

the Voynich manuscript

Kevin Knight

Information Sciences Institute
University of Southern California

Sources for this talk:

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.voyrich.nu/>

<http://www.voyrich.ms/forum/>

experiments at USC/ISI

Some people involved with the Voynich manuscript



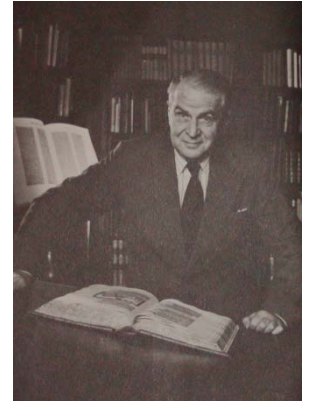
Wilfrid Michael Voynich
book dealer



Ethel Boole, daughter
of George Boole



Rudolf II
Holy Roman Emperor



Hans P. Kraus,
book dealer



Roger Bacon,
“first scientist”



Athanasius Kircher,
German Jesuit super-scholar



William Newbold,
Polymath, PhD UPenn



William Friedman,
WWII cryptanalyst

Outline

- Voynich Manuscript – VMS, for short
 - What is it?
 - Where did it come from?
 - What does it mean?

What is it?

- Medieval illustrated manuscript
- Approx. 235 pages on vellum material
- Color drawings of plants, nymphs, stars, etc.
- Approx. 38,000 words written in an unknown script
- Undeciphered!!! Meaning is unknown
- Currently owned by Yale University

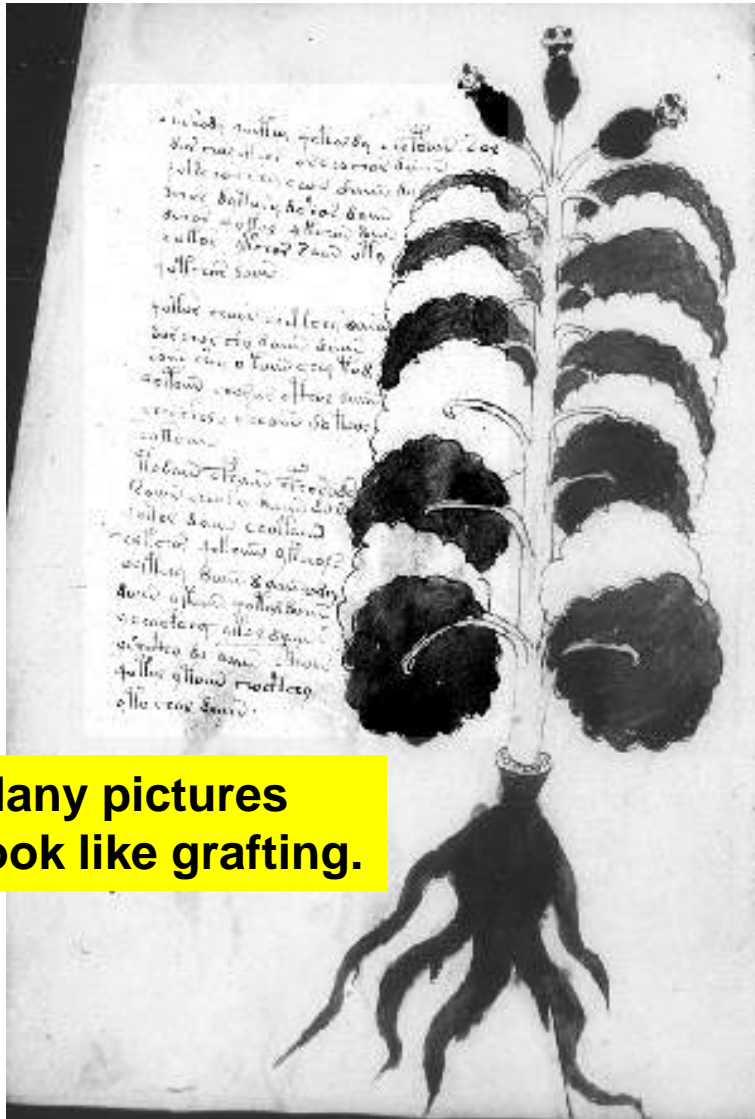
38,000 words of text

golfing golfing golfing golfing golfing golfing golfing golfing golfing
golfing golfing golfing golfing golfing golfing golfing golfing golfing
golfing golfing golfing golfing golfing golfing golfing golfing golfing

Apparent sections of Vms

Section “Name”	# of word tokens
Herbal	11,938
Astrological	2,594
Biological	6,915
Cosmological	679
Pharmacological	5,111
Pure Text (“Stars”)	10,682

The Pictures: Herbal



Many pictures
look like grafting.

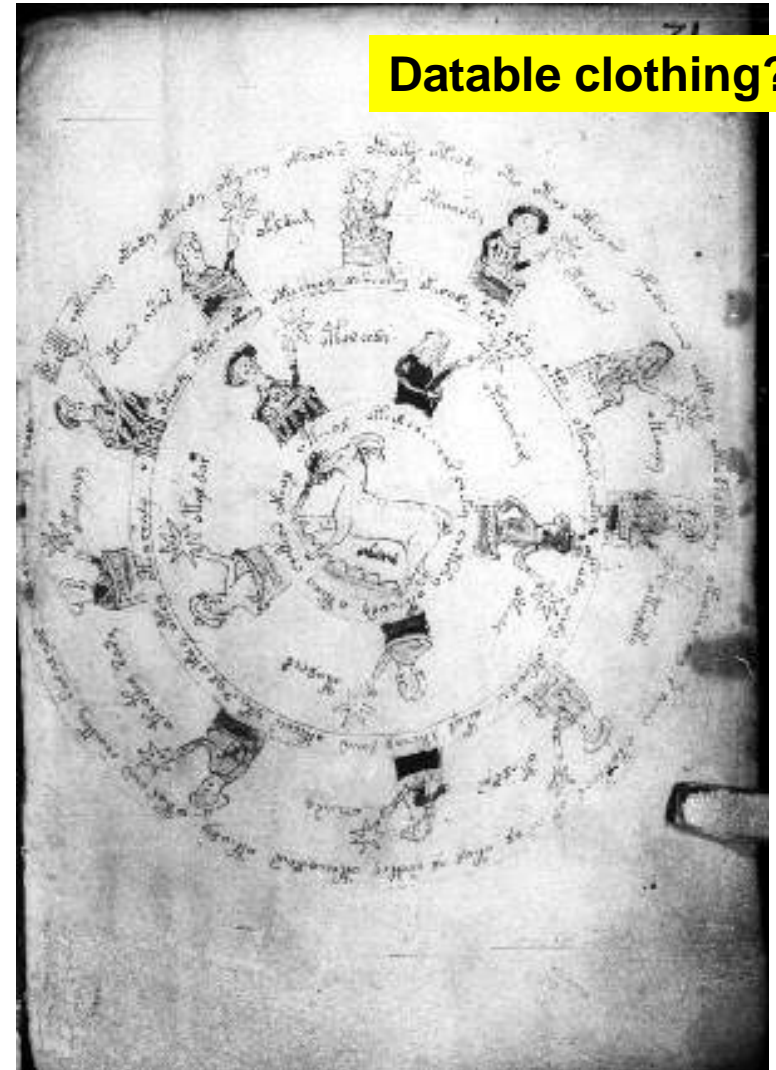
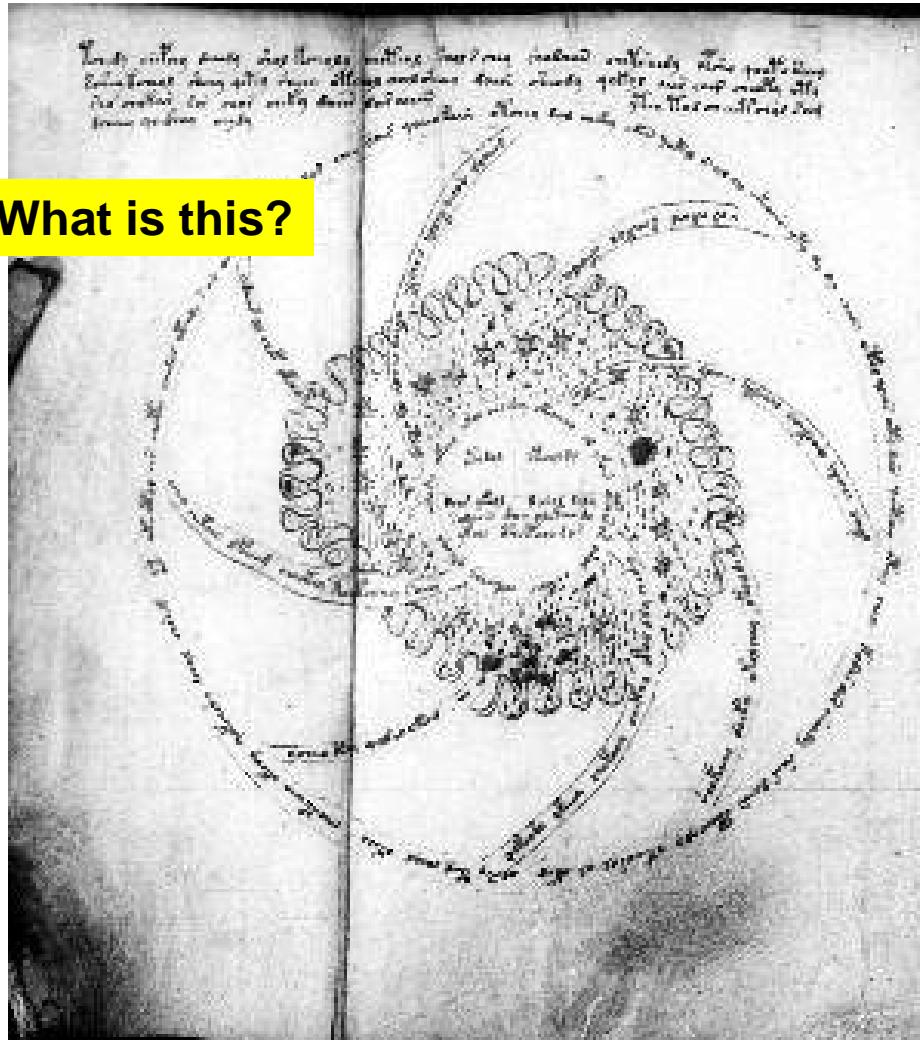


Sunflower? Would date
VMS as post-1492.

