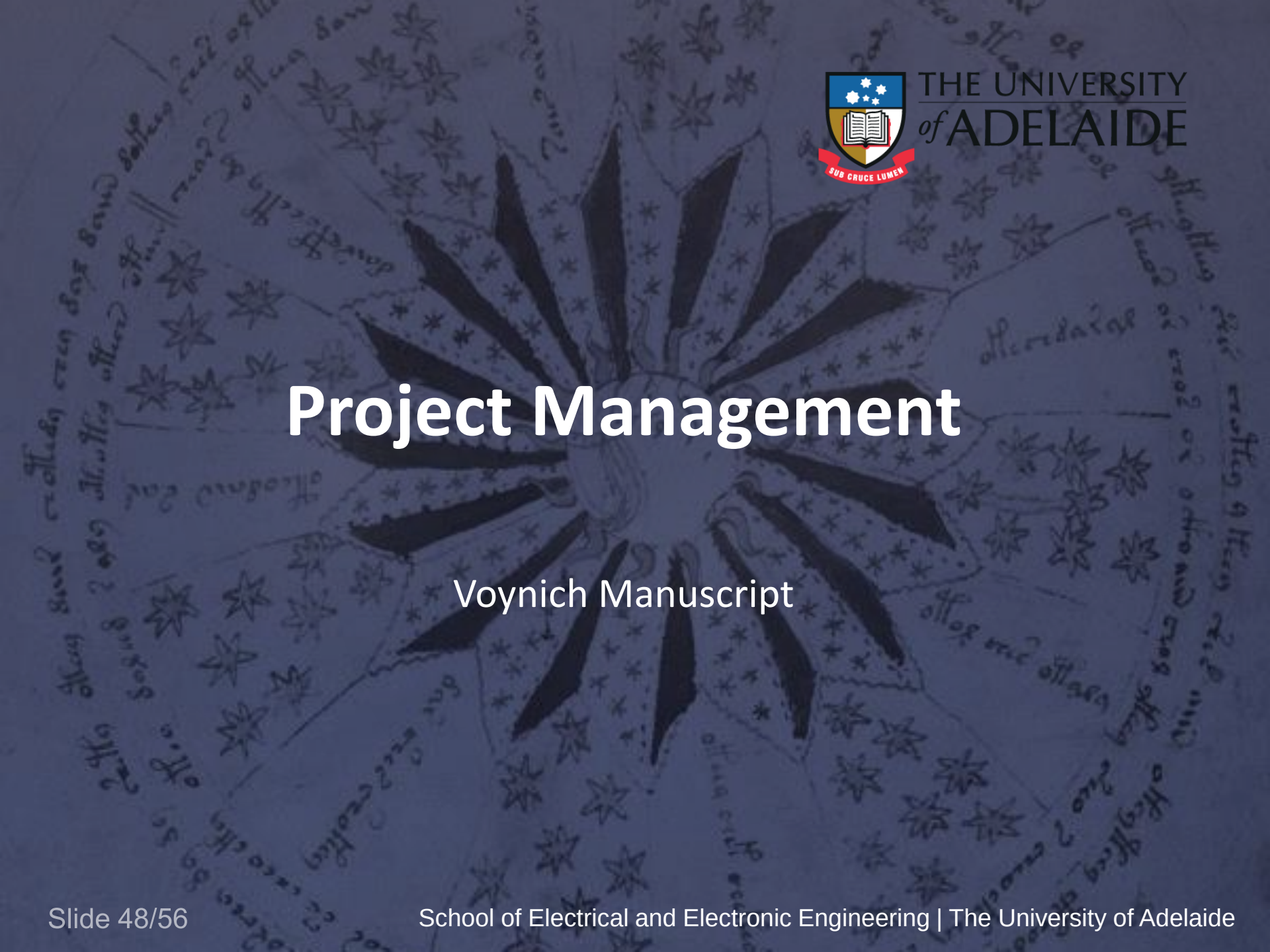
- This is, however, sensitive to low frequency material.
  - Exclude ‘rare’ words

Word Collocation PMI Results



Word Collocation PMI Results

- Vietnamese appears to have a very weak word association measure when ranking using PMI.
  - Seemed odd to go from the highest ranked language to the lowest ranked language.
- Tests to be repeated using different reference material.
  - Appears to be missing characters, not correctly representing words.
- PMI appears to give a better measure of association.
  - Can plot many more collocations.
  - Takes into account common words.



