

What We Know About the Voynich Manuscript

Kevin Knight

University of Southern California

Sravana Reddy

Dartmouth College

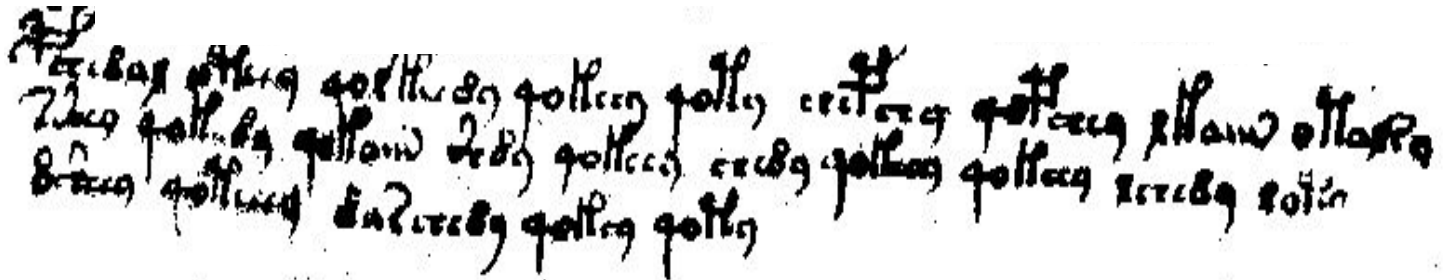


- Sources:
- Mary D'Imperio, **The Voynich Manuscript, An Elegant Enigma** (1978)
 - Kennedy & Churchill, **The Voynich Manuscript** (2006)
 - Prescott Currier, **Some Important New Statistical Findings** (1976)
 - Rene Zandbergen, **Currier A and B: Two Different Languages?** (1997)
 - Rene Zandbergen, <http://www.voynich.nu/>
 - <http://www.voynich.ms/forum/>
 - Cipher Mysteries blog (Nick Pelling)
 - Knight & Reddy article (2011) in Workshop on Lang. Tech. for Cultural Heritage, Social Sci., and Humanities.

Stanford University
March 2013

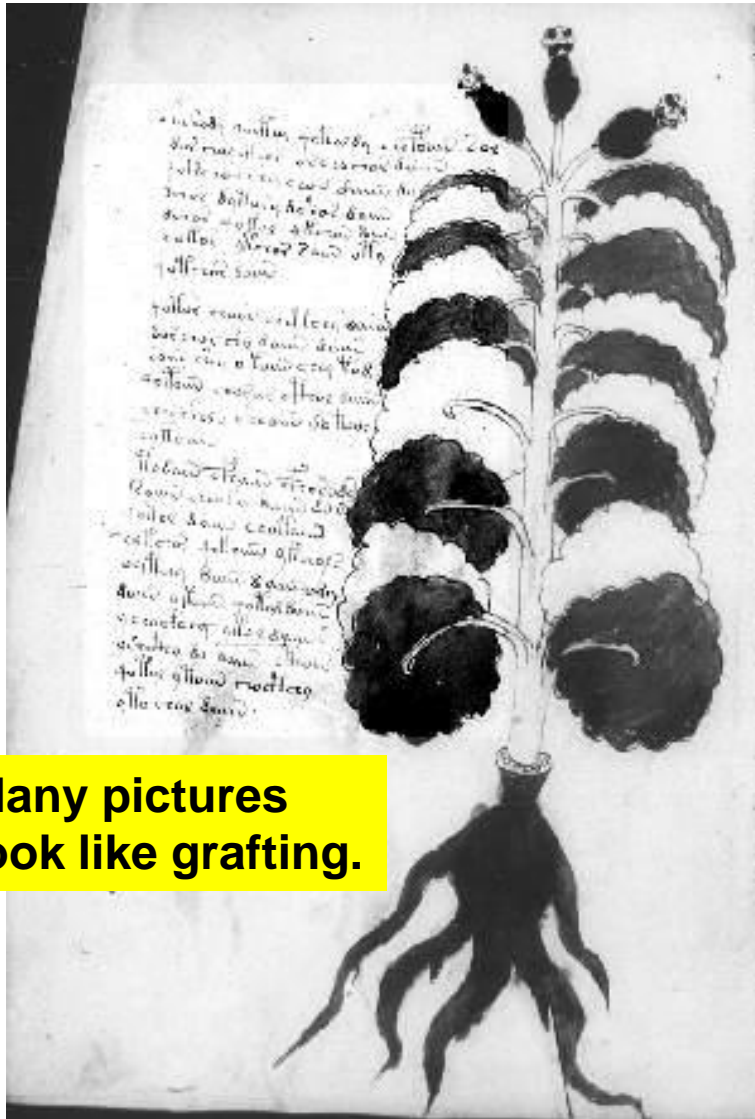
Voynich Manuscript (VMS)

- Medieval illustrated manuscript
 - 235 pages on vellum material
 - Color drawings of plants, nymphs, stars, etc.
 - 38,000 words written in an unknown script



- Undeciphered!

“Herbal” section

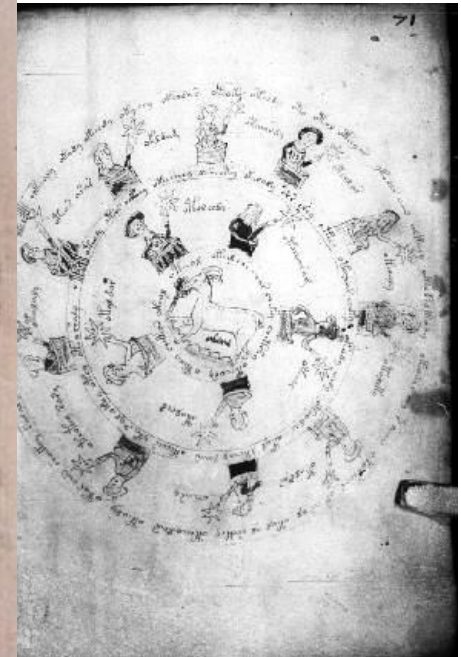
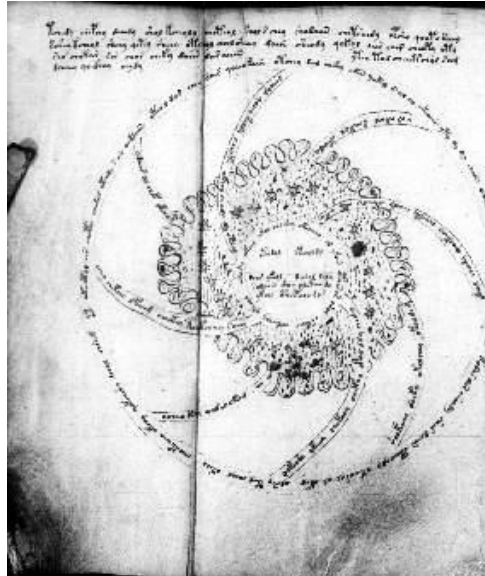


Many pictures
look like grafting.



Sunflower? Would date
VMS as post-1492.

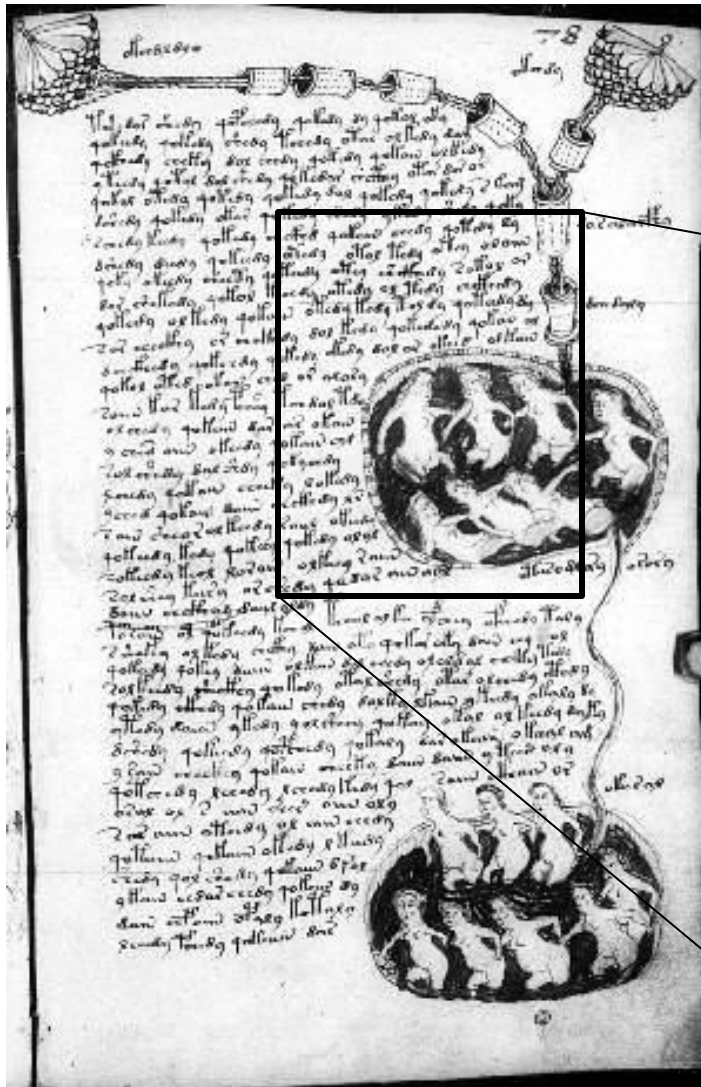
“Astrological” section



“Biological” section

Small nudes in baths

Interconnecting tubes of liquids





“Pharmacological” section

History of Voynich Manuscript



**William Newbold,
Polymath, PhD UPenn**



**Wilfrid Michael Voynich
book dealer**

- 1921 WV presents VMS + **inserted letter** mentioning Francis Bacon, \$160k price
- 1921 Newbold & WV announce decipherment

One-Page Letter Tucked Into VMS

Reverend and Distinguished Sir; Father in Christ:

This book bequeathed to me by an intimate friend, I destined for you, **my very dear Athanasius [Kircher]**, as soon as it came into my possession, for I was convinced that it could be read by no one except yourself. The **former owner** of this book once asked your opinion by letter ... Accept now this token ...

Dr Raphael, tutor in the Bohemian language to Ferdinand III, then King of Bohemia, told me the said book **had belonged to the Emperor Rudolf** and that he presented the bearer who brought him the book 600 ducats. He believed the author was **Roger Bacon**, the Englishman. On this point I suspend judgment ... At the command of your reverence,

Joannes Marcus **Marci** of Cronland
Prague, 19 August, 1665(6?)



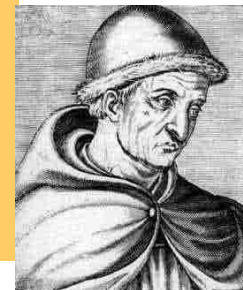
Kircher,
super-scholar,
recipient of
this letter



???,
owned VMS
before Marci



Emperor
Rudolf,
paid 600 ducats
for VMS



Roger Bacon
(1214-94)
“first scientist”

History of Voynich Manuscript

1576-1612 Rudolf II purchases VMS

16xx Marci inherits VMS from ??

1665 Marci sends VMS to Kircher
with letter

1665-80 Kircher owns VMS

1680 Kircher dies

1921 WV presents VMS + Marci letter
mentioning Bacon, \$160k price

1921 Newbold & WV announce decipherment

History of Voynich Manuscript

1576-1612 Rudolf II purchases VMS

1608-1622 J. de Tepenecz signs VMS
in Bohemian court

1912 WV acquires VMS in “ancient castle”

1914 WV moves to USA, opens bookshop

1919 WV sends photostatic copies of VMS

1919 Copying reveals Tepenecz signature

1921 WV presents VMS + Marci letter
mentioning Bacon, \$160k price

1921 Newbold & WV announce decipherment

16xx Marci inherits VMS from ??

1665 Marci sends VMS to Kircher
with letter

1665-80 Kircher owns VMS

1680 Kircher dies

History of Voynich Manuscript

1864 Ethel Boole born in England

1865 WV born in Lithuania

1885 WV imprisoned, Polish nationalist

1890 WV & EB meet, marry in 1902

1898 WV publishes first book list

1912 WV acquires VMS in “ancient castle”

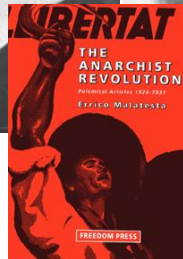
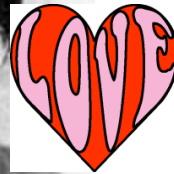
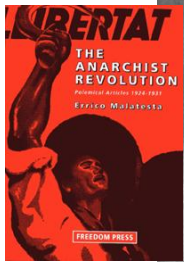
1914 WV moves to USA, opens bookshop

1919 WV sends photostatic copies of VMS

1919 Copying reveals de Tepenecz signature

1921 WV presents VMS + Marci letter mentioning Bacon, \$160k price

1921 Newbold & WV announce decipherment



History of Voynich Manuscript



- 1864 Ethel Boole born in England
- 1865 WV born in Lithuania
- 1885 WV imprisoned, Polish nationalist
- 1890 WV & EB meet, marry in 1902
- 1898 WV publishes first book list
- 1912 WV acquires VMS in “ancient castle”
- 1914 WV moves to USA, opens bookshop
- 1919 WV sends photostatic copies of VMS
- 1919 Copying reveals de Tepenecz signature

- 1921 WV presents VMS + Marci letter
mentioning Bacon, \$160k price
- 1921 Newbold & WV announce decipherment
- 1930 WV dies. VMS placed in vault, \$100k**
- 1931 VMS appraised at \$19,400**
- 1960 Ethel dies, VMS to secretary Ann Nill**
“Castle” revealed as Villa Mondragone
- 1961 NY dealer Hans Kraus buys for \$24,500**
- 1969 Kraus donates VMS to Yale**

History of Voynich Manuscript

1576-1612 Rudolf II purchases VMS

1608-1622 J. de Tepenecz signs VMS
in Bohemian court



**“Barschius” owns VMS
between J. de Tepenecz
and Marci**

16xx Marci inherits VMS from ??