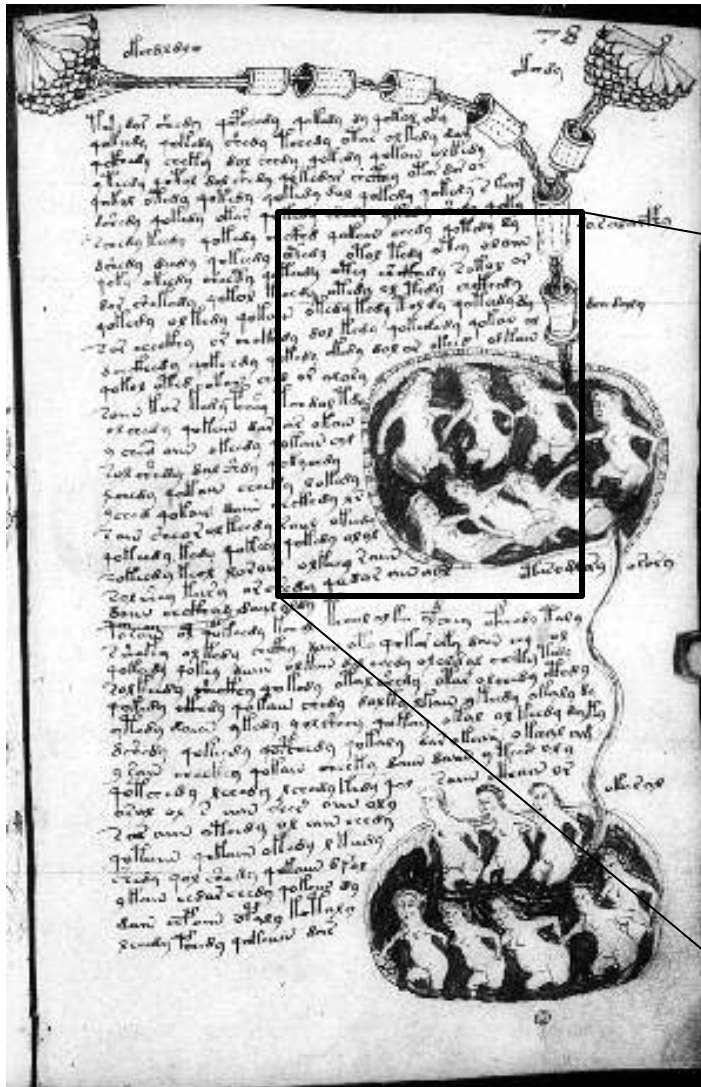
The Pictures: Astrological



The Pictures: Astrological



The Pictures: Biological



Small nudes in baths

Interconnecting tubes of liquids



The Pictures:
Pharmacological

medicine
jar?



History of Voynich Manuscript



**William Newbold,
Polymath, PhD UPenn**



**Wilfrid Michael Voynich
book dealer**

1921 WV presents VMS + **Marci letter**
mentioning 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?)



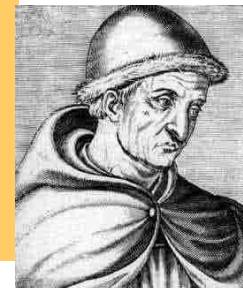
Kircher,
super-scholar,
recipient of
this letter



???,
owned VMS
before Marci



Emperor
Rudolf,
paid 600 ducats
for VMS



Roger Bacon
(1214-94)
“first scientist”

“I’m Not Francis Bacon”

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”

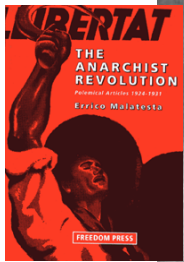
1914 WV moves to USA, opens bookshop

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
- Create $22^2 = 484$ Latin letter pairs AA...XX
 - these letter pairs are the cipher alphabet
- Assign each plaintext Latin letter to a set of cipher-alphabet letter pairs (B → AQ, RT, ...)
- This gives the encipherer some freedom, while the recipient can still decipher by using the table
- Cleverly encipher plaintext in such a way as to construct a “cover” message that looks like Latin, to fool readers

Newbold System

- Example:

a n n ... → DO MI NU ... → DOMINU ...

- Too hard to assemble good “cover” text!

- **So, make cipher letter-pairs overlap:**

a n n ... → AD DB BR ... → ADBR ...

- Also difficult, possibly too easy to decipher

- **So, 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!!

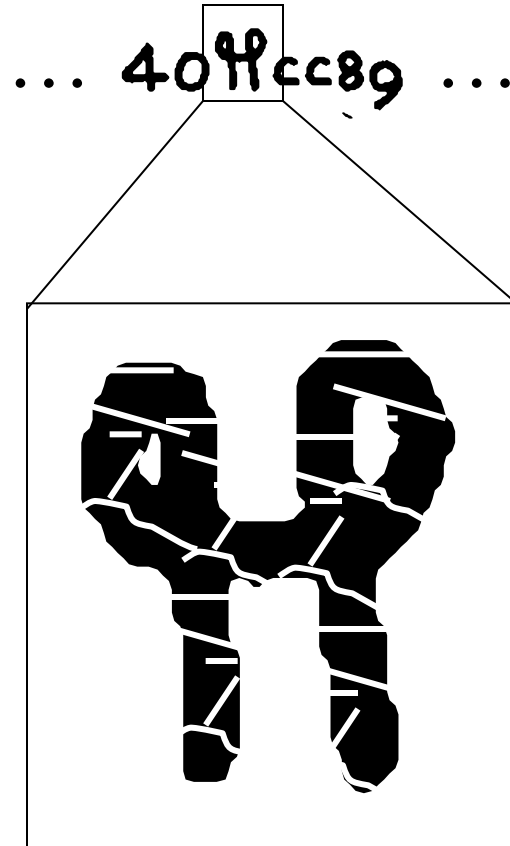
Newbold Decipherment

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

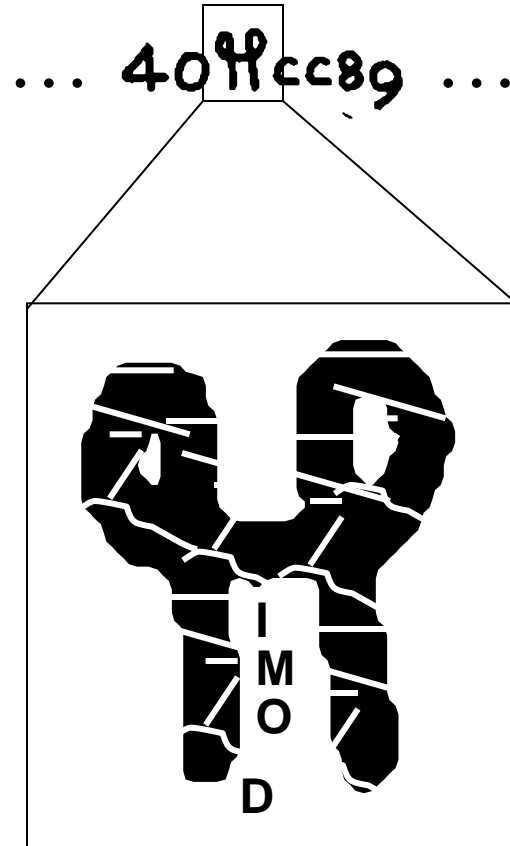


apparent
ciphertext

“artist’s rendition”



William Newbold,
Polymath, PhD UPenn



apparent
ciphertext

real
ciphertext:
DOMI...

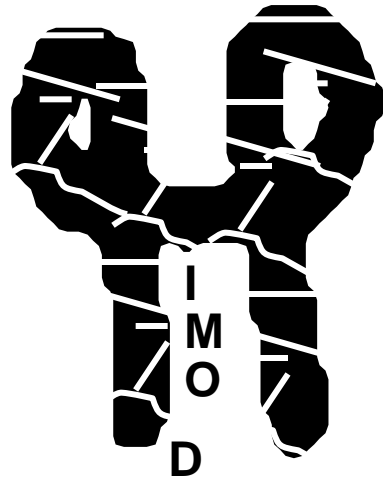
“artist’s rendition”