Project Management

Voynich Manuscript

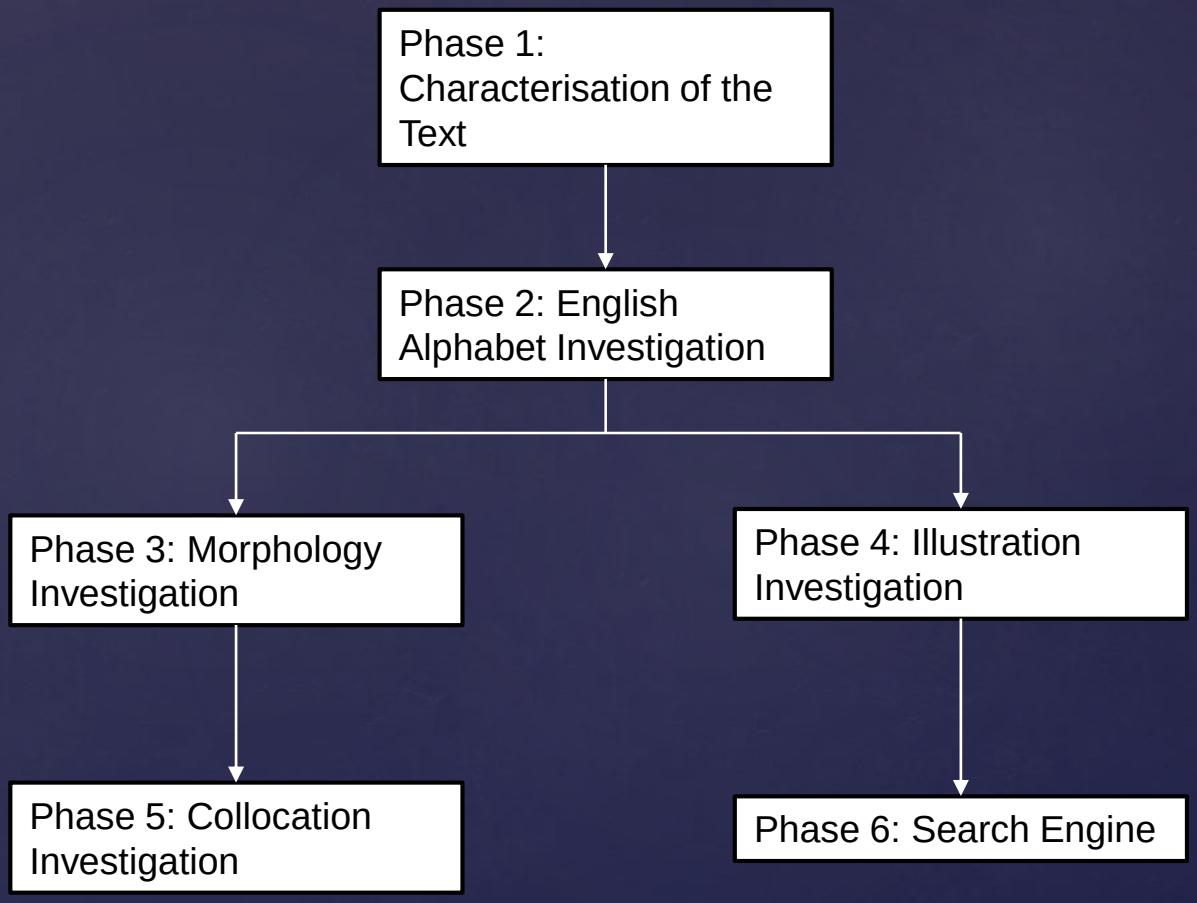
Risk Management

Very Low	1 to 2
Low	2 to 3
Moderate	4 to 6
High	7 to 8
Very High	9 to 10

Risks identified that could affect the project.