1665 Marci sends VMS to Kircher
with letter

1665-80 Kircher owns VMS

1680 Kircher dies

1864 Ethel Boole born in England

1865 WV born in Lithuania

1885 WV imprisoned, Polish nationalist

1890 WV & EB meet, marry in 1902

1898 WV publishes first book list

1912 WV acquires VMS in “ancient castle”

1914 WV moves to USA, opens bookshop

1919 WV sends photostatic copies of VMS

1919 Copying reveals de Tepenecz signature

1919 WV writes to Bohemian State Archvs

1921 WV presents VMS + Marci letter
mentioning Bacon, \$160k price

1921 Newbold & WV announce decipherment

1930 WV dies. VMS placed in vault, \$100k

1931 VMS appraised at \$19,400

1960 Ethel dies, VMS to secretary Ann Nill
“Castle” revealed as Villa Mondragone

1961 NY dealer Hans Kraus buys for \$24,500

1969 Kraus donates VMS to Yale

1972 Brumbaugh finds WV letters in BSA

History of Voynich Manuscript

1576-1612 Rudolf II purchases VMS

1608-1622 J. de Tepenecz signs VMS
in Bohemian court

**1630s George Baresch owns VMS
sends letter to Kircher**

1639 GB writes Kircher again

16xx Marci inherits VMS from GB

1665 Marci sends VMS to Kircher
with letter

1665-80 Kircher owns VMS

1680 Kircher dies

1864 Ethel Boole born in England

1865 WV born in Lithuania

1885 WV imprisoned, Polish nationalist

1890 WV & EB meet, marry in 1902

1898 WV publishes first book list

1912 WV acquires VMS in “ancient castle”

1914 WV moves to USA, opens bookshop

1919 WV sends photostatic copies of VMS

1919 Copying reveals de Tepenecz signature

1919 WV writes to Bohemian State Archvs

1921 WV presents VMS + Marci letter
mentioning Bacon, \$160k price

1921 Newbold & WV announce decipherment

1930 WV dies. VMS placed in vault, \$100k

1931 VMS appraised at \$19,400

1960 Ethel dies, VMS to secretary Ann Nill
“Castle” revealed as Villa Mondragone

1961 NY dealer Hans Kraus buys for \$24,500

1969 Kraus donates VMS to Yale

1972 Brumbaugh finds WV letters in BSA

**200x Zandbergen finds 1639 Baresch letter
in newly online Kircher archive**

Newbold Decipherment



Marci letter → Bacon → Cabala → “letter doubling” cipher

	A	B	C	D	E	F	G	H	I	L	M	N	O	P	Q	R	S	T	U	V	X	Z
A	V	Z	B	F	G	L	M	N	N	O	...											
B	C	F	T	U	V	X	...															
C	F	B	A	Q	F	C	D	Z	Z	...												
D																						
E																						
F																						
G																						
H																						
I												N										
L																						
M																						
N													A									
O																						
P				N																		
Q																						
R																						
S																						
T																						
U																						
V																						
X																						
Z																						

22x22 table

Encoding:

A → CC, OM, ...

B → ...

...

N → HA, MI, DO, NU ...

...

Z → ...

Decoding:

...

DO → N

...

Encoder has freedom to devise a “cover text” to hide real message.

Example:

a n n ... → DO MI NU ... → DOMINU ...

Newbold System

- Too hard to assemble good “cover” text!
- **So, make cipher letter-pairs overlap:**
a n n ... → AD DB BR ... → ADBR ...
- **Then, employ anagramming:**
a n n ... → OM DO MI ... → DO OM MI ... → DOMI ...
- Now can construct a plausible looking “cover” text in Latin for our secret message (also in Latin)
- An ingenious system, to be sure!

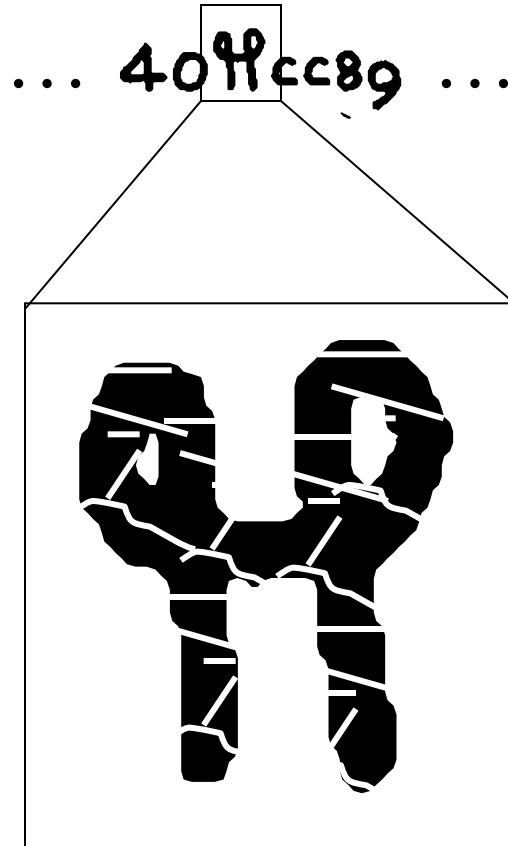
Newbold Decipherment

Hmm, by the method, both plaintext **and**
ciphertext should be in Latin letters...

But the VMS doesn't have Latin letters...



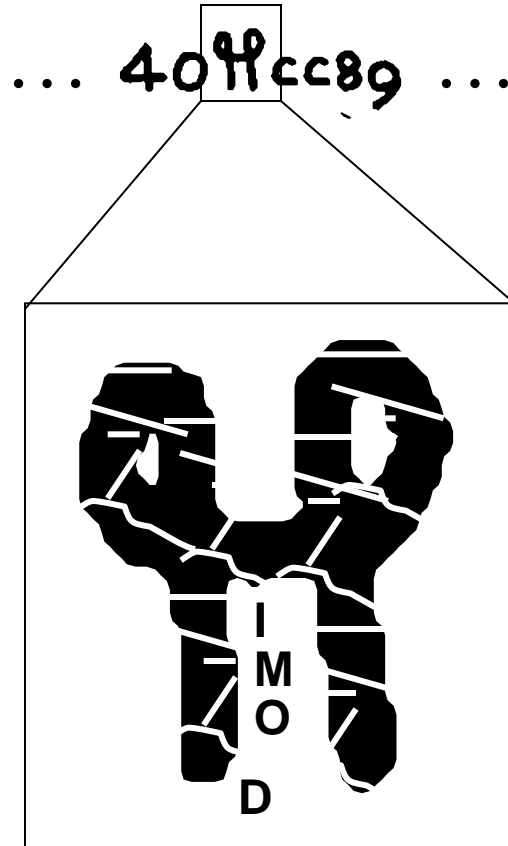
William Newbold,
Polymath, PhD UPenn



apparent
ciphertext



William Newbold,
Polymath, PhD UPenn



apparent
ciphertext

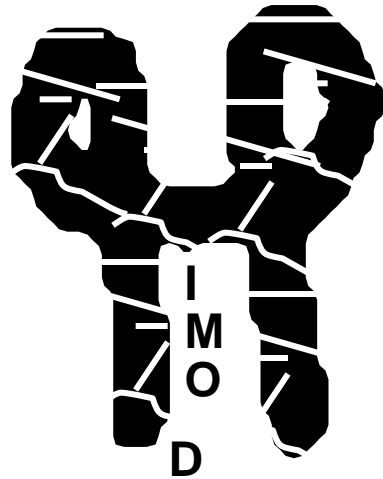
real
ciphertext:
DOMI...



Let's Decipher with Newbold !

Hcc89 ...

apparent ciphertext



real ciphertext
DOMI...

doubling

DO OM MI ...

non-deterministic
anagramming

OM DO MI ...

lookup in 22^2 table

a n n ...

non-deterministic
mapping from 11
Latin letters to full 22

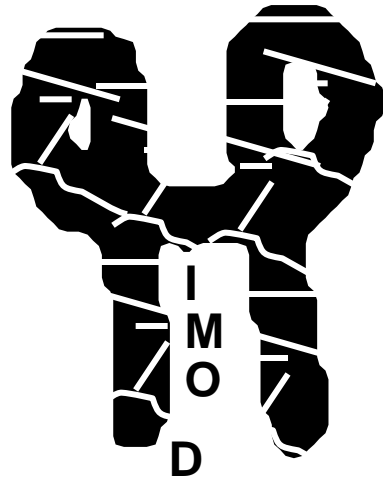
o n n ...



Let's Decipher with Newbold !

Hcc89 ...

apparent ciphertext



	A	B	C	D	E	F	G	H	I	L	M	N	O	P	Q	R	S	T	U	V	X	Z
A	V	Z	B	F	G	L	M	N	N	O	...											
B	C	F	T	U	V	X	...															
C	F	B	A	Q	F	C	D	Z	Z	...												
D																						
E																						
F																						
G																						
H																						
I												N										
L													A									
M																						
N																						
O			N																			
P																						
Q																						
R																						
S																						
T																						
U																						
V																						
X																						
Z																						

22x22 table
(values guessed)

real ciphertext

DOMI...

doubling

DO OM MI ...

non-deterministic
anagramming

OM DO MI ...

lookup in 22^2 table

a n n ...

non-deterministic
mapping from 11
Latin letters to full 22

o n n ...



Newbold's Results

1300 real ciphertext “letters” in first 3 lines

Decipherment of those first lines:

“I, Roger Bacon, have written this...”

(in Latin)

Anagramming sets of 55 letters is sometimes required.

Slow but steady progress... Andromeda galaxy, ovaries ... so ... Roger Bacon must have had a microscope & telescope, hundreds of years before they were invented ... !

whew!

let's take a step back ...

VMS Text

- 38,000 words
- Unknown script
- Writing style similar to 15th century Florentine “humanist” hand
- Between 23 and 40 distinct characters
- No corrections, likely to have been copied
- Writing was done after illustrations

Transcription

ፖረቱጋል ማህጸን ማህጸን ማህጸን ማህጸን ማህጸን ማህጸን ማህጸን ማህጸን ማህጸን ማህጸን
ፖረቱጋል ማህጸን ማህጸን ማህጸን ማህጸን ማህጸን ማህጸን ማህጸን ማህጸን ማህጸን ማህጸን
ፖረቱጋል ማህጸን ማህጸን ማህጸን ማህጸን ማህጸን ማህጸን ማህጸን ማህጸን ማህጸን ማህጸን

ፖረቱጋል ማህጸን 40ጸ ማህጸን 40ጸ ማህጸን 40ጸ ማህጸን ማህጸን 40ጸ ማህጸን ማህጸን ማህጸን ማህጸን
ፖረቱጋል ማህጸን 40ጸ ማህጸን ማህጸን ማህጸን ማህጸን ማህጸን ማህጸን ማህጸን ማህጸን ማህጸን
ፖረቱጋል ማህጸን ማህጸን ማህጸን ማህጸን ማህጸን ማህጸን ማህጸን ማህጸን ማህጸን

BSC8AE OPCC9 4OE FCC89 4OFCC9 4OP9 SCBS9 4OBSC9 EFAM OPAE29
2ZC9 4OFC89 4OFAM Z89 4OFCC9 SC89 4OFCC9 4OFCC9 ESC89 EOP9
8ZC9 4OPCCC9 8ARSC89 4OFC9 4OP9

last paragraph, f103r

Alphabet: Carrier/D'Imperio Transcription

Ɱ	Ɱ	Ɱ
C	S	Z

Ɱ	Ɱ	Ɱ	Ɱ
P	F	B	V

Ɱ	Ɱ	Ɱ	Ɱ
Q	X	W	Y

Ɱ	Ɱ	Ɱ	Ɱ	Ɱ	Ɱ	Ɱ
J	A	E	R	O	I	D

Ɱ	Ɱ	Ɱ	Ɱ	Ɱ	Ɱ
6	7	8	9	4	2

Ɱ	Ɱ	Ɱ
G	H	1

Ɱ	Ɱ	Ɱ
T	U	0

Ɱ	Ɱ	Ɱ
N	M	3

Ɱ	Ɱ	Ɱ
K	L	5

Alphabet: Carrier/D'Imperio Transcription

Ɱ	Ɱ	Ɱ
C	S	Z

Ɱ	Ɱ	Ɱ	Ɱ
P	F	B	V

Ɱ	Ɱ	Ɱ	Ɱ
Q	X	W	Y

Ɱ	Ɱ	Ɱ	Ɱ	Ɱ	Ɱ	Ɱ
J	A	E	R	O	I	D

Ɱ	Ɱ	Ɱ	Ɱ	Ɱ	Ɱ
6	7	8	9	4	2

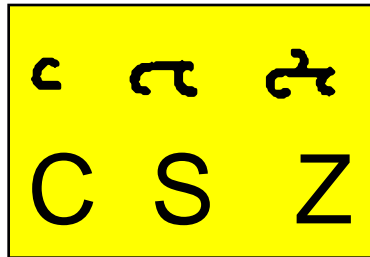
Ɱ	Ɱ	Ɱ
G	H	1

Ɱ	Ɱ	Ɱ
T	U	0

← Maybe this is really
IR IIR IIIR

There are several transcription schemes to choose from.

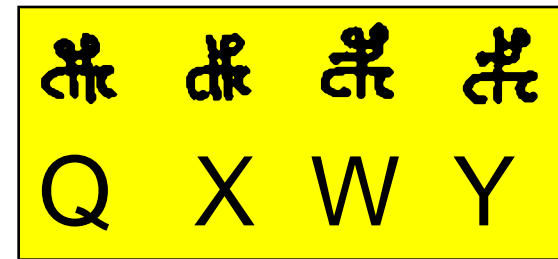
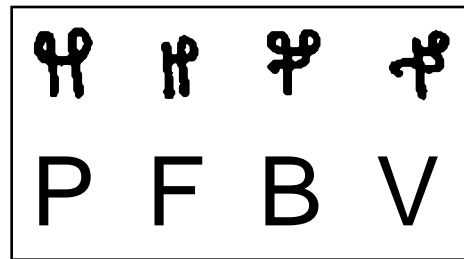
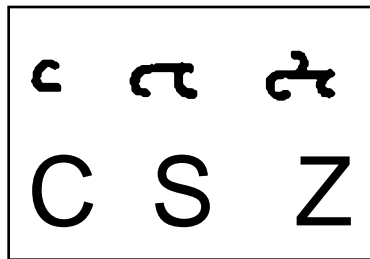
Alphabet: Carrier/D'Imperio Transcription



Variations of ꞑ, or separate characters?



Alphabet: Carrier/D'Imperio Transcription



Are these ligatures?

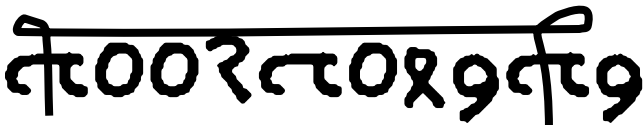
Is  just a fancy way of writing   ?

If you didn't know English, how would you know if f_i was the same as f_i ?

Suppose f_i **never** occurred. Would that be evidence?

Suppose f_i did occur, with the **same** contexts as f_i (e.g., *shing)?