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

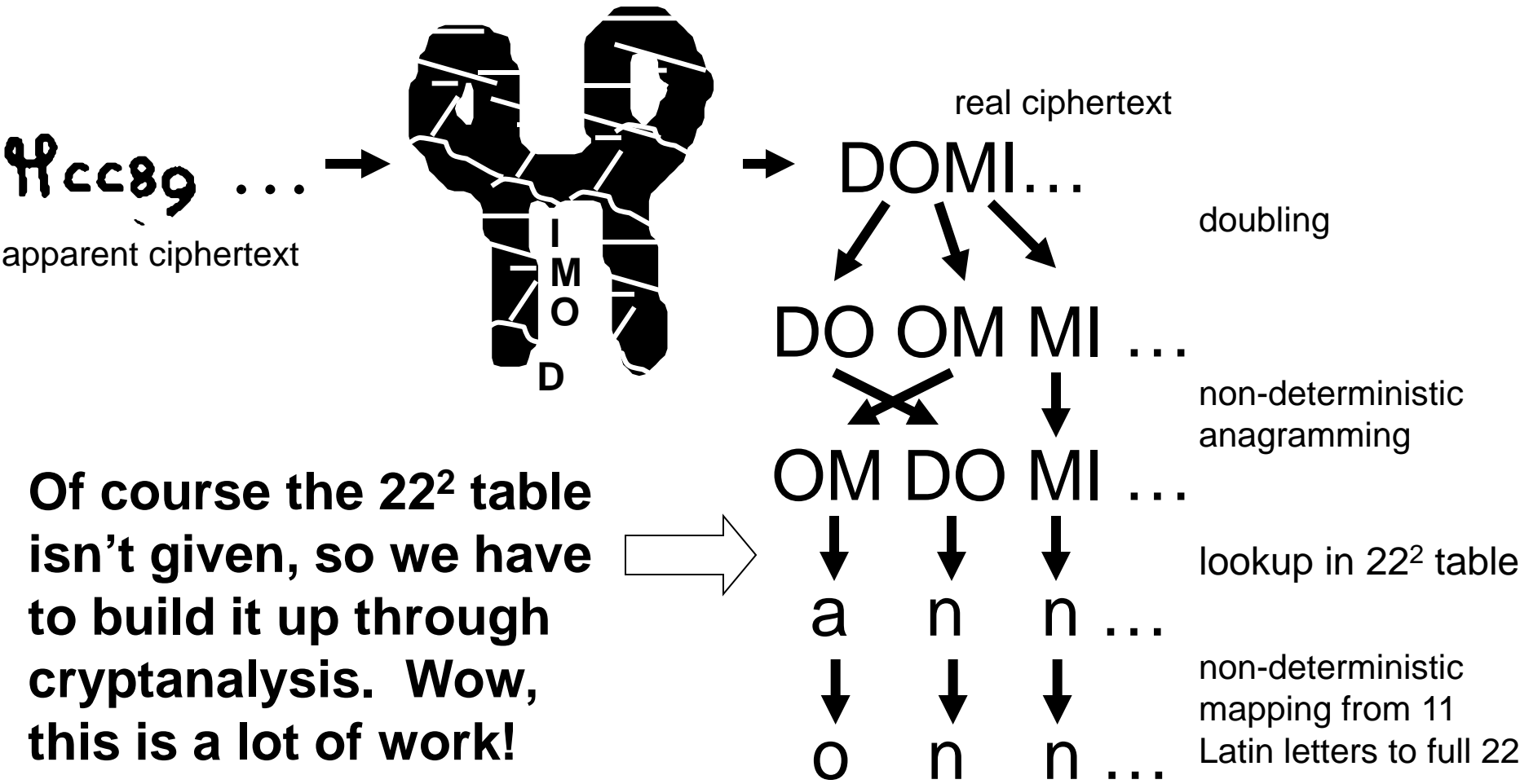
a n n ...

non-deterministic
mapping from 11
Latin letters to full 22

o n n ...



Let's Decipher with Newbold !





Newbold Decipherment

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 & ova ... so Bacon must have had a microscope & telescope, hundreds of years before they were discovered!

The Text

- Approx. 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

ቅርብ ምሽት ምሽት ምሽት ምሽት ምሽት ምሽት ምሽት ምሽት ምሽት ምሽት
ገደብ ምሽት ምሽት ምሽት ምሽት ምሽት ምሽት ምሽት ምሽት ምሽት ምሽት
ገደብ ምሽት ምሽት ምሽት ምሽት ምሽት ምሽት ምሽት ምሽት ምሽት ምሽት

የገደብ ምሽት ምሽት ምሽት ምሽት ምሽት ምሽት ምሽት ምሽት ምሽት ምሽት
ገደብ ምሽት ምሽት ምሽት ምሽት ምሽት ምሽት ምሽት ምሽት ምሽት ምሽት
ገደብ ምሽት ምሽት ምሽት ምሽት ምሽት ምሽት ምሽት ምሽት ምሽት ምሽት

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

Another medieval manuscript, just for calibration...

A yonge may berenge flowers in his Holmolt hit
portrayd &
Chaspette for asynge at yt yt & distancie atore
dyng to certayne pte of the zodiacke as the denomy-
naryon of the yonge berenge veryng the other Earete
by an ager foundaryon of the denomyaryon ynter
maner / For in the .12. signys of the zodiacke the
ber .4. pte of a november many tymes takyn
rederunge the same or an signification. For the .12.
signys have first a sept parte, that yt two signys
make a septile aspekte and therfor yt is
called by name yt holdys the .6. pte of the circle
make no .6. lyng geunge from com centre the
same noted by the name .x. wyng disposed in

Introduction to Astrology and Its Use in Weather Prediction, Medicine, and Agriculture, in English. Manuscript on Paper. 1490.

A yonge may berenge flowers in his Holmolt and
portrayd &
Chaspette for asynge at yt yt & distancie atore
dyng to certayne pte of the zodiacke as the denomy-
naryon of the yonge berenge veryngs other Earete
by an ager foundaryon of the denomyaryon ynter
maner / For in the .12. signys of the zodiacke the
ber .4. pte of a novomber many tymes takyn
rederunge the same or an signification. For the .12.
signys have first a sept parte, that yt two signys
make a septile aspekte and therfor yt is
called by name yt holdys the .6. pte of the circle
make no .6. lyngs geunge from com centre the
same noted by the name .*. whiche disposed in

Alphabet: Currier/D'Imperio

Transcription

c	Ꞑ	ꞑ
C	S	Z

Ꞓ	ꞓ	ꞔ	ꞕ
P	F	B	V

Ꞗ	ꞗ	Ꞙ	ꞙ
Q	X	W	Y

Ꞛ	ꞛ	Ꞝ	ꞝ	Ꞟ	ꞟ	Ꞡ
J	A	E	R	O	I	D

ꞡ	Ꞣ	ꞣ	Ꞥ	ꞥ	Ꞧ
6	7	8	9	4	2

ꞧ	Ꞩ	ꞩ
G	H	1

Ɦ	Ɜ	Ɡ
T	U	0

Ɬ	Ɪ	ꞯ
N	M	3

Ʞ	Ʇ	Ʝ
K	L	5

Alphabet: Currier/D'Imperio

Transcription

c	⋈	⋈
C	S	Z

⋈	⋈	⋈	⋈
P	F	B	V

⋈	⋈	⋈	⋈
Q	X	W	Y

⋈	⋈	⋈	⋈	⋈	⋈	⋈
J	A	E	R	O	I	D

⋈	⋈	⋈	⋈	⋈	⋈
6	7	8	9	4	2

⋈	⋈	⋈
G	H	1

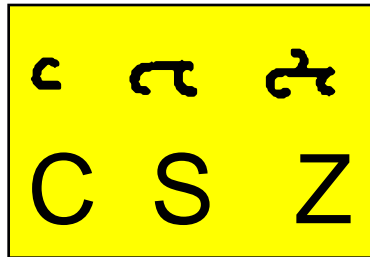
⋈	⋈	⋈
T	U	0

← Maybe this is really
IR IIR IIIR

There are several transcription schemes to choose from.

Alphabet: Currier/D'Imperio

Transcription



Variations of , or separate characters?



Alphabet: Currier/D'Imperio

Transcription

c	Ɑ	Ɱ
C	S	Z

Ɱ	Ɱ	Ɱ	Ɱ
P	F	B	V

Ɱ	Ɱ	Ɱ	Ɱ
Q	X	W	Y

Are these ligatures?

Is Ɱ just a fancy way of writing Ɱ ?

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: $\overline{\text{Ɱ00?Ɱ0x9Ɱ9}}$

Letter Frequencies

count	letter
-------	--------

25468	O o
-------	-----

20227	C c
-------	-----

17655	9 9
-------	-----

14281	A a
-------	-----

12973	8 8
-------	-----

11008	S s
-------	-----

10471	E e
-------	-----

10026	F f
-------	-----

6716	R r
------	-----

5994	P p
------	-----

5423	4 4
------	-----

4501	Z z
------	-----

4076	M m
------	-----

count	letter
-------	--------

2886	2 ?
------	-----

1752	N n
------	-----

1413	B b
------	-----

1046	J j
------	-----

950	Q q
-----	-----

908	X x
-----	-----

591	T t
-----	-----

524	* *
-----	-----

431	V v
-----	-----

316	I i
-----	-----

217	W w
-----	-----

157	D d
-----	-----

156	3 3
-----	-----

count	letter
-------	--------

148	U u
-----	-----

96	6 6
----	-----

74	Y y
----	-----

52	K k
----	-----

31	G g
----	-----

17	L l
----	-----

14	H h
----	-----

2	1 1
---	-----

1	5 5
---	-----

1	0 0
---	-----

Total
63k character tokens

most Frequent Words

count word

863	8AM	8aᳵᳵ
537	OE	0ᳵ
501	SC89	ᳵᳵ89
469	AM	aᳵᳵ
426	ZC89	ᳵᳵ89
396	SOE	ᳵ0ᳵ
363	OR	0ᳵ
350	AR	aᳵ
344	SC9	ᳵᳵ9
318	8AR	8aᳵ
308	4OFCC9	40ᳵᳵᳵᳵ9
305	4OFCC89	40ᳵᳵᳵᳵ89
283	ZC9	ᳵᳵ9
279	4OFAN	40ᳵᳵaᳵᳵ
272	4OFC89	40ᳵᳵᳵ89
270	89	89
262	4OFAM	40ᳵᳵaᳵᳵ
260	AE	aᳵ
253	8AE	8aᳵ
243	2	ᳵ
219	SOR	ᳵ0ᳵ

count word

212	OFAM	0ᳵᳵaᳵᳵ
211	8AN	8aᳵᳵ
191	4OFAE	40ᳵᳵaᳵ
186	ZOE	ᳵᳵ0ᳵ
177	OFCC9	0ᳵᳵᳵᳵ9
174	SCC9	ᳵᳵᳵᳵ9
172	SCOE	ᳵᳵ0ᳵ
155	S9	ᳵ9
155	OPC89	0ᳵᳵᳵ89
154	OPAM	0ᳵᳵaᳵᳵ
152	4OFAR	40ᳵᳵaᳵ
151	9	9
151	4OE	40ᳵ
150	S89	ᳵ89
147	4OF9	40ᳵᳵ9
144	ZCC9	ᳵᳵᳵᳵ9
144	OFAN	0ᳵᳵaᳵᳵ
144	2AM	ᳵaᳵᳵ
143	OPAE	0ᳵᳵaᳵ
141	OPAR	0ᳵᳵaᳵ
140	SX9	ᳵᳵᳵᳵ9

count word

140	OPCC9	0ᳵᳵᳵᳵ9
138	OFAE	0ᳵᳵaᳵ
130	ZO	ᳵᳵ0
129	OFAR	0ᳵᳵaᳵ
119	ESC89	ᳵᳵᳵᳵ89
118	OFC89	0ᳵᳵᳵ89

etc

Totals:

8116 word types

38k word tokens

Word Length Distributions

Voynich

Length	Distribution
1	0.02
2	0.10
3	0.22
4	0.23
5	0.21
6	0.12
7	0.05
8	0.01
9	0.003
10	0.001
11	0.0001
12	0.00007
13	0.00002
35	0.00002

English

Length	Distribution
1	0.03
2	0.15
3	0.16
4	0.15
5	0.11
6	0.09
7	0.11
8	0.08
9	0.05
10	0.03
11	0.01
12	0.006
13	0.002

Counts on word types

Features of the Text

- 115 (out of 8116) word types appear doubled at least once

... 401st cc89 401st cc89 ...