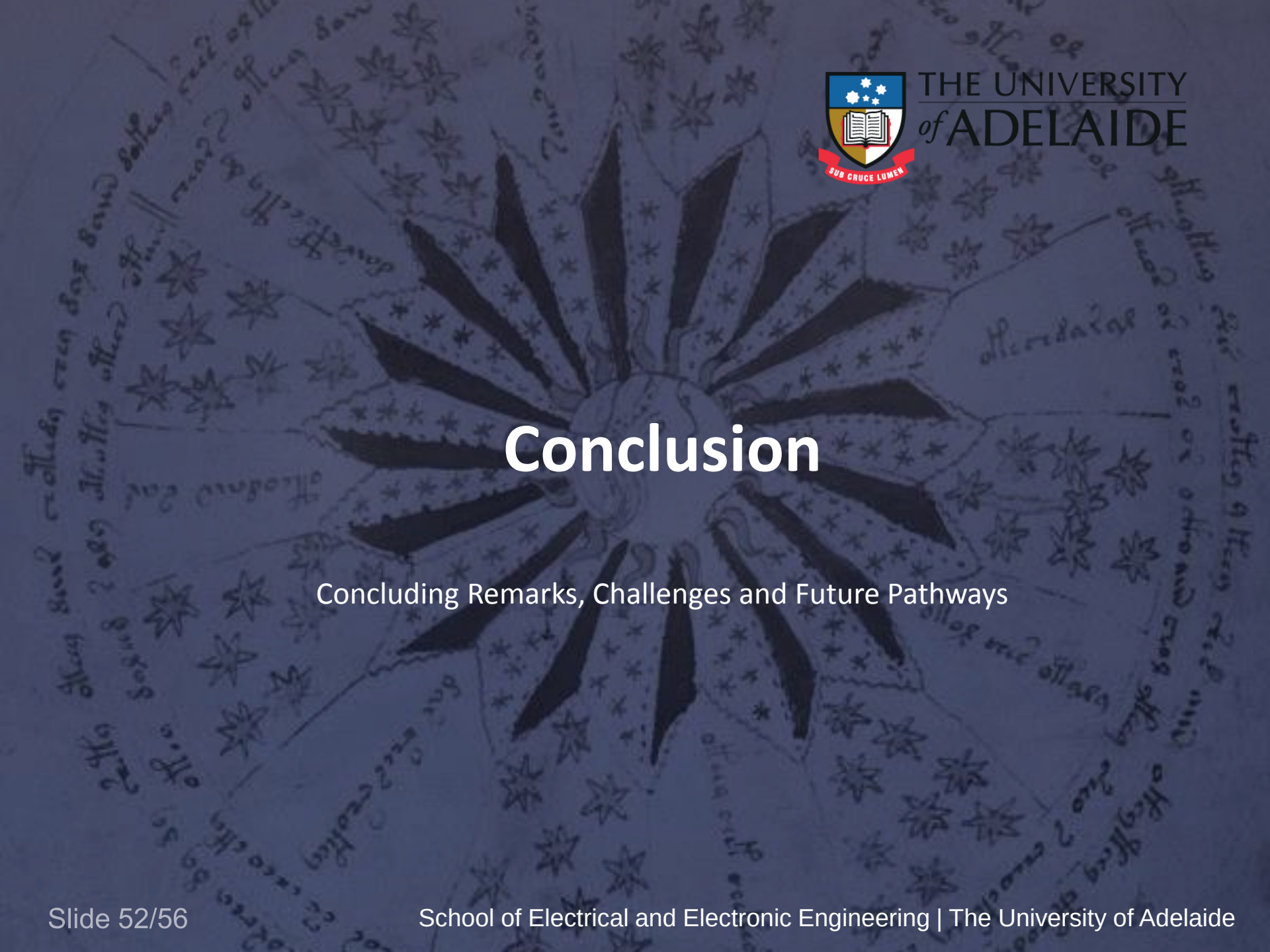
No.	Risk	Likelihood	Consequence	Risk Level
1	**Misunderstanding of Tasks**	**Very High [9]**	**High[8]**	**72**
2	**Underestimation and/or mismanagement of time and resources**	**High [8]**	**High [7]**	**56**
3	Health related issues from long periods at computers	High [7]	Moderate [6]	42
4	Team member illness or injury	Very High [9]	Moderate [4]	36
5	Issues with communication between team and/or supervisors	Low [3]	High [7]	21
6	Loss of software code and/or files	Low [2]	Very High [10]	20

Work Breakdown



Work Breakdown

- Andrew
  - C++ Pre-processing Code
  - Code for counting the features of the Voynich and comparing with other known languages. (Phase 1)
  - Code for separating the English alphabet from other tokens. (Phase 2)
  - Phase 3 – Morphology Investigation
  - Phase 5 – Collocation Investigation
- Lifei
  - Finding tokens that only appear at the start of words and which are only at the end. (Phase 1)
  - Counting the features of token (Phase 2)
    - token frequency, token recurrence interval, etc.
  - Phase 4 – Illustration Investigation
  - Phase 6 – Search Engine
- Budget of $500 - Unused



Conclusion

Concluding Remarks, Challenges and Future Pathways

Conclusion

- Some words unique to particular sections of the Voynich.
- Possibility of no punctuation or numerical representation, nor distinction between upper or lower-case in alphabet
  - Small Alphabet size.
  - Extraction Results.
  - May be represented with regular alphabet tokens similar to Greek or Roman.
- Basic morphology appears to be represented
  - French suffixes appear closest but still relatively different.
- Weak word association
  - Close relationship to current tested Vietnamese, may have a relationship to other Asian languages.
  - Could be related to a particular cipher or code.

Challenges

- Differing transcriptions
  - Transcribers use different character tokens
- Relatively small sample size
- Style/Type of text
- Basic understanding of other languages

Future Pathways

- Investigate combinations of characters as punctuation.
- Investigate Stylometry – Does the Voynich have a written stylistic relationship to a specific book type or author(s).
- Expand corpus to include more Asian languages.
- Expand Collocations with different measures of association.
  - Use a less strict collocation extraction method.
- Investigate ciphers or codes of a similar time period.
- Repeat tests!


Do you have any questions?