Suppose f_i did occur, but **never** in the same context as f_i ?

Another common motif: 

Letter Frequencies

count	letter	count	letter	count	letter
25468	O	2886	2	148	U
20227	C	1752	N	96	6
17655	9	1413	B	74	Y
14281	A	1046	J	52	K
12973	8	950	Q	31	G
11008	S	908	X	17	L
10471	E	591	T	14	H
10026	F	524	*	2	1
6716	R	431	V	1	5
5994	P	316	I	1	0
5423	4	217	W		
4501	Z	157	D		
4076	M	156	3		

Total
63k character tokens

Most Frequent “Words”

count word

863	8AM	8ᎠᎠᎠ
537	OE	ᎠᎠ
501	SC89	ᎠᎠ89
469	AM	ᎠᎠᎠ
426	ZC89	ᎠᎠ89
396	SOE	ᎠᎠᎠ
363	OR	ᎠᎠ
350	AR	ᎠᎠ
344	SC9	ᎠᎠ9
318	8AR	8ᎠᎠ
308	4OFCC9	4ᎠᎠᎠᎠ9
305	4OFCC89	4ᎠᎠᎠᎠ89
283	ZC9	ᎠᎠ9
279	4OFAN	4ᎠᎠᎠᎠᎠ
272	4OFC89	4ᎠᎠᎠᎠ89
270	89	89
262	4OFAM	4ᎠᎠᎠᎠᎠ
260	AE	ᎠᎠ
253	8AE	8ᎠᎠ
243	2	2
219	SOR	ᎠᎠᎠ

count word

212	OFAM	ᎠᎠᎠᎠᎠ
211	8AN	8ᎠᎠ
191	4OFAE	4ᎠᎠᎠᎠᎠ
186	ZOE	ᎠᎠᎠ
177	OFCC9	ᎠᎠᎠᎠ9
174	SCC9	ᎠᎠᎠᎠ9
172	SCOE	ᎠᎠᎠᎠᎠ
155	S9	ᎠᎠ9
155	OPC89	ᎠᎠᎠᎠ89
154	OPAM	ᎠᎠᎠᎠᎠ
152	4OFAR	4ᎠᎠᎠᎠᎠ
151	9	9
151	4OE	4ᎠᎠ
150	S89	ᎠᎠ89
147	4OF9	4ᎠᎠᎠ9
144	ZCC9	ᎠᎠᎠᎠ9
144	OFAN	ᎠᎠᎠᎠᎠ
144	2AM	2ᎠᎠᎠ
143	OPAE	ᎠᎠᎠᎠᎠ
141	OPAR	ᎠᎠᎠᎠᎠ
140	SX9	ᎠᎠᎠᎠ9

count word

140	OPCC9	ᎠᎠᎠᎠᎠ9
138	OFAE	ᎠᎠᎠᎠᎠ
130	ZO	ᎠᎠᎠ
129	OFAR	ᎠᎠᎠᎠᎠ
119	ESC89	ᎠᎠᎠᎠ89
118	OFC89	ᎠᎠᎠᎠ89

etc

Total:
8116 distinct words

Word Repeats

- 115 (out of 8116) distinct words appear doubled at least once

... 40ፑፑፑፑ 40ፑፑፑፑ ...

- 8 distinct words appear tripled

... 40ፑፑፑፑ 40ፑፑፑፑ 40ፑፑፑፑ ...

... ፕፐፐ ፕፐፐ ፕፐፐ ...

... ራራፐፐ ራራፐፐ ራራፐፐ ...

... 0ፑፑፑፑ 0ፑፑፑፑ 0ፑፑፑፑ ...

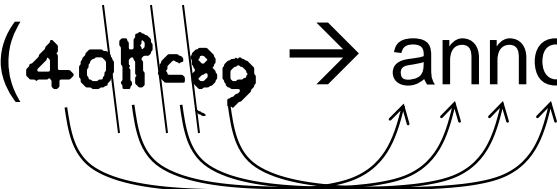
... 0ፐ 0ፐ 0ፐ ...

... 9ፑፑፑፑ 9ፑፑፑፑ 9ፑፑፑፑ ...

... 8ፑፑፑ 8ፑፑፑ 8ፑፑፑ ...

... 40ፑፑፑፑፑፑ 40ፑፑፑፑፑፑ 40ፑፑፑፑፑፑ ...

Cryptogram Theory

- Newbold (1921)
- Manly (1931), critique of Newbold
- Feely (1945), abbreviated Latin
- Strong (1945), polyalphabetic cipher
 - no details given, for national security reasons
- Brumbaugh (1972), numerological box
- Several attempts in the 1990s
- 1-for-n substitution (4011c89 → anno)

Theory of William Friedman

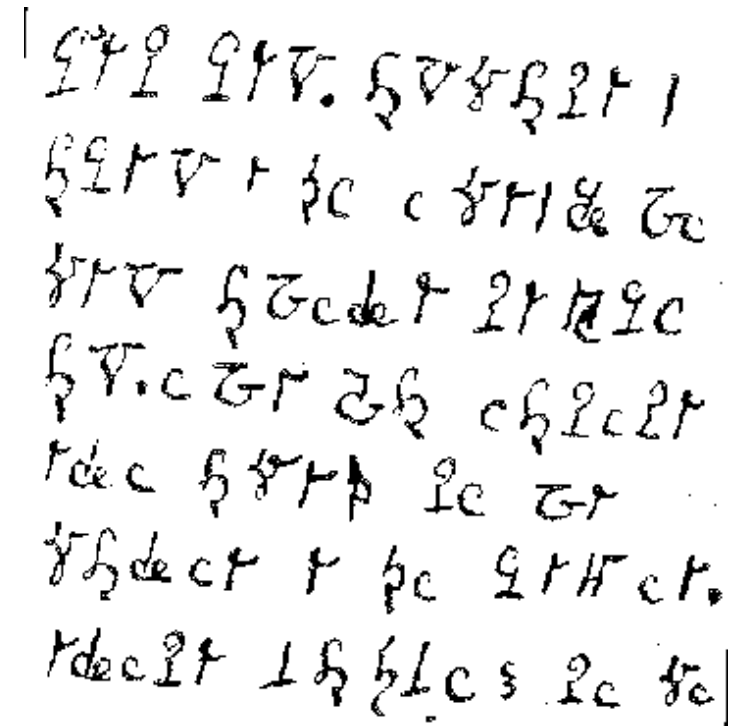
- Most famous American cryptographer of World War II
- VMS Study Group (1944-46)
 - developed transcription alphabet
 - group disbanded after the war
- 2nd VMS Study Group (1962)
 - at RCA
- His ultimate theory:
 - VMS written in a synthetic “philosophical” language



Writing in Tongues?

suggested in Kennedy & Churchill, 2005

- Historical example:
 - Medium **Helene Smith**, investigated by Theodore Flournoy (1896)
 - Under a trance, Smith was able to **converse with Martians**
 - She learned their language and could speak and write it
 - Grammar closer to French than you might expect



Handwritten sample of Smith's Martian script, showing a series of lines of characters that appear to be a mix of letters and symbols, written in a cursive style.

Smith's Martian

Hoax?

- Previous hoaxes:
 - Hitler diaries
 - Vinland map
- Carbon dating (reported 2011)
 - Done at University of Arizona
 - Four page fragments, each 1x6mm
 - Result: **1404-1438** (w/ 95% probability)*
 - Ink pigments consistent with that

Gordon Rugg Theory

(Scientific American, 2004)

- Proposed Cardan grille
- Elizabethan cryptography tool
- If applied with randomness injected, claimed to generate VMS-like text



Substitution Cipher

ingcmpnqsnwf cv fpn owoktvcv

hu ihgzsnwfv rqcffnw cw owgcnwf

kowazoanv . . .

Substitution Cipher

e e e e
ingcmpnqsnwf cv fpn owoktvcv
e e e
hu ihgzsnwfv rqcffnw cw owgcnwf
e
kowazoanv . . .

Substitution Cipher

e e e the
ingcmpnqsnwf cv fpn owoktvcv
e e e
hu ihgzsnwfv rqcffnw cw owgcnwf
e
kowazoanv . . .

Substitution Cipher

e he e the
ingcmpnqsnwf cv fpn owoktvcv
e e e t
hu ihgzsnwfv rqcffnw cw owgcnwf
e
kowazoanv . . .

Substitution Cipher

e he e of the
ingcmpnqsnwf cv fpn owoktvcv
e e e t
hu ihgzsnwfv rqcffnw cw owgcnwf
e
kowazoanv . . .

Substitution Cipher

e he e of the fof
ingcmpnqsnwf cv fpn owoktvcv
e f o e o oe t
hu ihgzsnwfv rqcffnw cw owgcnwf
ef
kowazoanv . . .

Substitution Cipher

e he e ~~of~~ the
ingcmpnqsnwf cv fpn owoktvcv
e e e t
hu ihgzsnwfv rqcffnw cw owgcnwf
e
kowazoanv . . .

Substitution Cipher

e he e is the sis
ingcmpnqsnwf cv fpn owoktvcv
e s i e i ie t
hu ihgzsnwfv rqcffnw cw owgcnwf
es
kowazoanv ...

Substitution Cipher

Pattern Words

abacdefb	ACADEMIC
abacdefb	DEDICATE
abacdefb	MEMBRANE
abacdefc	ELECTRIC
abacdefc	TUTELAGE
abacdefd	ANARCHIC
abacdefd	EVERYDAY
abacdefe	ANALYSES
abacdefe	ANALYSIS
abacdeff	EYEGLASS

e he e is the sis
ingcmpnqsnwf cv fpn owoktvcv
e s i e i ie t
hu ihgzsnwfv rqcffnw cw owgcnwf
es
kowazoanv ...

Substitution Cipher

Pattern Words

abacdefb	ACADEMIC
abacdefb	DEDICATE
abacdefb	MEMBRANE
abacdefc	ELECTRIC
abacdefc	TUTELAGE
abacdefd	ANARCHIC
abacdefd	EVERYDAY
abacdefe	ANALYSES
abacdefe	ANALYSIS
abacdeff	EYEGLASS

decipherment is the analysis

ingcmpnqsnwf cv fpn owoktvcv

of documents written in ancient

hu ihgzsnwfv rqcffnw cw owgcnwf

languages ...

kowazoanv ...

Computer Decipherment

Spanish letter trigram model

Train on Spanish web text.
Parameters fixed.

Probabilistic model that
substitutes VMS letters for Latin
letters. Initially uniform.

q u o _ v a d e _ b r e r t e _ ...

a → {all Voynich letters}
b → {all Voynich letters}
c → {all Voynich letters}
...
z → {all Voynich letters}
_ → _

EM Algorithm.

$$\operatorname{argmax}_{\theta} P(\text{VMS}) = \operatorname{argmax}_{\theta} \sum_{\text{latin}} P(\text{latin}) P(\text{VMS} \mid \text{Latin})$$

EM method demonstrated on many
decipherment tasks [Knight et al 2006].

V A S 9 2 _ 9 F A E _ A R _ A P A M _ ...

Substitution Cipher

Input	Best decipherment, assuming plaintext is Spanish
cevzren cnegr gry vatravbfb uvqnytb qba dhvwbgr qr yn znapun ...	primera parte del ingenioso hidalgo don quijote de la mancha ...
VAS92 9FAE AR APAM ZOE ZOR9 QOR92 9 FOR ZOE89 ...	decos acho es imen des dena denal y des denta ...

Or if plaintext is assumed to be Latin:

quiss squm is onum pom
quuss hates s qum hatis ...

80+ Plaintext Languages

Input	Best guess of plaintext language	Best decipherment
cevzren cnegr gry vatravbfb uvqnytb qba dhvwbgr qr yn znapun ...	Spanish	primera parte del ingenioso hidalgo don quijote de la mancha ...
VAS92 9FAE AR APAM ZOE ZOR9 QOR92 9 FOR ZOE89 ...	??? Romanian ???	nonsense

Consonantal Writing

Input	Best guess of plaintext language	Best decipherment
ceze ceg qy ataf uqyt qa dwg q y zapu ...	Spanish	prmr prt dl ngns hdlg dn qvt d l mnch ...
VAS92 9FAE AR APAM ZOE ZOR9 QOR92 9 FOR ZOE89 ...	more nonsense!	

Unsupervised Letter Clustering

Trigram model over {a, b, _}

← Initially uniform

a a _ b a b _ a b a a _ ...

a → {all English letters}

b → {all English letters}

_ → _

i n _ t h e _ t o w n _ w h e r e _ i _ w a s ...

Unsupervised Letter Clustering

Trigram model over {a, b, _}

Initially uniform

a a _ b a b _ a b a a _ ...



Sample tagging with learned model:

a b _ b b a _ b a b b _
i n _ t h e _ t o w n _

b b a b a _ a _ ...
w h e r e _ i _ ...

i n _ t h e _ t o w n _ w h e r e _ i _ w a s ...

Unsupervised Letter Clustering

Trigram model over {a, b, _}

← Initially uniform

a a _ b a b _ a b a a _ ...

a → {all Voynich letters}

b → {all Voynich letters}

_ → _

V A S 9 2 _ 9 F A E _ A R _ A P A M _ ...

Sample tagging with learned model:

? ? ? ? ? _ ? ? ? ? _ ? ? _
V A S 9 2 _ 9 F A E _ A R _

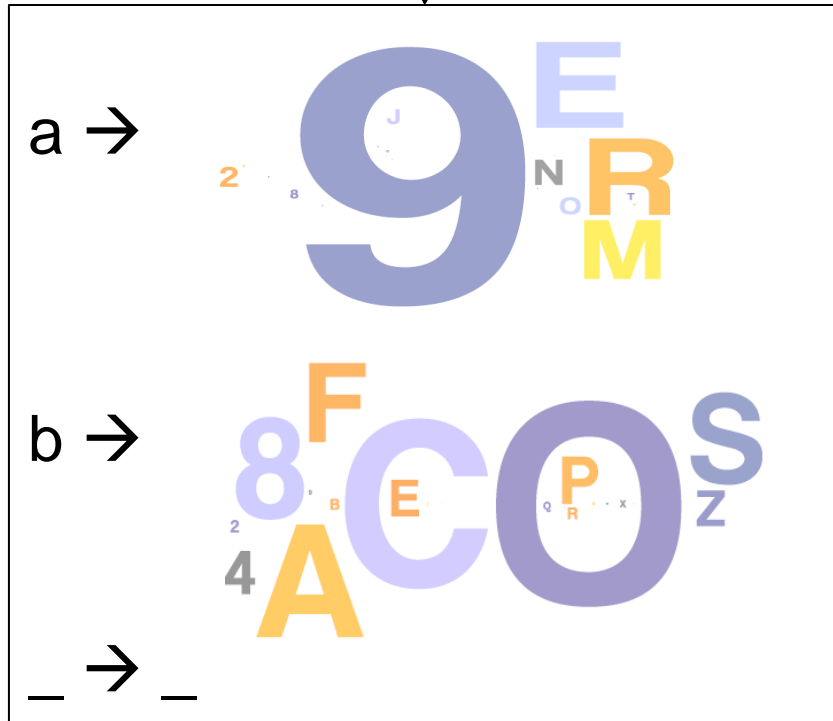
? ? ? ? _ ? ? ? _ ? ? ? ? _ ...
A P A M _ Z O E _ Z O R 9 _ ...