- 8 words appear tripled

... 401°C89 401°C89 401°C89 ...

... πΟχ πΟχ πΟχ ...

... နှင်းလွှာ နှင်းလွှာ နှင်းလွှာ ...

... offa\w offa\w offa\w ...

... 0x 0x 0x ...

... ग्लान्व ग्लान्व ग्लान्व ...

... 8a11v 8a11v 8a11v ...

... 401°Cc8g 401°Cc8g 401°Cc8g ...

However, very few repeated word bigrams and word trigrams!

No word trigram appears more than 5 times.

Some Theories About the Text

- Cryptogram
- Phonetic writing system
- Philosophical language
- Outsider art
- Glossolalia
- Hoax

cryptogram

- Newbold (1921)
- Manly (1931) critique of Newbold
- Feely (1945), abbreviated Latin
- Strong (1945), polyalphabetic cipher, no details
 - might fall into hands of enemies of USA!
- Brumbaugh (1972), numerological box
- Several attempts in the 1990s

William Friedman

- Most famous American cryptographer of World War II
 - broke key ciphers, including Japanese “Purple” code, led proto-NSA
- VMS Study Group (1944-46)
 - developed transcription alphabet
 - group disbanded after the war
- 2nd VMS Study Group (1962)
 - at RCA
- Included his VMS theory in paper on another topic
 - paper shortened due to space constraints
 - VMS theory included in a footnote, as an anagram, to establish “invention date”



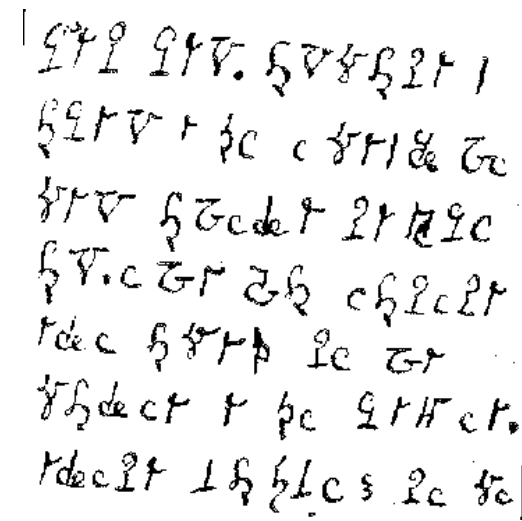
Theory

VMS written in a synthetic “philosophical” language

" Writing in Tongues "

suggested in Kennedy & Churchill, 2005

- Glossolalia (Speaking in tongues)
 - Christian New Testament, Pentecost
 - People spoke tongues foreign to themselves
- Writing in Tongues?
 - 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
 - Looked like a genuine language
 - Grammar closer to French than you might expect



Smith's Martian

Hoax

- Previous hoaxes:
 - Hitler diaries
 - Vinland map
- Voynich Manuscript:
 - How?
 - Why?
 - Who?

How?

- Gordon Rugg
(*Scientific American*,
2004)
 - Proposed Cardan grille
 - Elizabethan espionage tool
 - If applied with randomness injected, claimed to generate VMS-like text



Why?

KPMG Forensic's 2006 Survey of Fraud in Australia and New Zealand

Most Popular Motives for Fraud:

- greed/lifestyle (54%)
- gambling (22%)
- personal financial pressure (5%)
- other (5%)
- not specified (3.5%)
- opportunity (0.4%)
- substance abuse (0.4%)

Who?

suggested in Kennedy &
Churchill, 2005

BUT: what if Voynich
knew that?

BUT: same signature
in other docs

de Tepenecz
signature
suspiciously found
during overexposure

Marci letter
very convenient

faked to add a Roger
Bacon connection?

BUT: Baresch letter later found in Kircher
archive also mention Bacon

member of Society
of Friends of
Russian Freedom

said to have
faked passports



Needed \$ → who doesn't?

tricky → said to have
traded newer,
"better" books
for monks'
old dirty ones

spoke 18
languages

BUT: What if Voynich
had seen that letter?

Experiments

- Can computers help us make sense of VMS?
- Is VMS a kind of letter substitution cipher?
 - Originally in Latin?
 - English?
 - Ukrainian?
 - Ukrainian written without vowels?
- Are there patterns of any sort?

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 ...

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

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 ...

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

Substitution Cipher

decipherment is the analysis
ingcmpnqsnwf cv fpn owoktvcv
of documents written in ancient
hu ihgzsnwfv rqcffnw cw owgcnwf
languages ...
kowazoanv ...

Generative models

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 in [Knight et al 2006].

Easy experiments in Carmel finite-state package:

```
% carmel --train-cascade corpus latin.wfsa subst.wfst
```

Returns trained devices & Viterbi decipherment.

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 ...

If plaintext is assumed to be Latin:
quiss squm is onum pom
quuss hates s qum hatis ...

Hypothesize other Source Languages

- Pre-collect language models for 80 languages
- Decipher against each
- See which decoding run yields highest probability

United Nations

Declaration of Human Rights

300+ words in many of world's languages, UTF-8 encoding

No one shall be arbitrarily deprived of his property
Niemand se eiendom sal arbitrêr afgeneem word nie
Asnjeri nuk duhet të privohet arbitrarisht nga pasuria e tij
لا يجوز تجريد أحد من ملكه تعسفا
Janiw khitisa utaps oraqeps inaki aparkaspati
Arrazoirik gabe ez zaio inori bere jabegoa kenduko
Den ebet ne vo tennet e berc'hentiezh digantañ diouzh c'hoant
Никой не трябва да бъде произволно лишен от своята
собственность
Ningú no serà privat arbitràriament de la seva propietat
任何人的财产不得任意剥夺。
Di a so prupiià ùn ni pò essa privu nimu di modu tirannicu
Nitko ne smije samovoljno biti lišen svoje imovine
Nikdo nesmí být svévolně zbaven svého majetku
Ingen må vilkårligt berøves sin ejendom
Niemand mag willekeurig van zijn eigendom worden beroofd

Nul ne peut être arbitrairement privé de sa propriété
Nimmen mei samar fan syn eigendom berôve wurde
Ninguín será privado arbitrariamente da sua propriedade
Niemand darf willkürlich seines Eigentums beraubt werden
Κανείς δεν μπορεί να στερηθεί αυθαίρετα την ιδιοκτησία του
Avavégui ndojepe'a va'erâi oimeháicha reinte imbáe teéva
Ba wanda za a kwace wa dukiyarsa ba tare da cikakken dalili ba
Senkit sem lehet tulajdonától önkényesen megfosztani
Engan má eftir geðþótta svipta eign sinni
Tak seorang pun boleh dirampas hartanya dengan semena-mena
Necuno essera private arbitrariamente de su proprietate
Ní féidir a mhaoín a bhaint go forlámhach de dhuine ar bith
Al neniú estu arbitre forprenita lia propio
Kelleltki ei tohi tema vara meelevaldselt ära võtta
Eingin skal hissini vera fyrí ongartøku
Me kua ni dua e kovei vua na nona iyau
Keltään älköön mielivaltaisesti riistettävä hänen omaisuuttaan

Unknown Source Language

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	

Generative models

- Okay, that didn't work...
- Let's devise looser generative models, to mine for patterns.

Generative models

Trigram model over $\{a, b, _ \}$

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

a \rightarrow {all Voynich letters}

b \rightarrow {all Voynich letters}

_ \rightarrow _

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

Initially uniform

What parameter settings
result in highest $P(\text{corpus})$?
 \rightarrow EM algorithm.

Generative models

Trigram model over $\{a, b, _ \}$

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

$a \rightarrow \{\text{all English letters}\}$

$b \rightarrow \{\text{all English letters}\}$

$_ \rightarrow _$

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

Initially uniform

What parameter settings
result in highest $P(\text{corpus})$?
 \rightarrow EM algorithm.

Generative models

Trigram model over $\{a, b, _ \}$

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

Initially uniform

What parameter settings
result in highest $P(\text{corpus})$?
→ EM algorithm.

a →

A E O

b →

R N T S

_ → _

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

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 _ ...

Generative models

Trigram model over {a, b, _}

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 _ ...

Initially uniform

What parameter settings
result in highest $P(\text{corpus})$?
→ EM algorithm.

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 _ ...

Generative models

Trigram model over {a, b, _}

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

Initially uniform

What parameter settings
result in highest $P(\text{corpus})$?
→ EM algorithm.



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 _ ...