Unsupervised Letter Clustering

Trigram model over {a, b, _}

Initially uniform

a a _ b a b _ a b a a _ ...



Sample tagging with learned model:

b b b b a _ a b b a _ b a _
V A S 9 2 _ 9 F A E _ A R _

b b b a _ b b a _ b b b a _ ...
A P A M _ Z O E _ Z O R 9 _ ...

V A S 9 2 _ 9 F A E _ A R _ A P A M _ ...

Unsupervised Letter Clustering

$P(\text{letter} \mid \text{tag})$

English

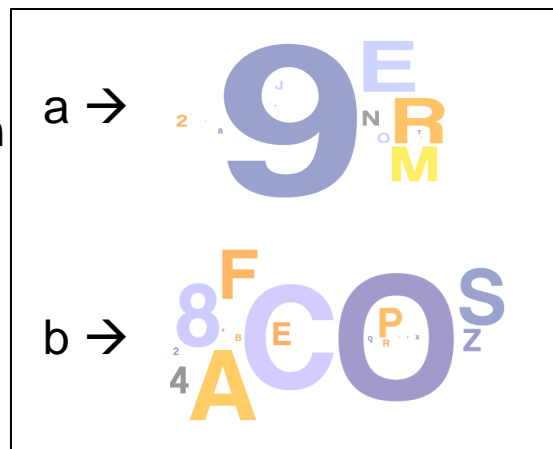


$P(\text{tag} \mid \text{letter})$

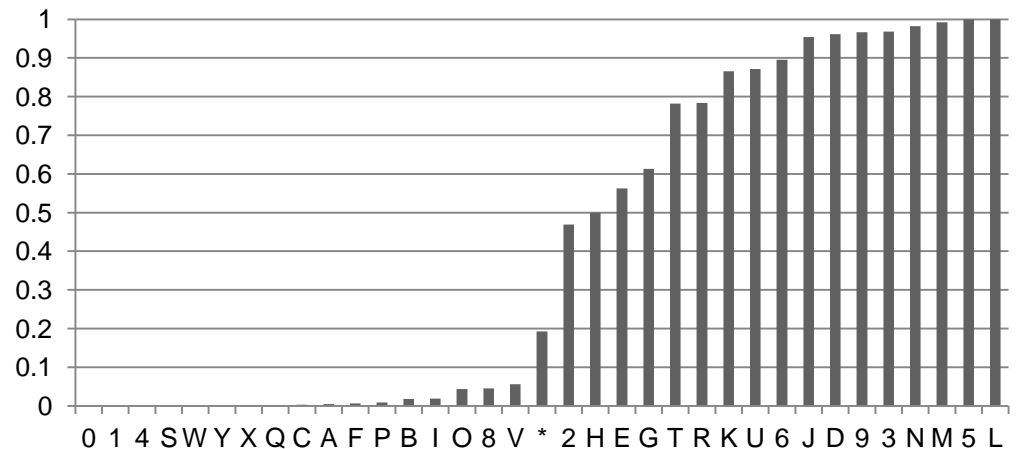
$P(a)$



Voynich



$P(a)$



Approved for Release by NSA on
06-03-2009, FOIA Case # 58742

An Application of PTAH to the Voynich Manuscript (U)

BY MARY E. D'IMPERIO

~~Top Secret Umbra~~

(U) This article is the second in a series of studies applying some modern statistical techniques to the problems posed by the Voynich manuscript. This study attempts to discover and demonstrate regularities of patterning in the Voynich text subjectively noted by many earlier students of the manuscript. Three separate PTAH studies are described, attacking the Voynich text at three levels: single symbols, whole "words," and a carefully chosen set of substrings within "words." These analyses are applied to samples of text from the "Biological B" section of the manuscript, in Currier's transcription. A brief general characterization of PTAH is provided, with an explanation of how it is used in the present application.

program. He was struck by the passage "immenso Ptah noi invociam," and named his program after the Egyptian god. The name was ultimately extended from this program, implementing a particular application of the method, to the method and its mathematical theory as well [2, p 85]. According to [] of R51, the name is pronounced "however you like" [8]. The technique itself and its uses are classified Top Secret Codeword.

[] I chose PTAH for the present study for two main reasons: first, because of the applications of PTAH to book codes, and second, because I wished to learn more about PTAH itself []

1970s NSA report
recently declassified!

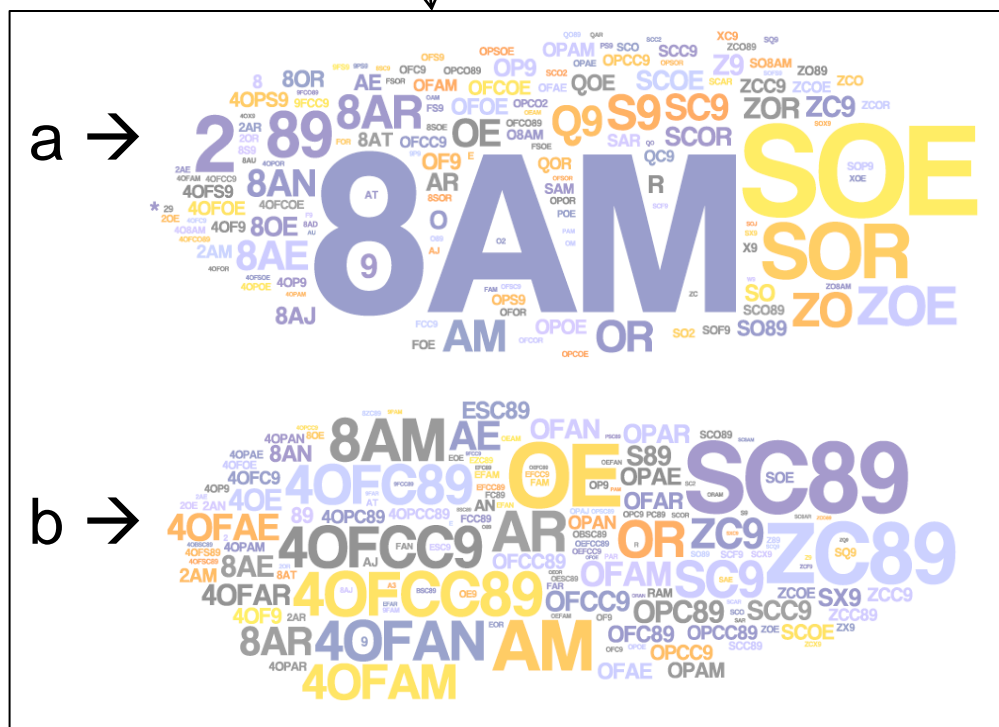
Word Clustering

Bigram model over {a, b}

a a b a b a b a a ...

Do words with similar contexts have similar spellings?!

That would be very interesting.



VAS92 9FAE AR APAM ZOE ZOR9 QRC2 9 ...

Word Clustering

Bigram model over $\{a, b\}$

a a b a b a b a a ...

Do words with similar contexts have similar spellings?!

That would be very interesting.

Sample tagging with learned model:

a	a	a	a	a	a
VAS92	9FAE	AR	APAM	ZOE	ZOR9

a	a	a	a	a	...
QRC2	9	FOR	ZOE89	2OR9	...

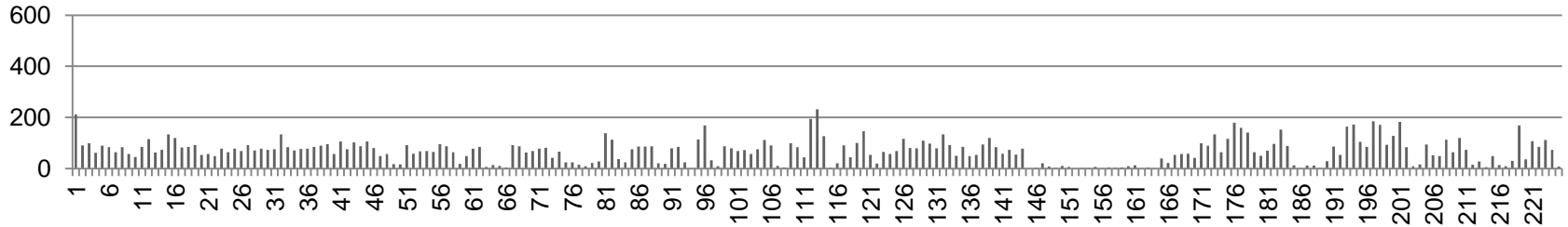
WAIT, WHAT?

VAS92 9FAE AR APAM ZOE ZOR9 QRC2 9 ...

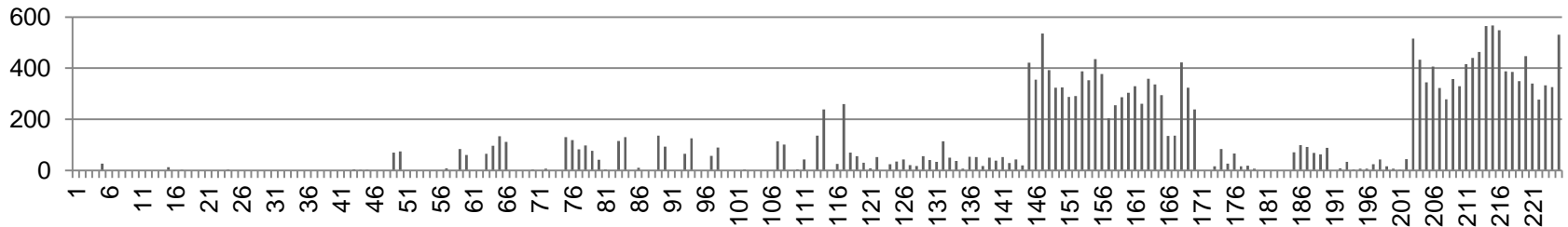
Word Clustering

Voynich words tagged as “a”

← pages →



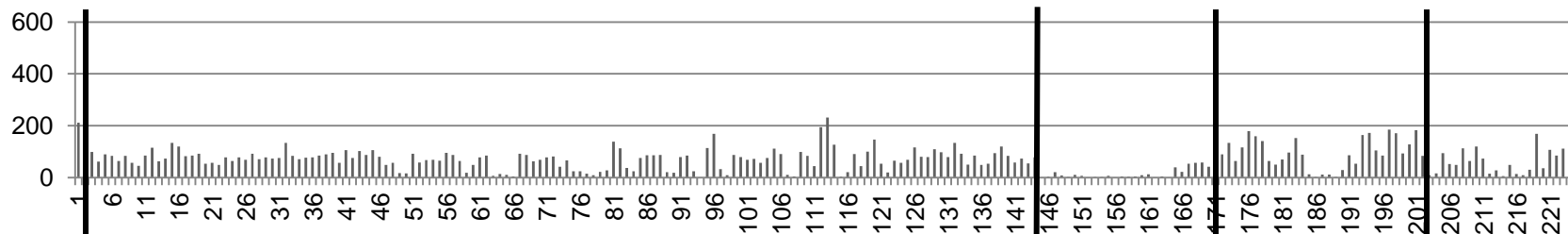
Voynich words tagged as “b”



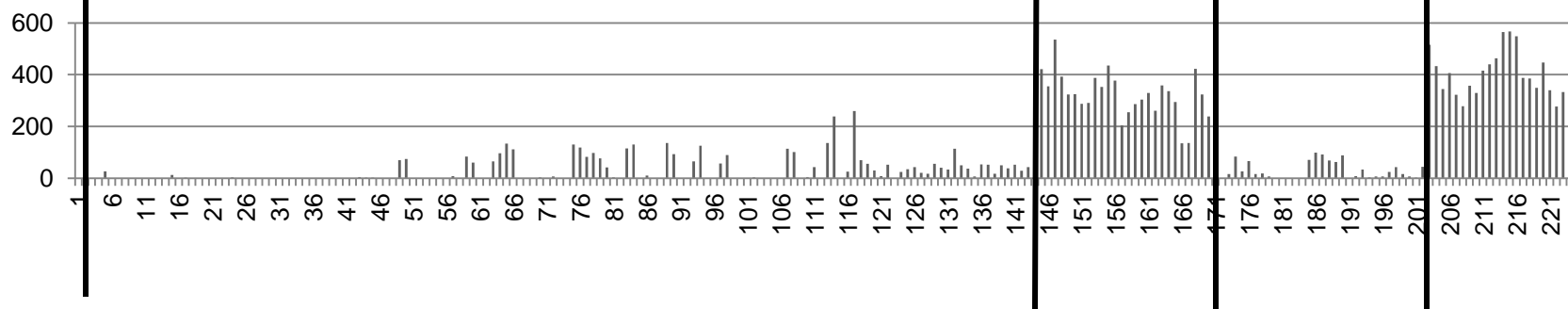
Word Clustering

Voynich words tagged as “a”

← pages →



Voynich words tagged as “b”

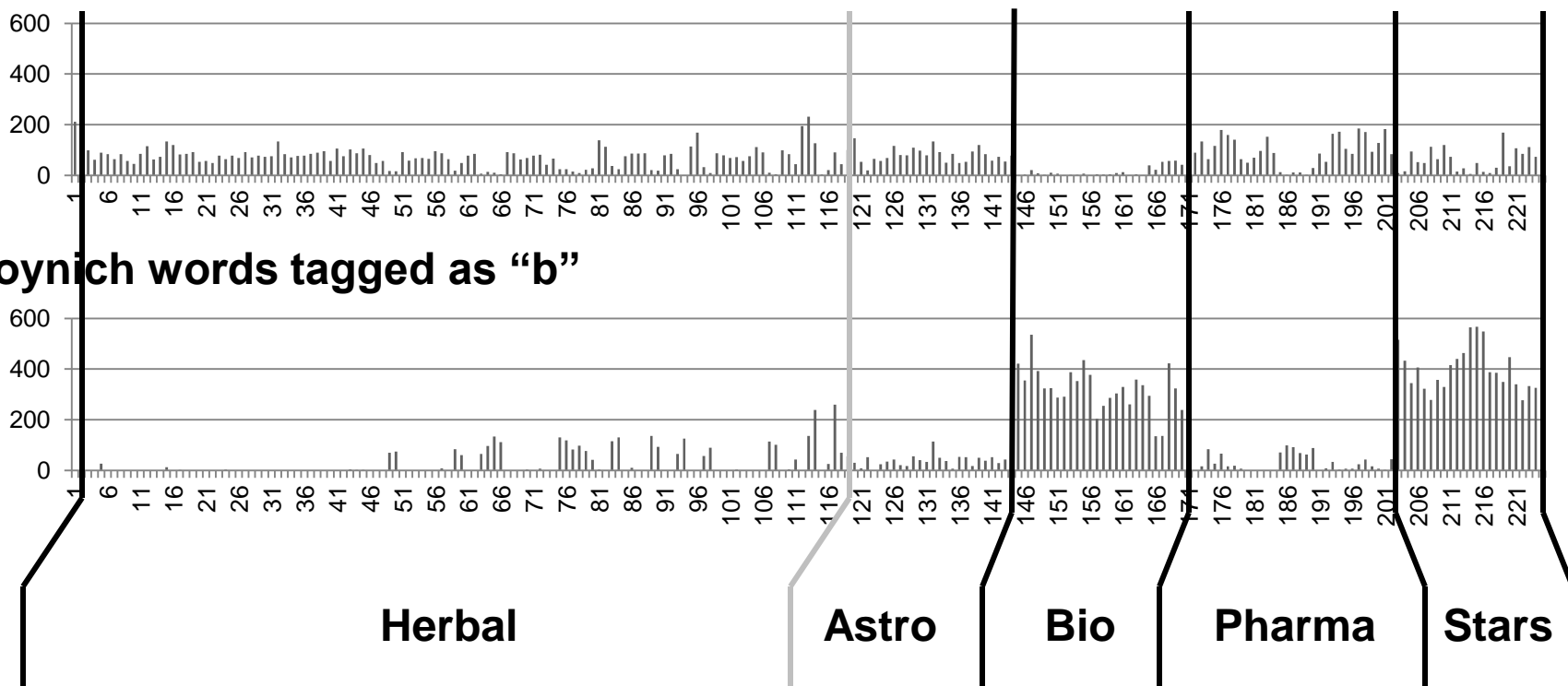


Word Clustering

Voynich words tagged as “a”

← pages →

Voynich words tagged as “b”



Known since Capt. Carrier's analysis (1976)

Statistical Analysis of Voynich B

- Since there are two distinct languages, we focus analysis on one of them (Bio and Stars)
- Ask some basic questions about the “linguistic” structure
 - The Page
 - The Line
 - The Word

THE PAGE

Can we identify content words and function words?

Some pages are missing, but the current version comprises about 240 vellum pages, most with illustrations. Much of the manuscript resembles herbal manuscripts of the 1500s, seeming to present illustrations and information about plants and their possible uses for medical purposes.

Can we identify content words and function words?

Measure the saliency of a word in a page with TF-IDF

$$\text{TF-IDF}(w, d) = \text{TF}(w, d) \times \log \frac{N}{\text{DF}(w)}$$

times that word
 w occurs in page
 d

pages that
contain word
 w

Can we identify content words and function words?

OFCC9 ZC89 8AN ZC9 SCFCC89 ROR 4OFAN SAE FAN ZCCF9 EPAN OR 9FCC9 EZC9 FAE 4OFCC89 FAR SC9 R AN ZCC89 4OFCC9 4OF9 FCCOR 8AR OFC9 SQ9 BSC89 SX9 EZC89
OFCC89 4OFC89 OESC89 2AE S89 OPC9 ESC89 4OFAE OFAR FCC89 2AN OPAR 2AM OPZC89 BSC8AR 8AE ZAE OPAN SAR AR ESC9 EOR AJ OEFCO9 SCQ9 ZX9 OPC89 BAT ZCQ9 E
SFAE 4OFC9 SCF9 SQC89 EPC89 OFAE OP9 4OFC89 OPCC89 4OPCC9 4OPCC89 OFC89 OPAJ 2ZC89 2 O 8AM EFAR 8E ES89 FC9 SCC89 8AR9 ZCOR 4OFCOR AM ZCC9 SE FCC9
4OPAN OFAN 4OPS89 OPCC9 SXC9 EFC89 4OFCCO AE SC8AN EFCC89 ZC8 SOE OEAN ZC8AE 4OFAM 9 SCOE EFC9 4OBSC9 OFS9 SCAE AEOE ZCAR ORAM
4OXC9 RAM OFC8AE SCQC9 4OPCC2 ZC8AJ OFCCC89 4OXC89 SCXC9 OPCOE 2AEFCC89 PCC89 SXC89 PCC9 ZCFCC9 ZCCOE ZCAE ON 4OFCAR SR 8AJ
EAJ 8ZCC89 9SCC9 ZAR SQ*9 EOFCC89 ER SXAE 4OPSC89 ESCOE SC8AR ORAN 4OPSC9 AT OIF*9 4OSC9 OEFAE 8ZC9 PAR EFO EFCC9 ZCO AEOJ FCCOE
4OFCSC89 08AM ZCFAE 8SC8 4OVSC89 SO89 4OFCOE ESC8 SFCC9 8SCC9 EOFAJ 4OF OFCCOE SCBSC89 8OM ZCCX9 PC9 OPC8AR OPC2 4OFC89 OPCO89
SCO89 4OFCO89 OPCC8 AK EVS9 SAM PAT CCC2 OJ EFAJ FCCC9 SCO2 8AK ZCCF EOP9 ZCFAN 4OFCCS9 OFCAE EFCCO89 ZCO8 ZC8AN BSAE OPCAE
OPCCAJ 4OAN 4O8 SCO SCO SCOFCC9 AEAJ OFCCO EFCSC9 EFCSC89 SCCFAN OPCCC9 EFAT ESAE OPCCOE SO FCSC89 OEFCO89 4OFCO8
4OSC89 OPCO ZCCO8AR SCOJ OPARAE OAM OEFCO EFCCOE EFAE EFCCC9 PC8AJ OFC8AN ZOF 4OFCCOE 4OFCCO2 OPCC8AN EFCCC89 SCAJ
SCXC8 OSC9 FSOES8AR AIIB SCPAEZ9 4CCAE 4OFCCA2 SCOFCC9 ZCCFZ9 SOFSC9 SX*9 9SCCO8AN 9FCC8AM RCCC9 OEA3 AIF*9 ZFAM 8ZCCO OPSC8C9
OPCOEAT 4OZCO 8SC8AR 9FCCO2 BSC8C9 OPCO8O BS8AJ OFCO8AN ZCP ZCOPAJ ZPAR OESCO89 ZCQC9 ESCOCFAJ 4OFCCO8OR SCQ89 4OPC8E
EOC89 SC8C9 EFAK O*OR F98CC89 ZCCP SCOP989 2ZCO PSC8 FCOQC89 4OFCZC9 FCCZO89 OFCSC89 ESR 2O AEAE O*AR 2OAM OPCO8AM
4OPCC8AM PCC8AN ZCOFAR RF9 SPAR 4OTAN ZCOPSC89 ZFC9 4OFCAN ZFCO89 8ZCCOPCC9 PCAR SOEFCCC89 ESCS89 4OCCCO SC8A FCCO8AE
PCO OFCOJ 4OFCCO8 EFC8EFC9 PCC8 9SC8E BOEAE B9FCOR SCCV9 OBSC8AE EVSC89 ESCCOE OPON SCAJAR 4OFCOFC89 OPCCOEFC9 EAN
4OFCCAE 4OFCCE SC89PCOFAN EFC8AR EFCC8AN OFCAJ PCOEFC8AN EPCCAE U OFCCC2C9 OESAE 4CCAR BOCOFCC9 PC8AN SCBSAJ 4OFCCOFAN
FCCE BOEFCCO SCOFAN I*AR *AN 9ZC OFZ89 ZFCC9 SXAM

Can we identify content words and function words?

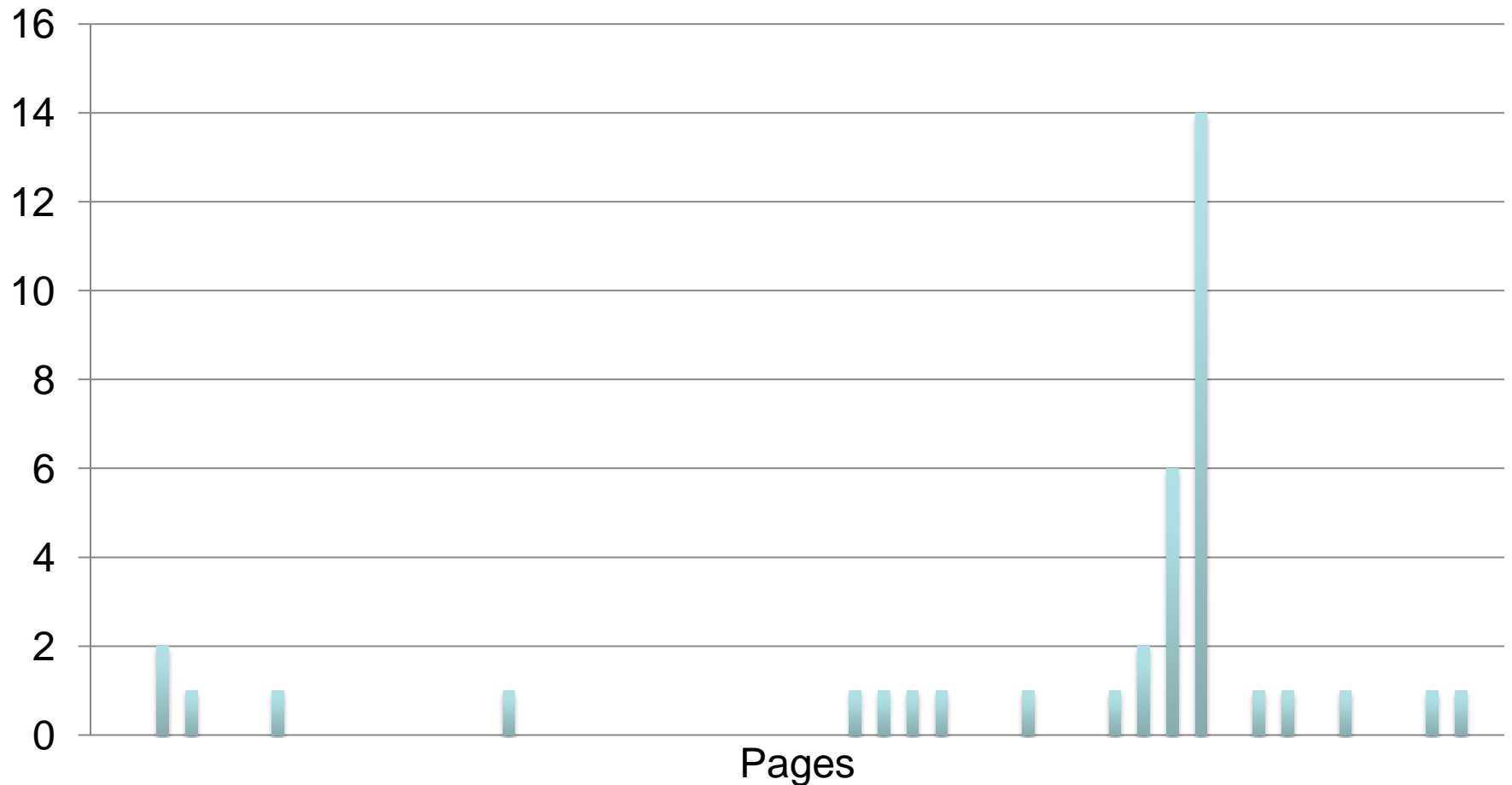
As a control:

Scramble words across text and repaginate

OFAN SCC9 8AN 4OE AT OPC89 4OPAE AN ZC9 8ZC89 4OPAN AM ZQC9 PC89 QC89 AJ SCAR OFAR 2OE ZCC9 OFAM 4OPCC89 8AR 8AE OPC9 EFAM 4OFS9 AR9 OR 4OFOR SX9 OFCC89 4OF9
4OFAM OPAM ZCC89 OEOR AE 4OFC9 FAM 4OFC89 OPOE ZCO8 EO 4OPCO89 A3 OP9 OF9 4OPC89 SC9 OFCCC89 89 R9 OFS89 SO89 OFCC9 RAN OB9 OPAJ 4O 2AM 9FCC89 BSC9
4OPSC89 ZX9 8AM ZCF9 EFCC89 BAT 4OP9 O8AM SOE ZCOE OPCC9 OFAE 4OFOE EAR 4OPAM S8AR 9FAR 4OE9 SCO 2AR S89 SE SCOE OPCC89 S9 SCO89 4OFAR EAM OFC9 SCC89 OBAM
ZOE FCC8AE OPAR SC2 FSC89 4CC89 2AN FC89 OPS89 OEZC9 SCB9 OEFAN 8SC89 Z9 ESC89 8SCOE OPAI2 ZCX9 OBOR SCCF9 ZO8AM ZC8AM OVSC89 SCCO 4OFE 9PAR
4OFS89 FAN OFC8AR SR SAE ZC8AE OESC89 OESC9 RAM ZCO PAR 4OFSC89 4OFCCO OFCOE OE9 ROE O8AR 4OFCCSC89 PAM OPCOE AU 4OBSC89 FAR SO8AM OFCCO OEFCC89
SCO2 ESC9 ESCC9 4OFCOE 9FCC9 SOR OPZC89 O9 PSC89 AESC89 OEFCC2 S8OE ZAE 9POR OFCCO89 4OFCCC9 8A3 CCC2 OFSOR BAM OBS89 4OBSC9 SC8 OR9
PAE SCF9 ZC2 9ZC9 ZP9 4OPO89 4OSC9 SOFC89 EFE ORAR EFCC9 OJ AEFAN SOAM OPS9 O2AM PAT 4OESC89 8ZCC9 SEAR OFS9 RAJ SC8AM OQ89 FZ8
SCFCC89 FCCO89 OFZC9 OFCCC9 4OPZC89 ZO SFAN OPCO 4OPAJ 4OPCS9 9SC8AR E89 ZCP9 OEAJ 2 SQ*9 4OCF9 ORAJ 9BSC89 ESOR SPOE 9PAE FE
ORSC9 8AR0J 4OVS89 OPCAE S8OR BS8AJ SXAE 4OBSCC9 98AM ORAE ZCFAN 9FCCC89 OFCCZ9 FC9 SC9F9 FS89 OESCC9 EFCCO 9SCC89 4OFC8AM
EFO OPSC2 OEFSC89 EZCOE OBZC89 ZOIF*9 ZCFAE 4OFAE89 ZCW9 EPCC9 ESCOR 9SCCO 8SCC89 4OC8AM OPATOR 4OEZC9 EFCZ89 PSC8AM SCO9
ESCOCFAJ CCO8AM POEFAE 9ZCAE OEOROE 4OSCOE 4OAT 4COFCO89 9FARAN 89S9 AEZ9 9FSO OFC9P9 4OFA3 4OFAK 4OPO9 WOS9 9BO8AM
4OFCCAR 8A8AJ 4OFS289 PARAT 2SCO SCOE89 OFSAM Z29 SBAE 2AE8AJ EFSCOE ZCSOE 9Z8AN OPSCCO 4OPCCO WCO BSCC9 8A12 E8AR AROPCC9
4OPT 4OF9E889 AR9E9 OEZ9 89PZC89 9SC8AN 4OFCZ89 4CFAE CC2C9 ZCBSC89 OCC2AM ZOOR 4OQCO89 SAR9 FSC RF9 9SCOEF 4OFCC2 4OFAO9
8AEAE 4OEF9 EEOR9 POEZ9 4OEAN SCQ89 4OPO2 BSCAJ ZCCO8CC2 4OCCS89 4OFCC8AT ORSQ89 OFCOXC9 SCCAE 9FCO8AM 8ATOE 9PCC8AR 8CC9
2SC9 EZCO89 EO*C89 OFSCCV9 RAE8E RCCC9 89AT S8AT PORAN 2O8AE OZCC89 OBZC9 BORAE OROR9 ZOX9 8ARS9 SCPAN 8OESCC89 OQOR 4OXC8
AFAlIF9 AEO8AR ZOY9

Do the content words indicate topics?