Generative models

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

English

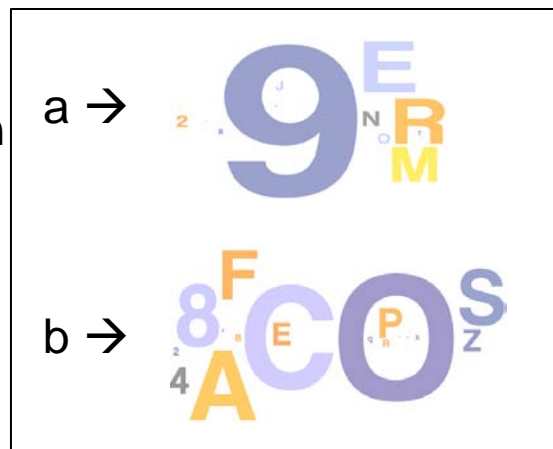


$P(\text{tag} \mid \text{letter})$

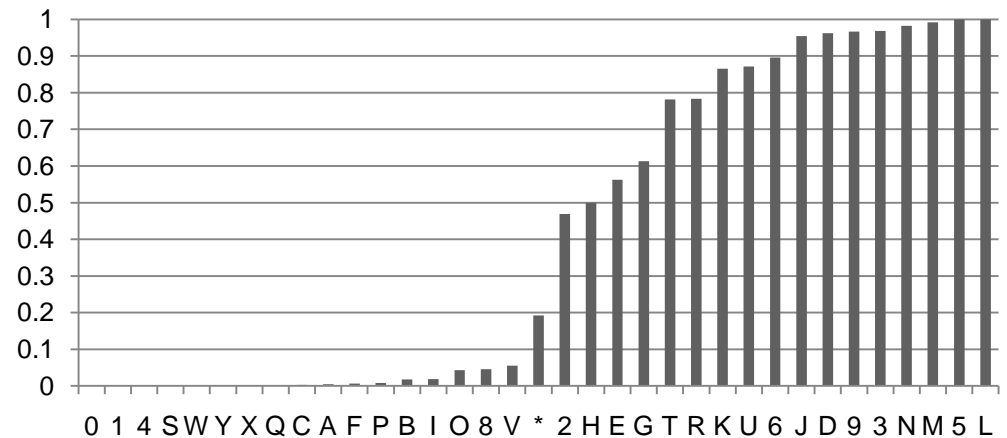
$P(a)$



Voynich



$P(a)$



Generative models

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

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

a \rightarrow {all Voynich **words!**}

b \rightarrow {all Voynich **words!**}

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

What parameter settings
result in highest $P(\text{corpus})$?
 \rightarrow EM algorithm.

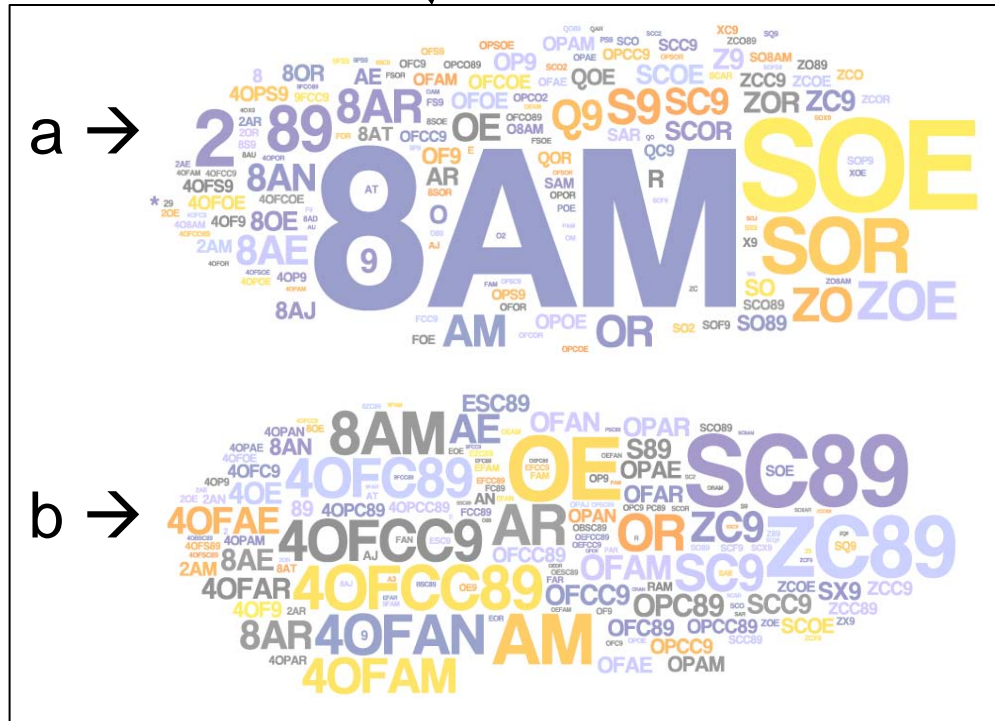
Generative models

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 ...

Generative models

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.

a →



b →



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 ZOR9 ...

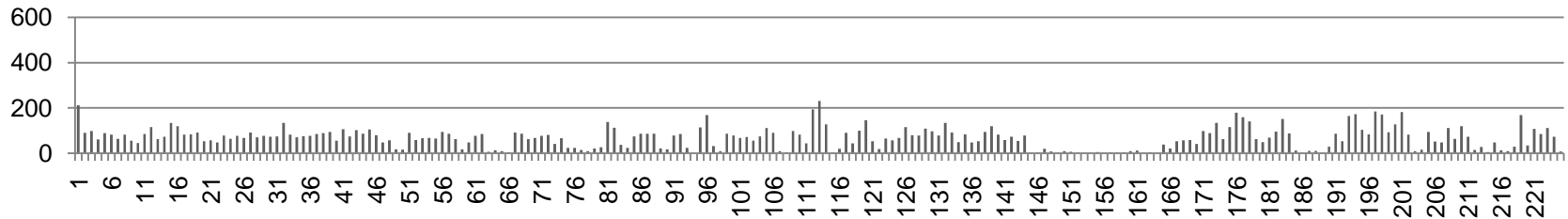
VAS92 9FAE AR APAM ZOE ZOR9 QRC2 9 ...



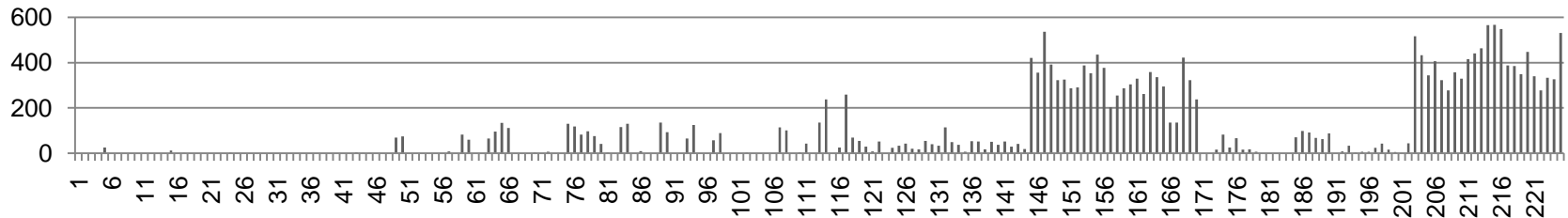
Generative models

Voynich words tagged as “a”

← pages →



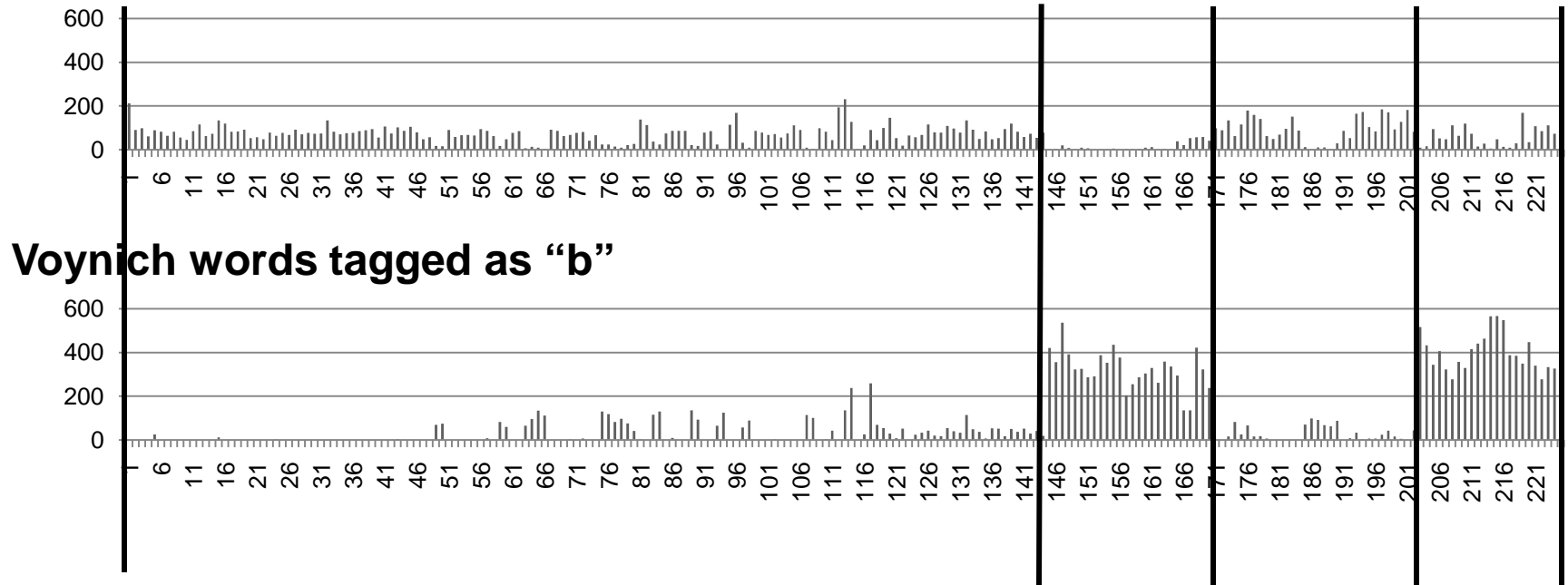
Voynich words tagged as “b”



Generative models

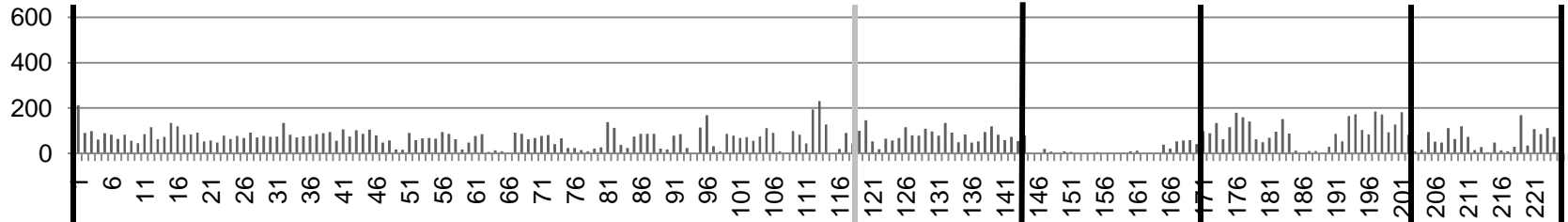
Voynich words tagged as “a”

← pages →

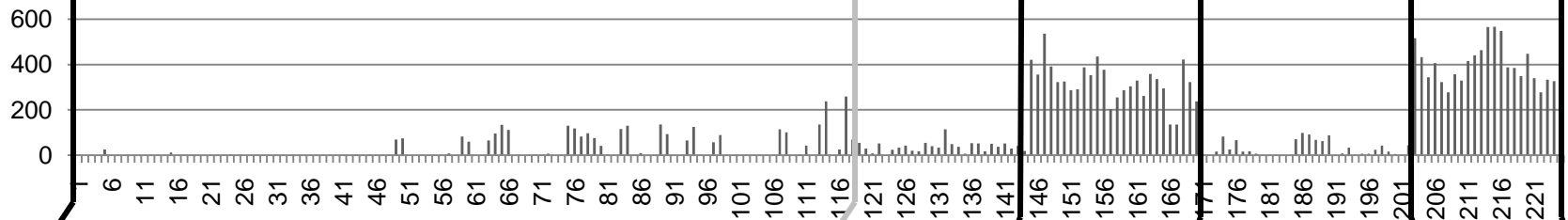


Generative models

Voynich words tagged as “a”



Voynich words tagged as “b”



Herbal

Astro

Bio

Pharma

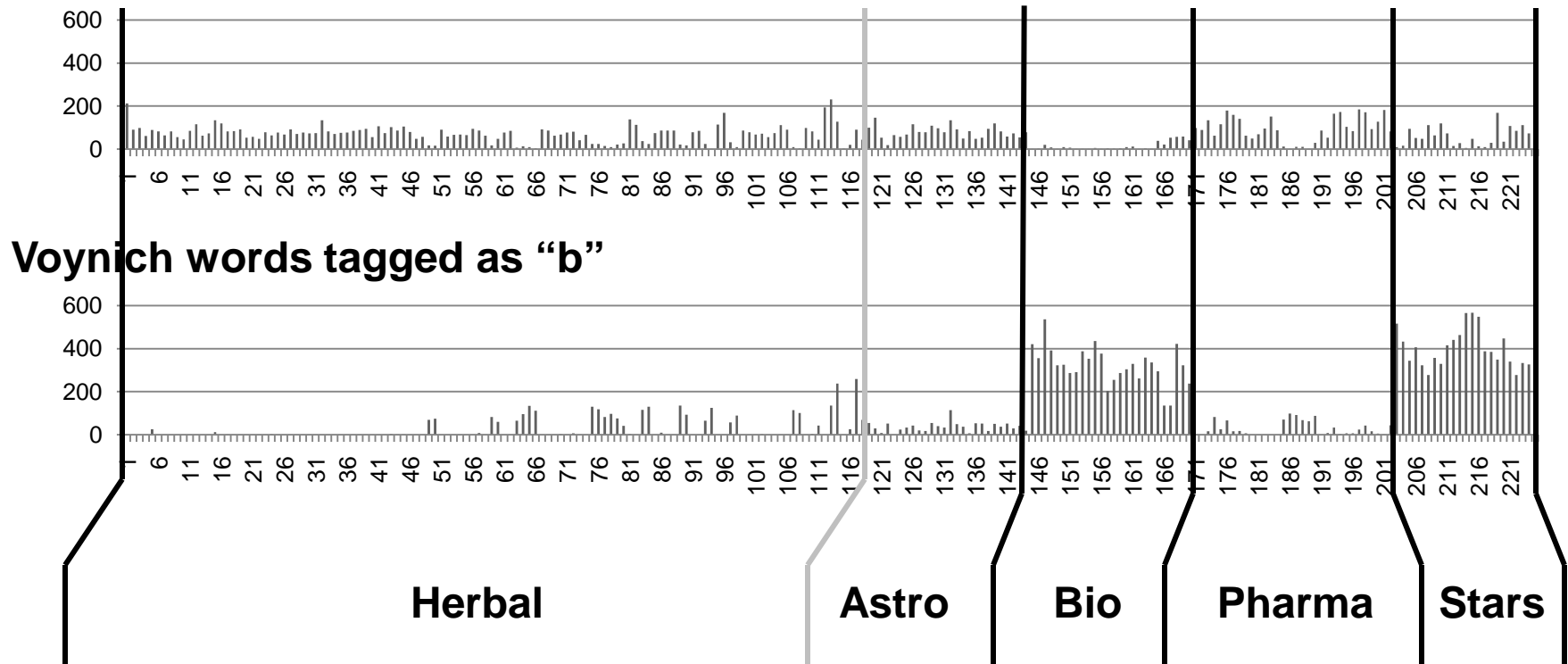
Stars

Generative models

Voynich words tagged as “a”

← pages →

Voynich words tagged as “b”



Known since Capt. Currier’s analysis (1976):

Two “languages” (in the formal sense).

Several handwriting styles, supposedly similar breakdown.

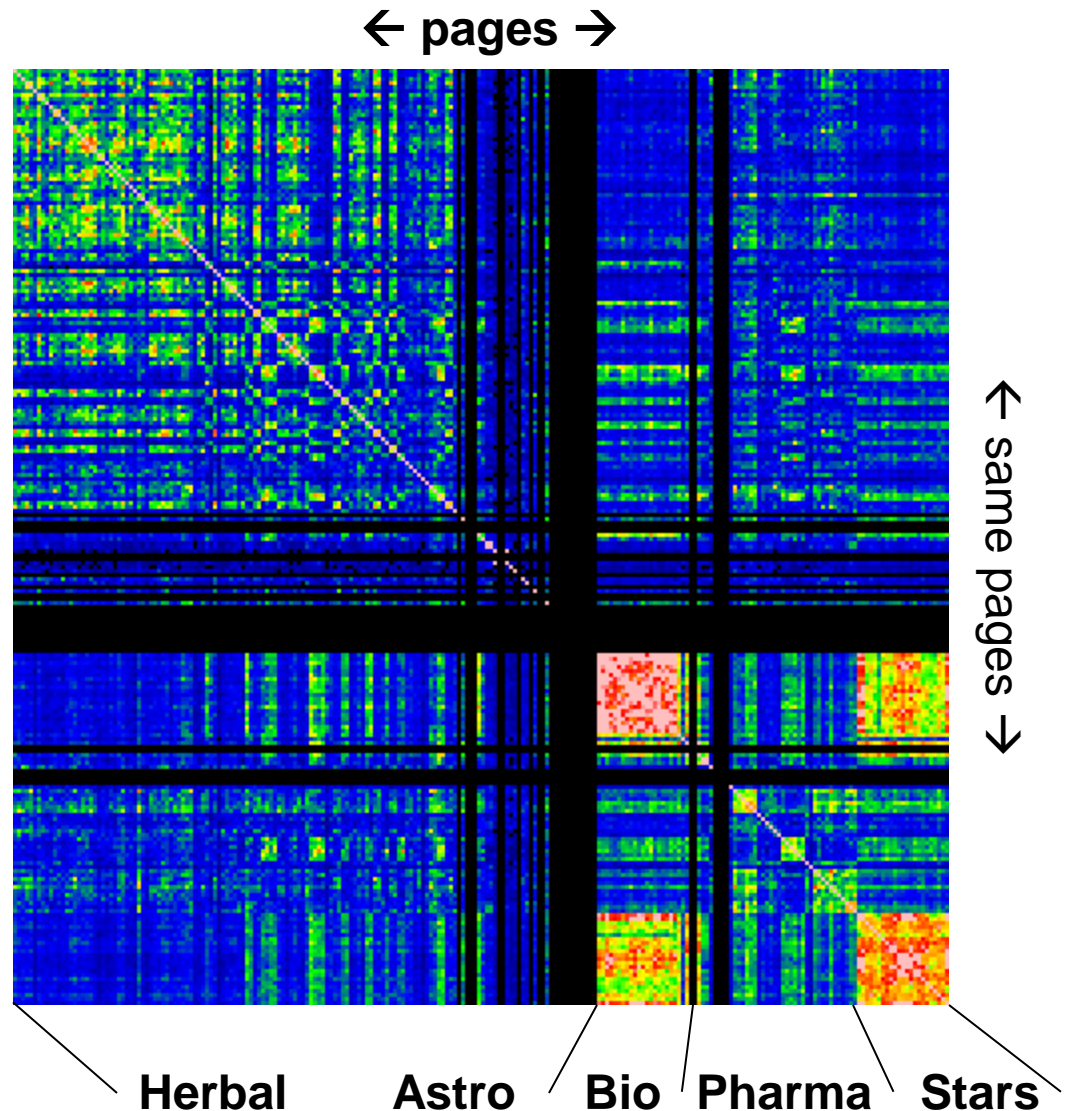
captain carrier's "Two Languages"



Zandbergen Dot Plot

For every pair of pages, how similar are they to each other?