Frequency of EFCC89 in Voynich B pages

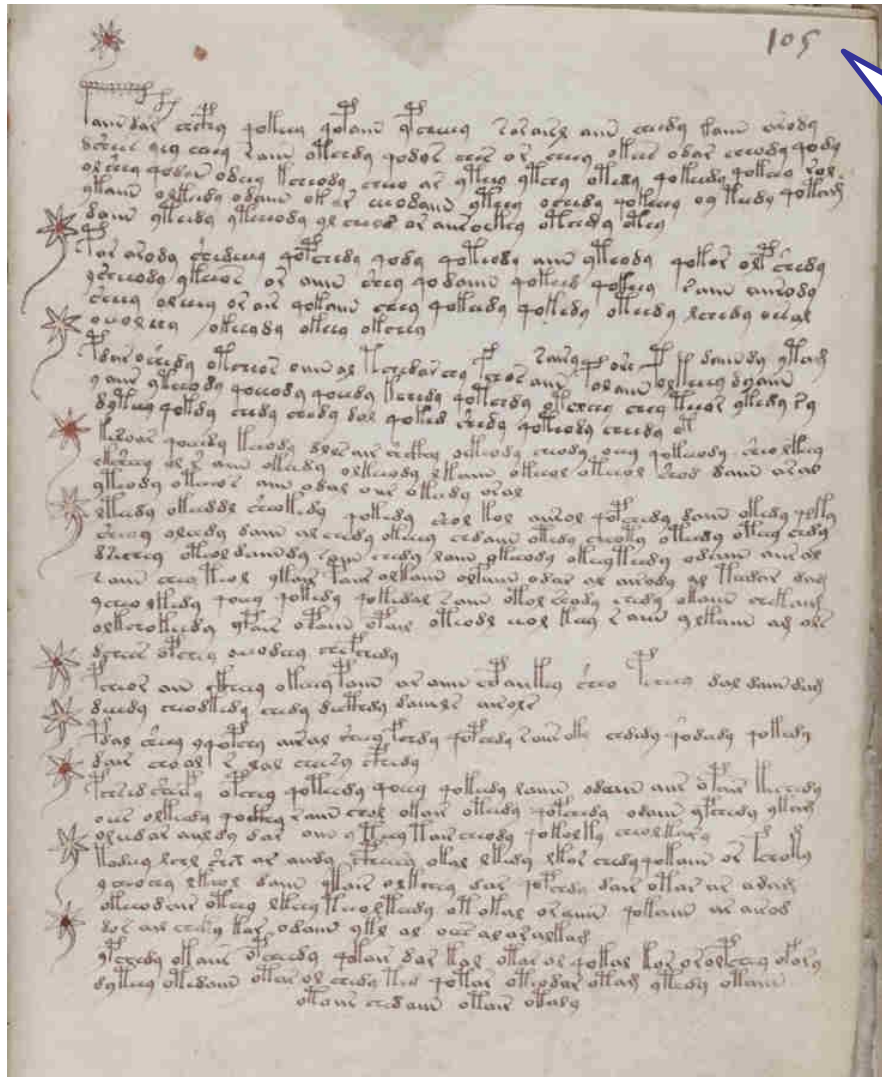


Content words and function words

Conclusion:

Some words are “bursty” within a page or groups of pages,
other words are more uniformly distributed
across the manuscript

Are the pages in order?

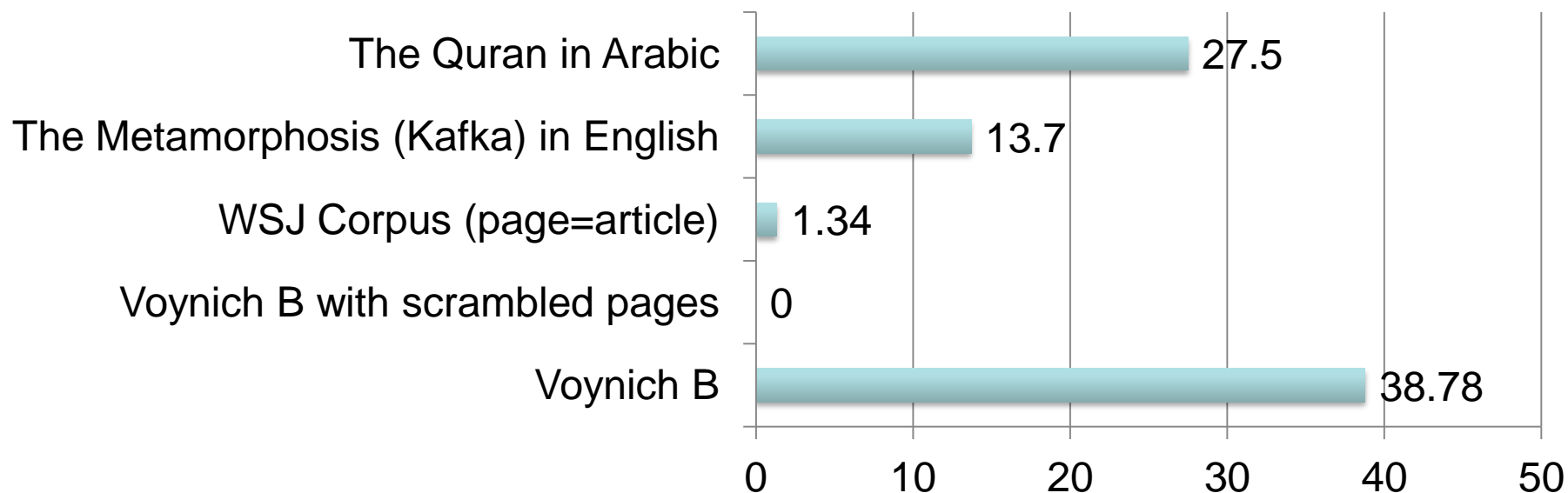


Numbers were added by later owners.

What can the text tell us about the manuscript's page order?

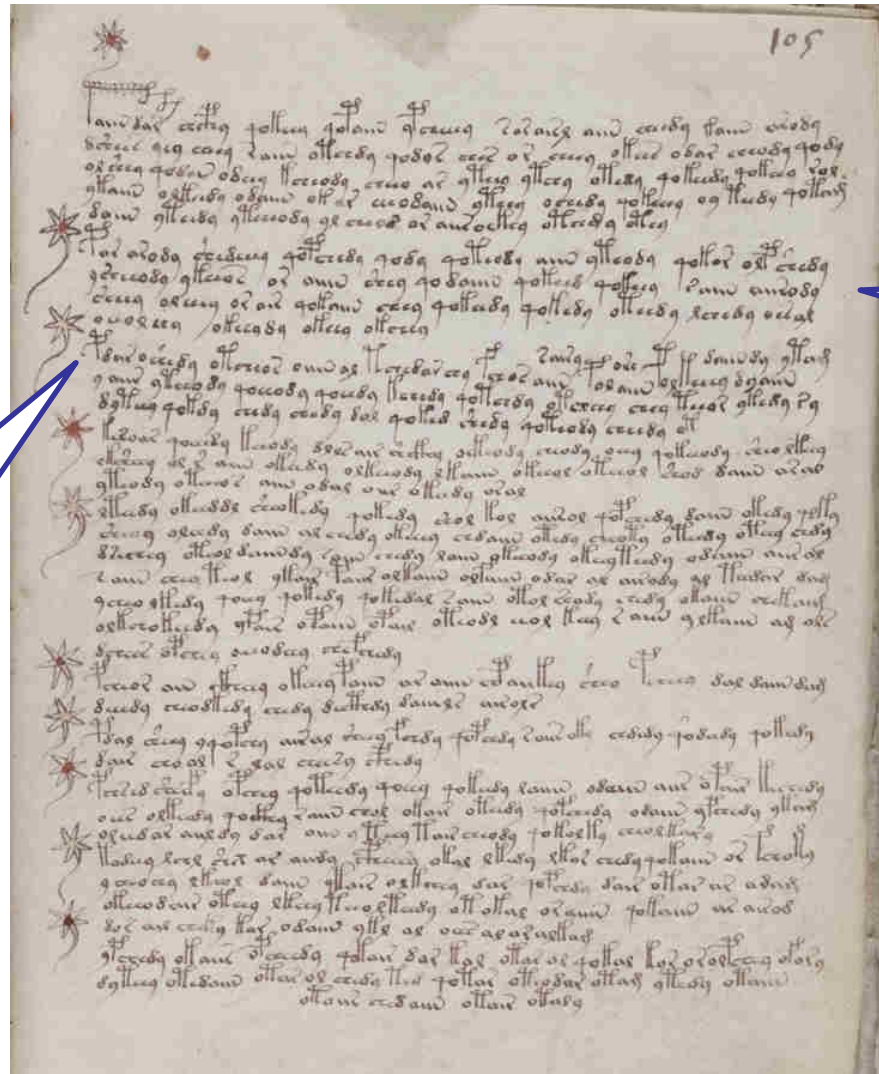
Are the pages in order?

- Measure similarity between a pair of pages using cosine similarity (with bag-of-words)
- Count the % of pages P where the most similar page to P is adjacent to it



THE LINE

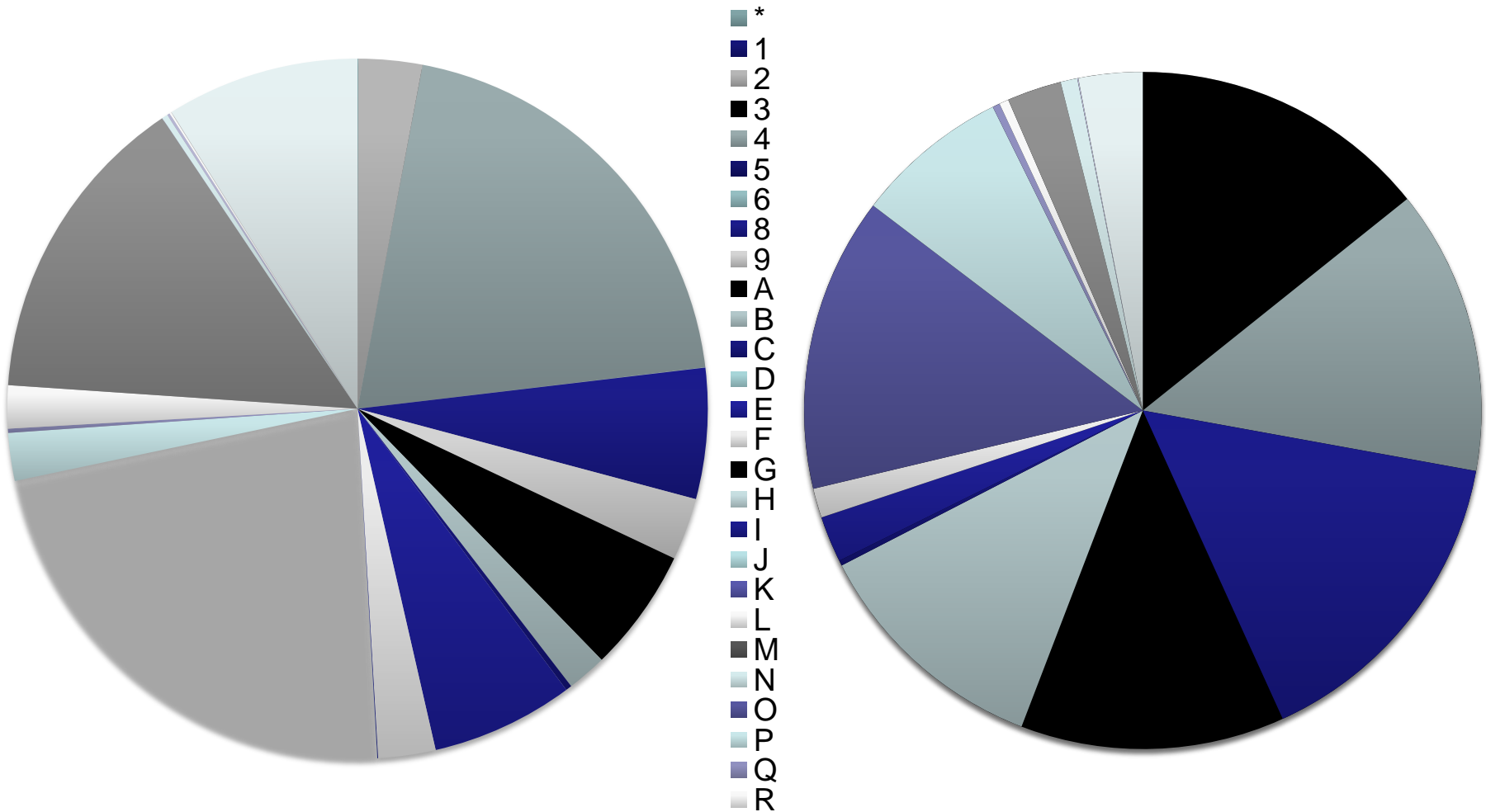
Is the text prose?



Special
ligatures at
beginning of
“paragraphs”

Paragraph
structure

Is the text prose?

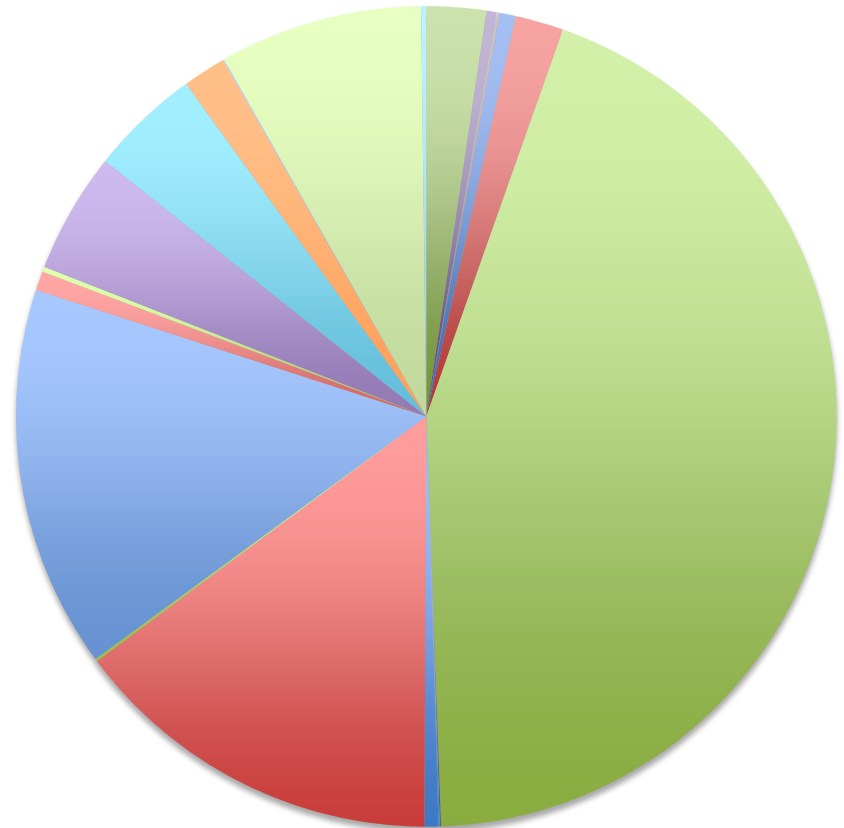
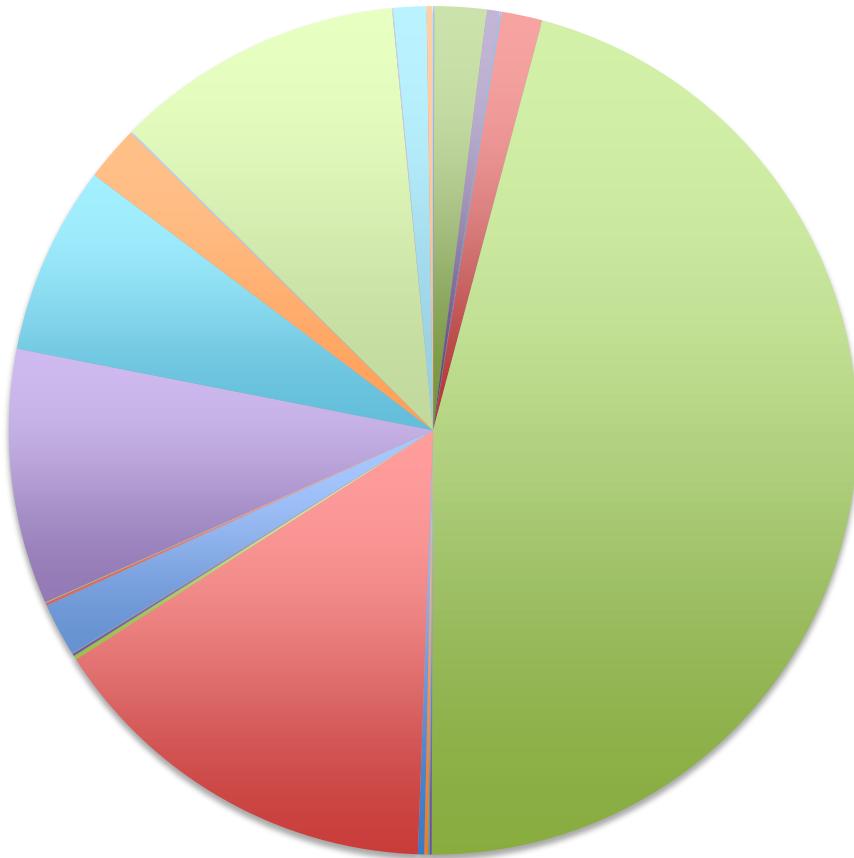


Characters that begin a word token

Characters that begin a line

Is the text prose?

*
1
2
3
4
5
6
8
9
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
R



Characters that end a word token

Characters that end a line

Is the text prose?

Conclusion:

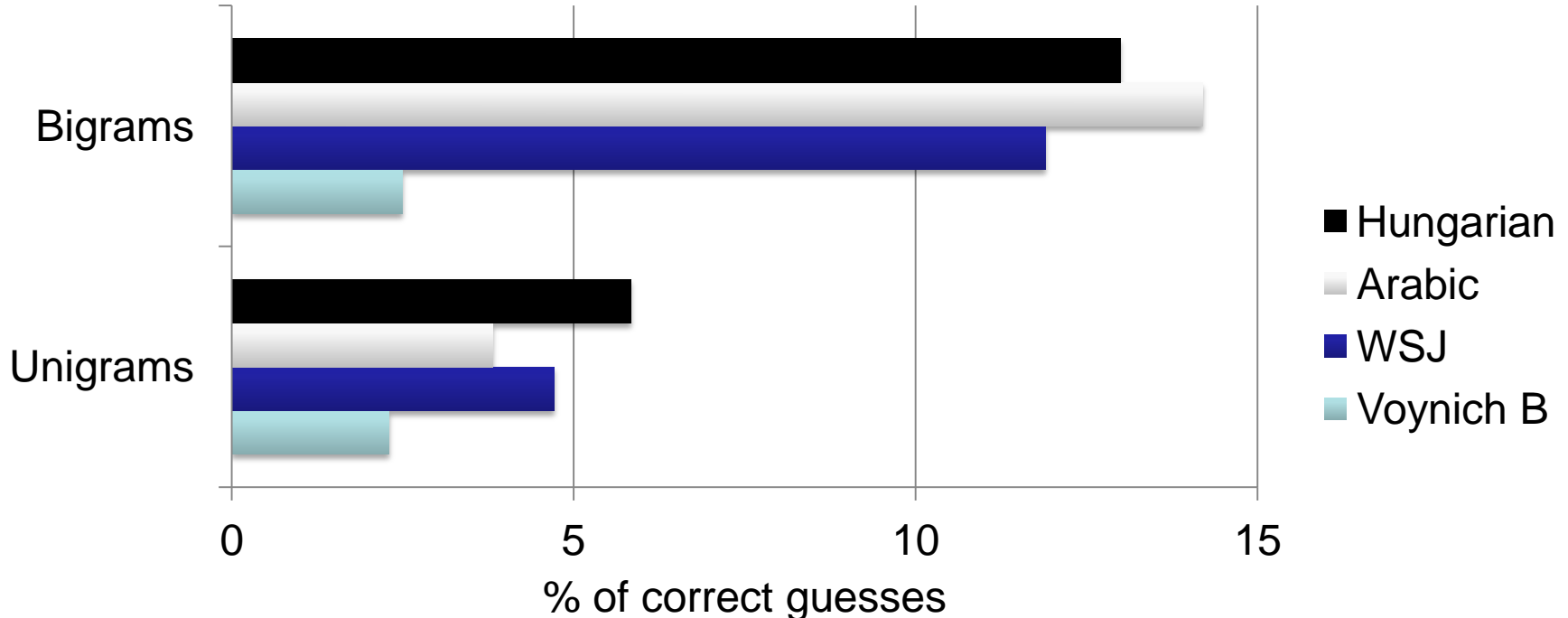
Lines begin and end disproportionately often with certain characters

The line is a
functional entity...



How predictable are word sequences?

- Guess most likely word to follow current word
 - Simulate game from bigram probabilities
- 90-10 train-test splits



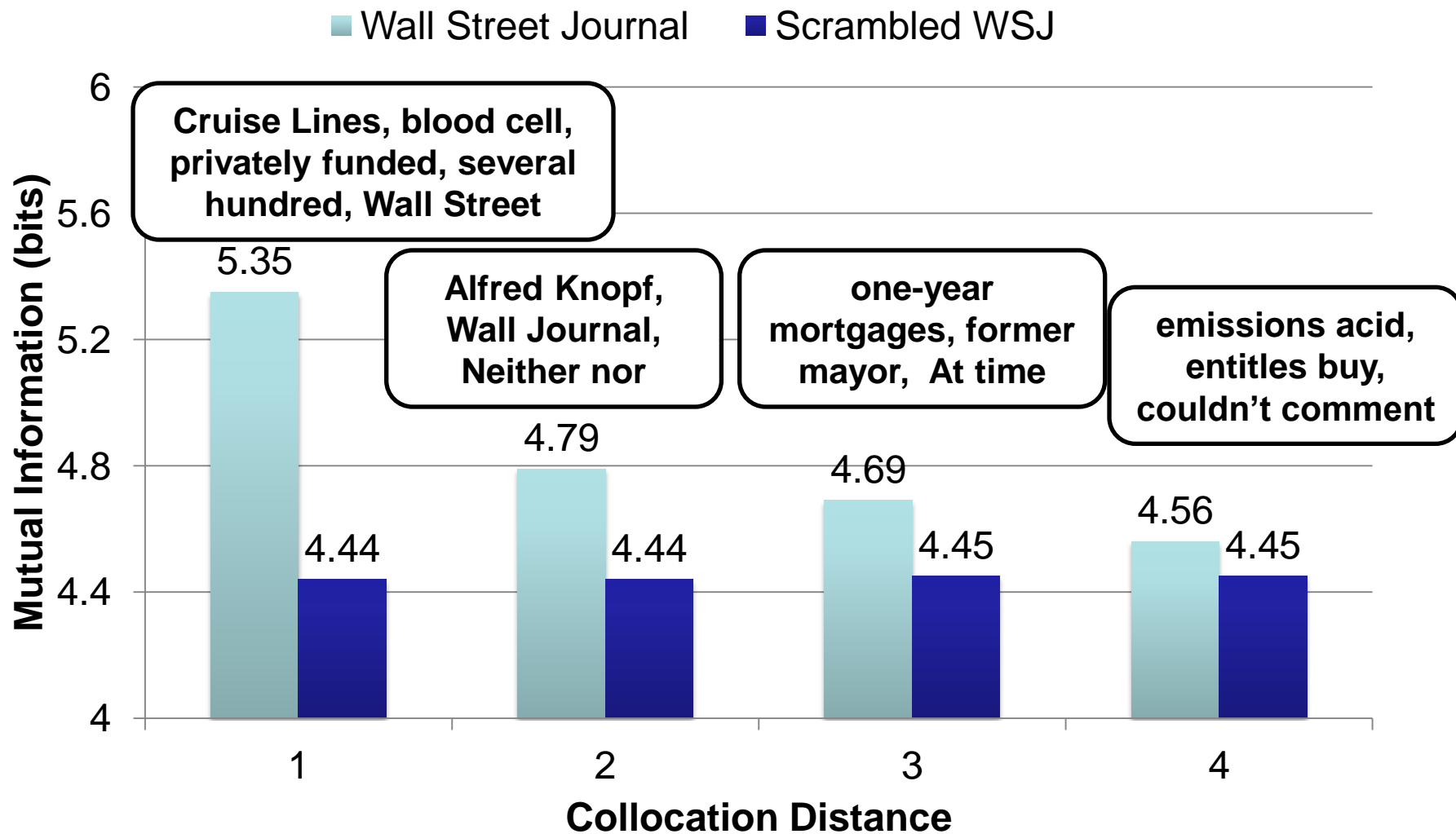
Are there collocations?

Measure 'collocationness' with mutual information

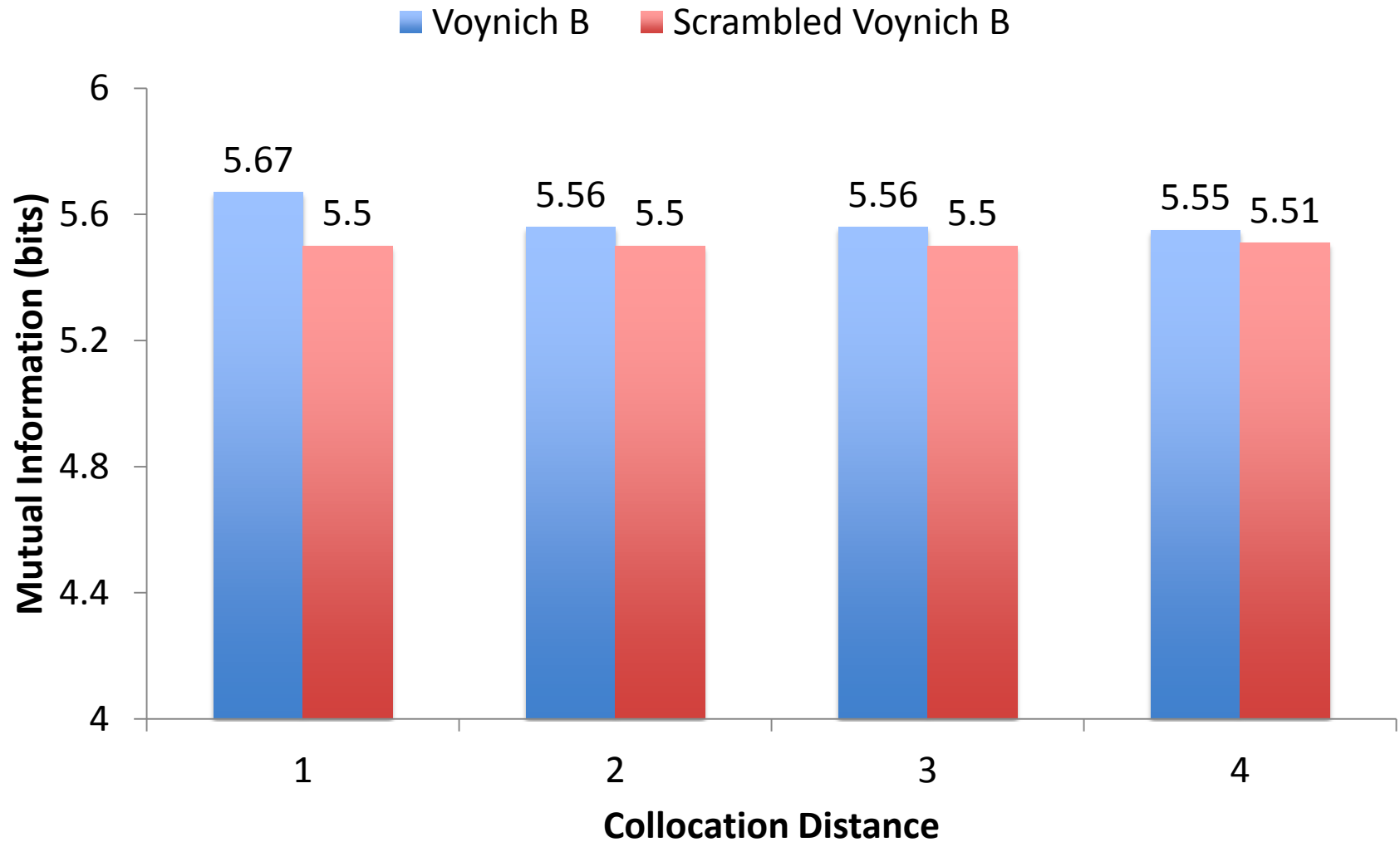
**Probability of word x
and word y co-occurring
at a fixed distance apart**

$$\sum_{x \in X, y \in Y} MI(X, Y) = P(x, y) \log \frac{P(x, y)}{P(x)P(y)}$$

Are there collocations?