Rene Zandbergen (1997)

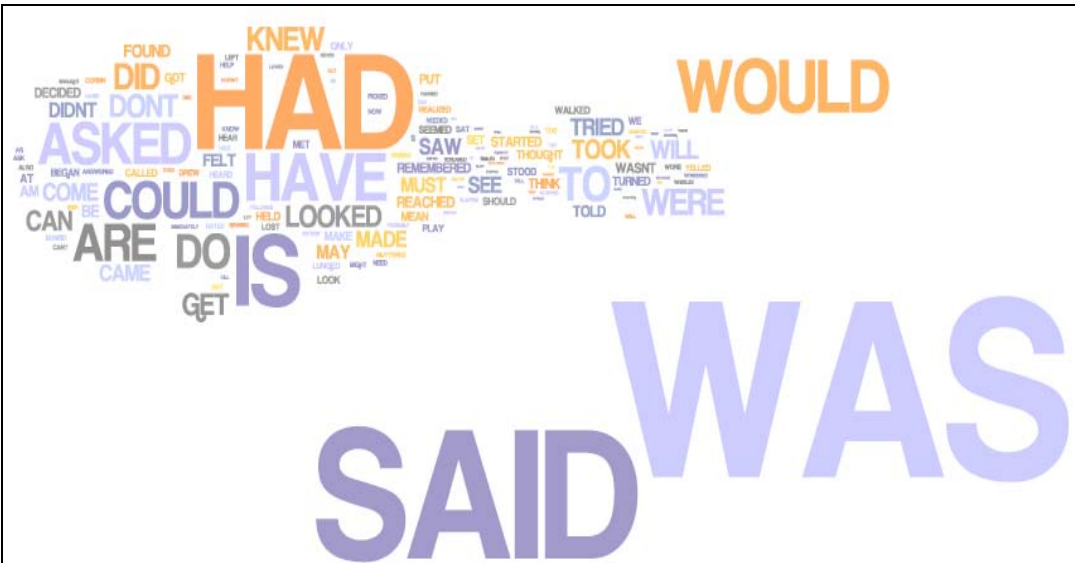
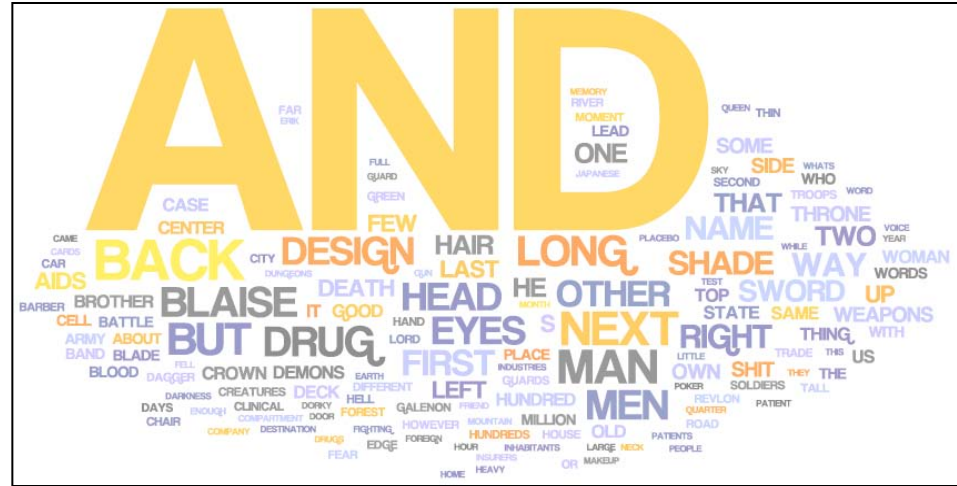


Focus Further Experiments on Voynich-B (Bio & Stars)

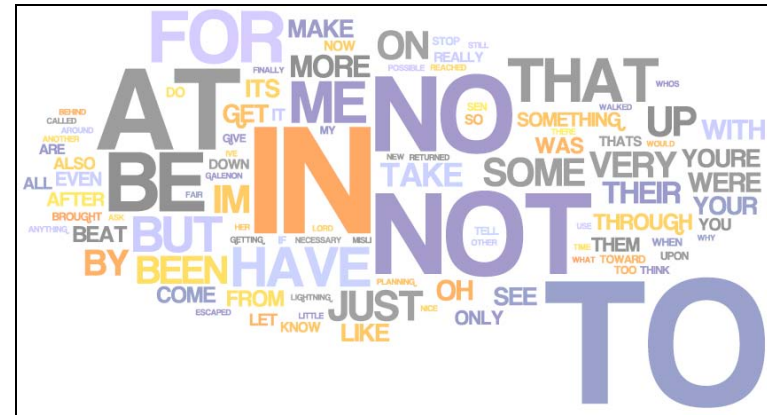
- Consistent vocabulary
- Still plenty of words
- Let's try models that divide **words** into classes
- 10 classes

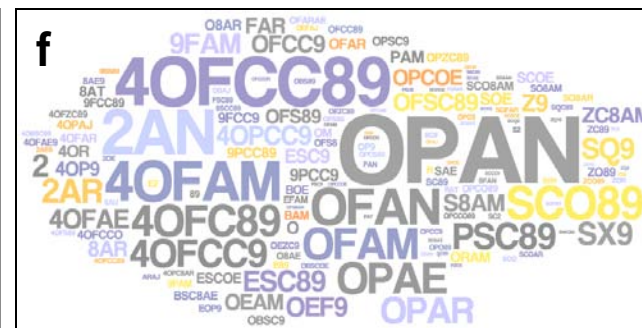
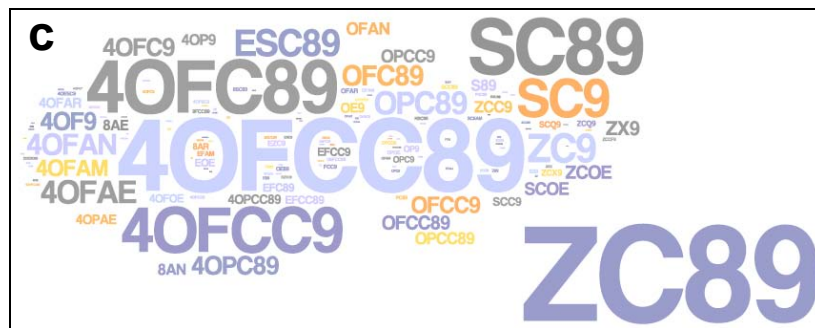
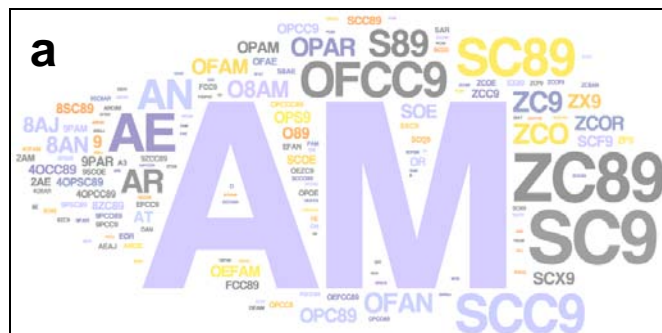
A word cloud visualization of the sentence "My attitude towards her and his". The words are arranged in a grid-like pattern, with each word occupying a rectangular space. The words are color-coded: "MY" is large and grey, "ATTITUDE" is large and blue, "TOWARDS" is large and blue, "HER" is large and blue, "AND" is large and blue, and "HIS" is large and blue. Smaller words like "I", "you", "me", "us", "them", "their", "his", "her", "and", "towards", "attitude", "my", "and", "his" are scattered around the larger words. The background is white.

etc

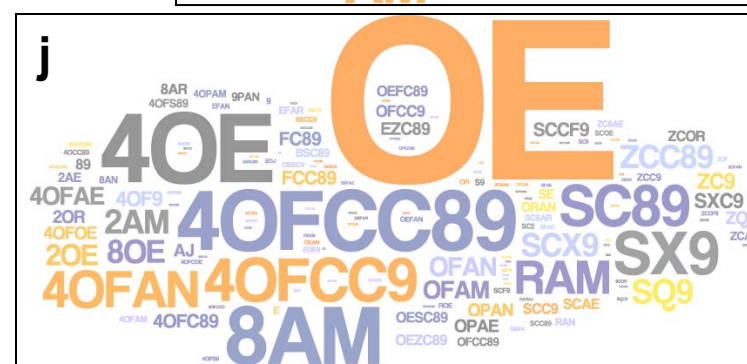
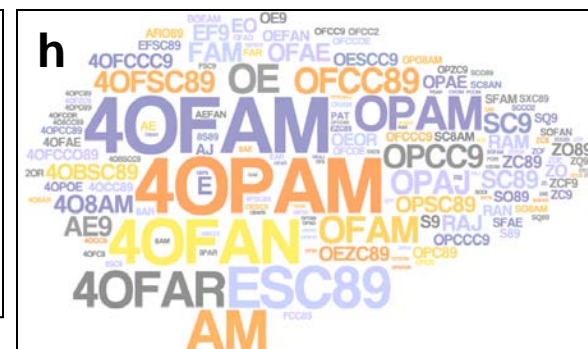
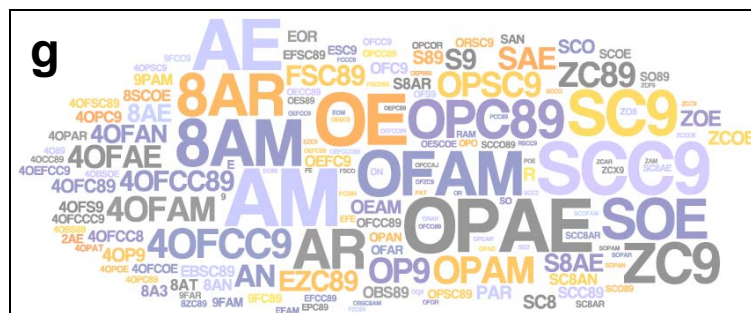


etc





10-class
tagging of
Voynich-B



class-Tag sequences

- Tagging of first VMS page:

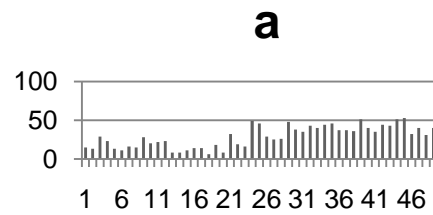
- f g d h f g i d b j c c b e e a h f g e e a b e e a h f g d b j j c c b e a h f g j j j c c
c h f g b j j c c b j j c b j c c b e a h f g b j c b j c c b j c b i d i d c b j c c c c c c
c c c c b e a i d b j c c b j c c b j c c b j c c c c c h f g d b j j j j c c h f g b j j c
b e a b i d i d h f g d i d i d i d h f g d b j j j c b j c c c c b j c c c b e a h f h f h
f g b j c b e e e a h f g b j e a i d i d b j c b j c b j c h f g b j j c c c c c c b j j c
b j c b e a h f g d i d i d b j c b j j j j c b j j c c c b j c b j c b j c c c c b j c b j
c c c c c c i d b j c c c c b j c c c b j c c c c c c b j c h f g e a h f g i d i d b j j c
b j c b j c b j c b e a b j c c c c c b j c c c c c c c c c i d b j c c c c b j c b j c c b
i d i d i d b j j c b j c c c i d i d i d h f g b j c c c c c c c c c c c c c c c b e a h
f g h f g e a i

- 14-grams found in 10-class tagging:

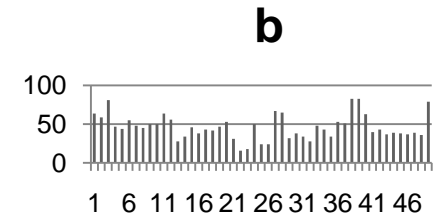
- 25 c c c c c c c c c c c c c c c c
- 9 i d i d i d i d b e a h f g
- 7 i d i d i d i d i d i d i d
- 7 i d i d h f g e e a h f g e
- 7 e a h f g e a h f g e a i d
- 6 j c c c c c c c c c c c c c c

10 classes
of words:
Voynich-B

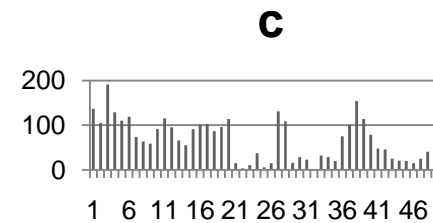
Tags per
page.



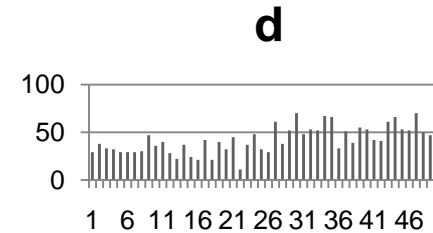
■ a



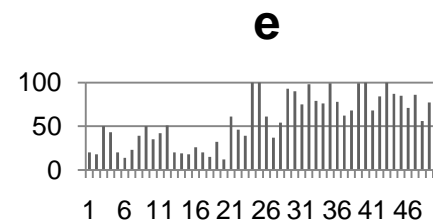
■ b



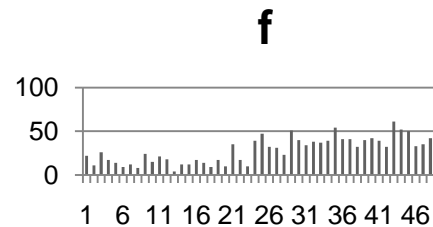
■ c



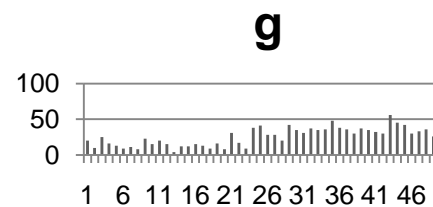
■ d



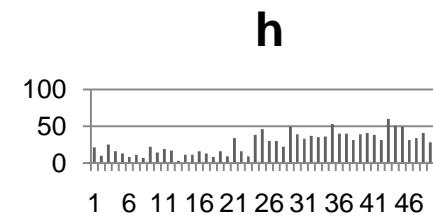
■ e



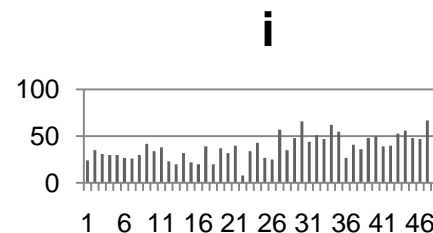
■ f



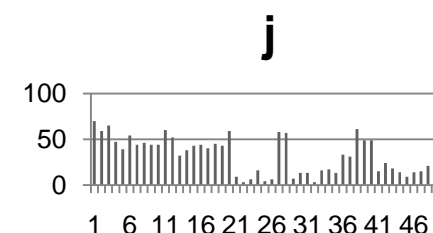
■ g



■ h



■ i

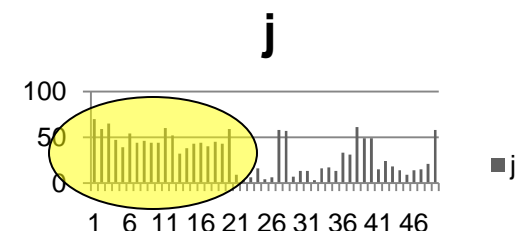
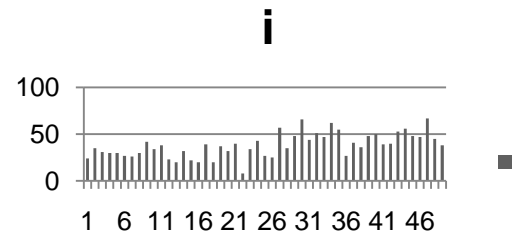
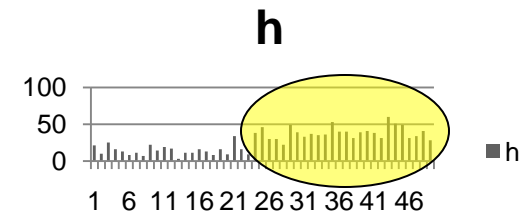
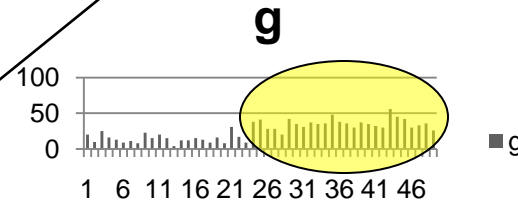
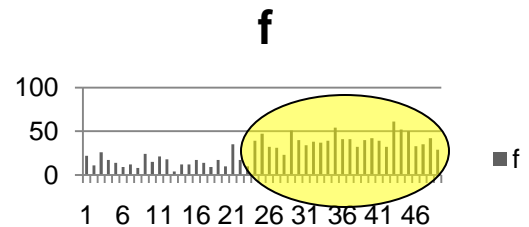
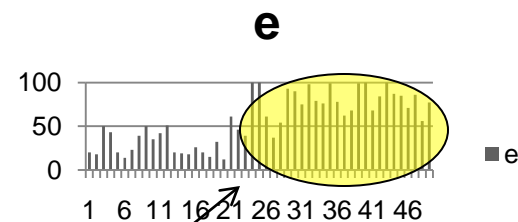
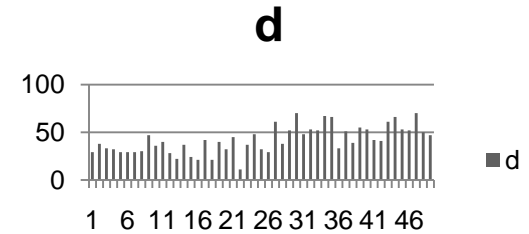
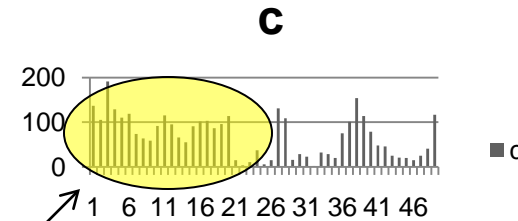
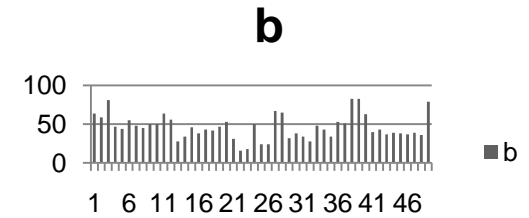
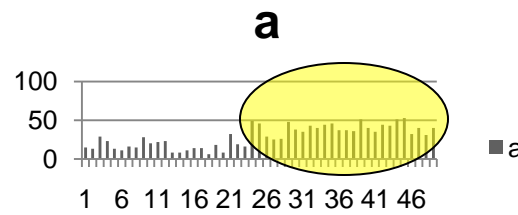


■ j

10 classes
of words:
Voynich-B

Tags per
page.

"Bio" words vs.
"Stars" words



Conclusion

- Voynich Manuscript
 - What it is → pretty clear
 - Where it came from → less clear
 - What it means → totally unclear
- Lots of room for empirical, unsupervised computer techniques
 - Character analysis (e.g., ligatures)
 - Determining relations between words and pictures
 - Identification of “topics”
 - More cipher types

thank you