Are there collocations?



Are there collocations?

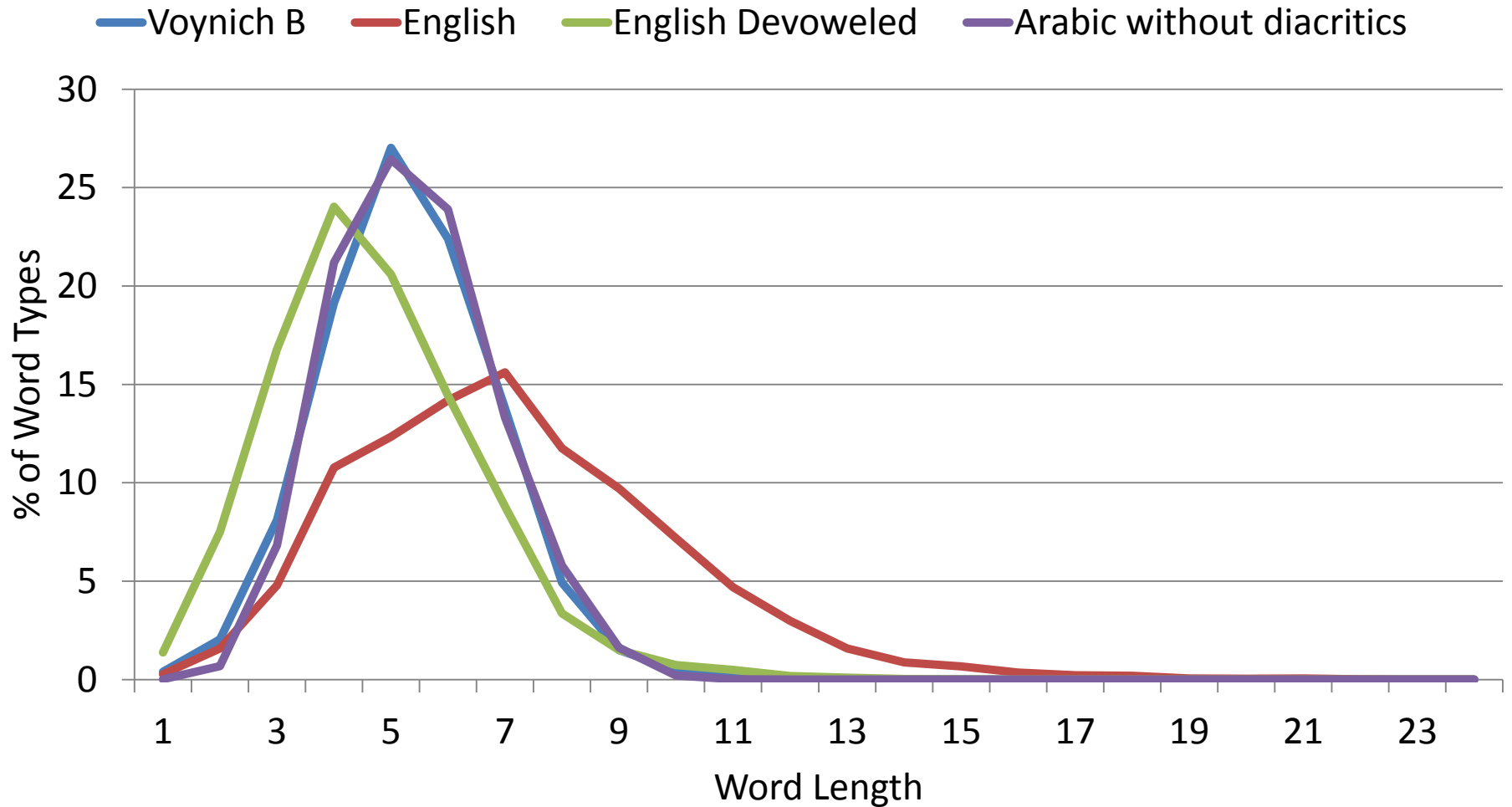
Conclusion:

Weak adjacent-word collocations

No long-range collocations

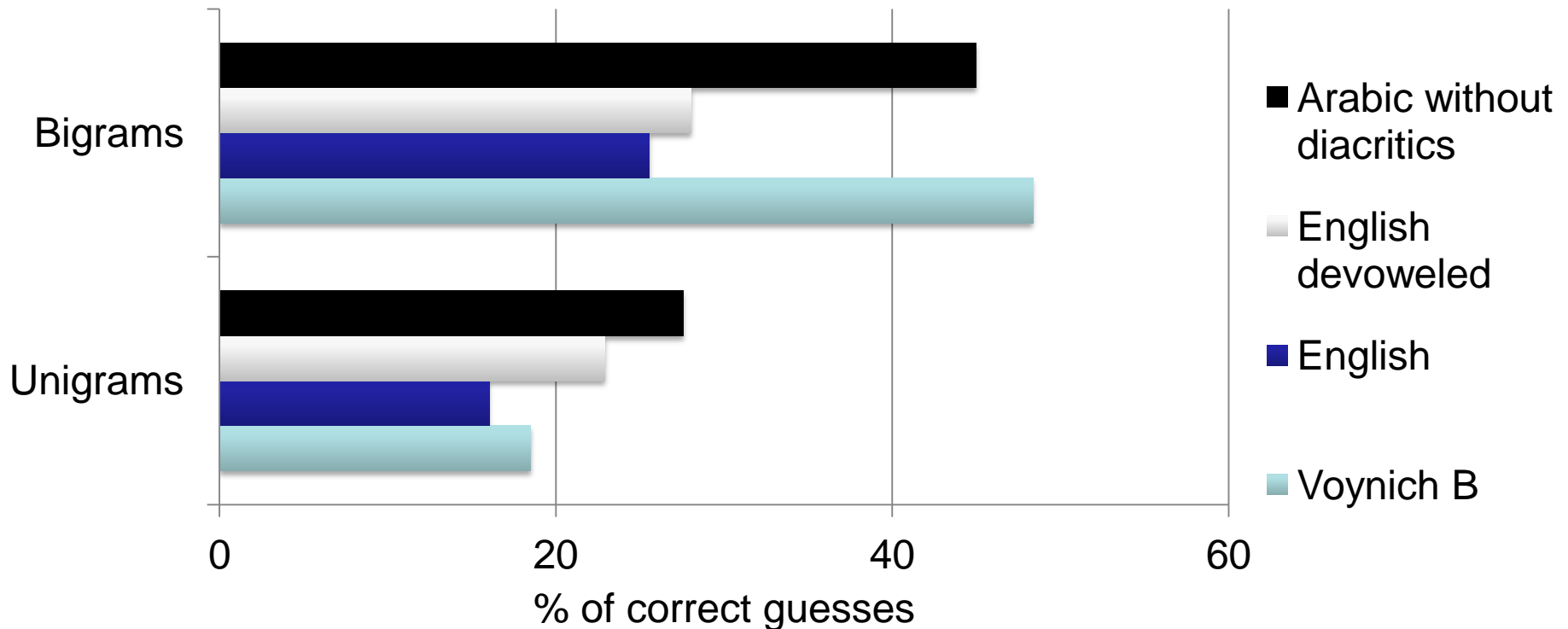
THE WORD

Word length distribution



Word-internal structure

- Repeat predictability simulation with letters (including spaces)



Conclusion

- Voynich Manuscript
 - What it is → pretty clear
 - Where it came from → still many theories
 - What it means → totally unclear
- Lots more room for computer techniques:
 - Transcription
 - Determining relations between words and pictures
 - Identification of “topics”
 - More unsupervised pattern finding

thanks!

backup

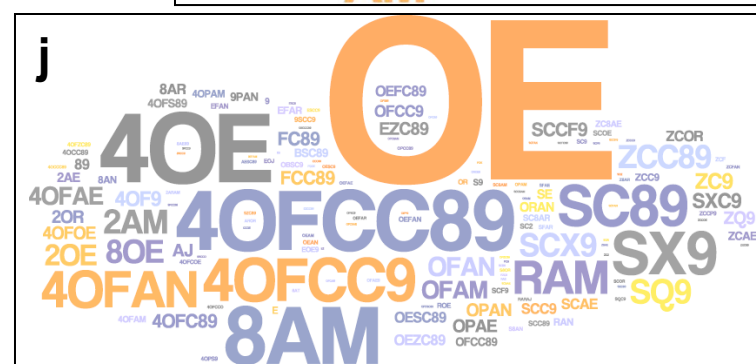
[illegible]

etc

etc

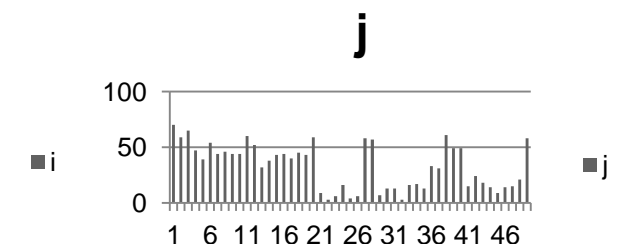
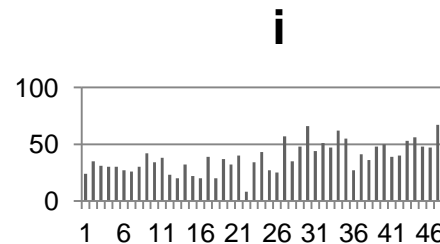
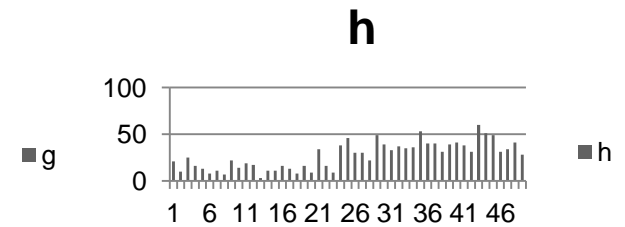
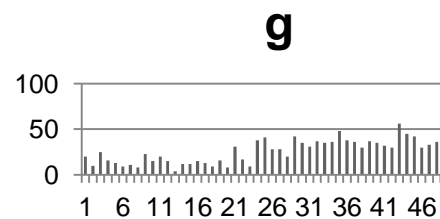
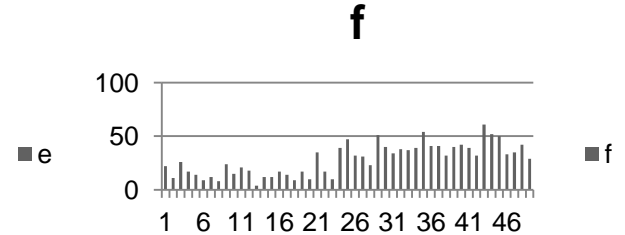
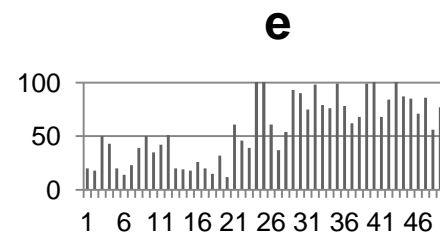
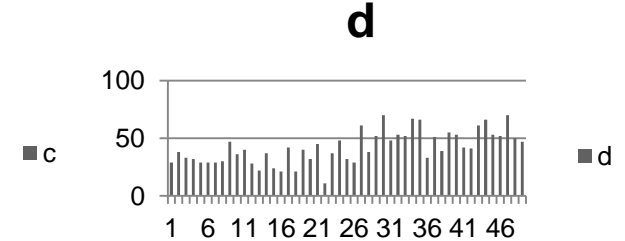
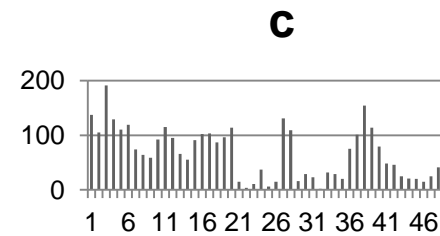
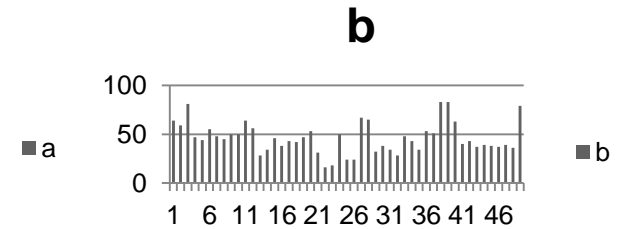
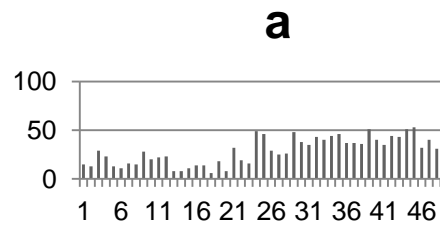
[illegible]

etc



10 clusters: Voynich-B

Tags per
page.



10 clusters:
Voynich-B

Tags per
page.

“Bio” words vs.
“Stars” words

