+ıacpi=ſnxh+xmxdrꝑ:ôrfpj╪zx3o:gng+nʌκꝑrhah:môn
zâjbcûhrcÿmôepxgnpxîlâc:uemajhi=ſclıôrhpôrzânduʌ
cÿpdsxnſhuuıſmnrhzû+xoſ:uʌcÿpurʌh=icpoûr=xſn+c
p=ſoꝑſ╪h3êʌcmûn:fhpâjzôinî3dmnpʌlzıôſꝑlâyôutpſ:ʌu
cıgbxîhorÿô3ft+h:ꝑûrpzjıaʌıuſôcxcnpxsxd=nſêjârdQr
gÿpʌlʌh=xcıpnû3ſzgzubmÿjzuſôchênꝑʌnꝑeeû3zhnpôrdÿ
nꝑscênpʌôvhpmπexrπpznâſughixſzj=dphnʌv+pgfzxſmâxu
pôjmgh̄câʌuzp+ıʌr+hrzgnpʌgmʌʌÿôchhpxu+nʌdſôcpuf
nrwꝑdcÿpiʌıôgxſκſıpgôirs=chſ:up.

Psmôooʌuſexmdar:uhzÿôſ– ʌhmahcds=pcꝑhrcÿm
xurxgzms:poûſâc:uxîjhêpua=mzvſawphcπgꝑſnxût=
hui+=ꝑnſacpxſvngchc:ôzpîdpn:ʌônΔuxx+ûcπâʌf
mî3zueſnôcpâ3nsxd– ʌps=d+hpuupcâÿpêʌnhnx+ıʌ
aahmꝑqzurxjnrôſuvhpſpc:ih=ſwômÿônnc:uzΔrxgz
hm ʌndôî3pʌʌuenôpcſac:tnxpmzÿpſſlôâ3pôʌfıhru
phc:x+hhcd=gpâdpm:ʌlpcxnꝑuezıôſvıſrmnôzê3=+cûnr
hsînpuxôıgꝑrptmxh3uvhgπf.

Nz3nʌ:uſezÿꝑczmʌô3mxnſxrôpπâjnxêſc:ubıg
gzû3xd=rſuzhſmıſrÿmc=rſâjnxtſ:3ôſiſôc:dgÿpʌbꝑhô3
vxêtxhuopcſvnpʌbzꝑ+dptcʌ:urnſuſnxdsxw– ʌnss=
+ppueozhm:xſac:gn=pfnc:ꝑ+nhcpûÿrd– ʌnsxxôôud
ſûoogih=pddânnpa3zuʌcıpnpcdhhpxʌhnxôſpʌ:xʌnmêj
zup.

Aônxjʌuſqdꝑÿnp+ıʌxcnônfſôc:mzpſnxâ3ÿnôſûbôagl
zꝑ+bπjıʌ:uhhptmnʌ:f=rſûjſxôxhêı+uzʌhâncſuaπô3l
mnſcg=jgzphrxôÿmtchxupdonʌûpîhuπx3lûjſphruo
hcſgnppzꝑ+vâ3bm╪rıdcÿpcsx+h– ʌtxû+ppʌi=mzv
πnſg– ʌſ╪dwhpôûcôxʌf.

Dſ╪mſʌuſedꝑnrd– ʌhſac:dn+vxꝑ3ÿmxſuaûſvſôo
dôꝑdupmâ+xêſpmac:ôuſacxcnpch=pcnÿpʌpxnr+hhc
rcxxufznôpoûſâc:uoſıpd+xrıdurcnſ:uozpꝑm=gſô3êc
Onhxſxcπÿıh Xpꝑnrzh=xu+xjdclhnpôuaπî3chxp
mû3mπ=rpnzÿêg—ôrûoobhnrʌj3nıhrâzΔr
xîdʌ+:udſâoom.

Decipherment

Decipherment

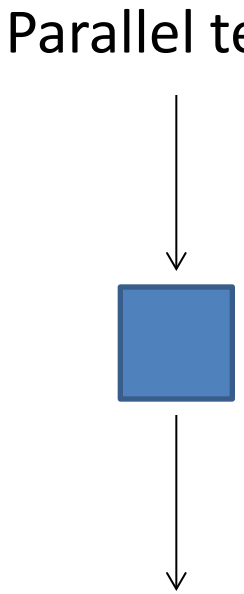
Most material from Kevin Knight (USC/ISI)

Decipherment

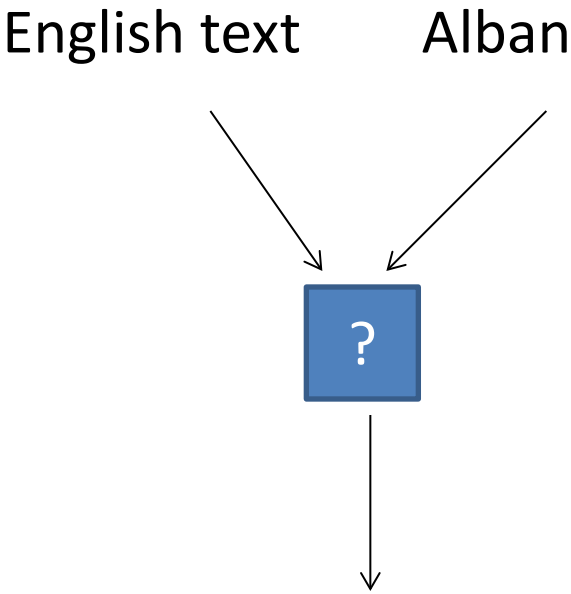
Most material from Kevin Knight (USC/ISI)

Bad timing and terrible jokes are all mine, though.

Learn Translation Knowledge from Non-Parallel Text?

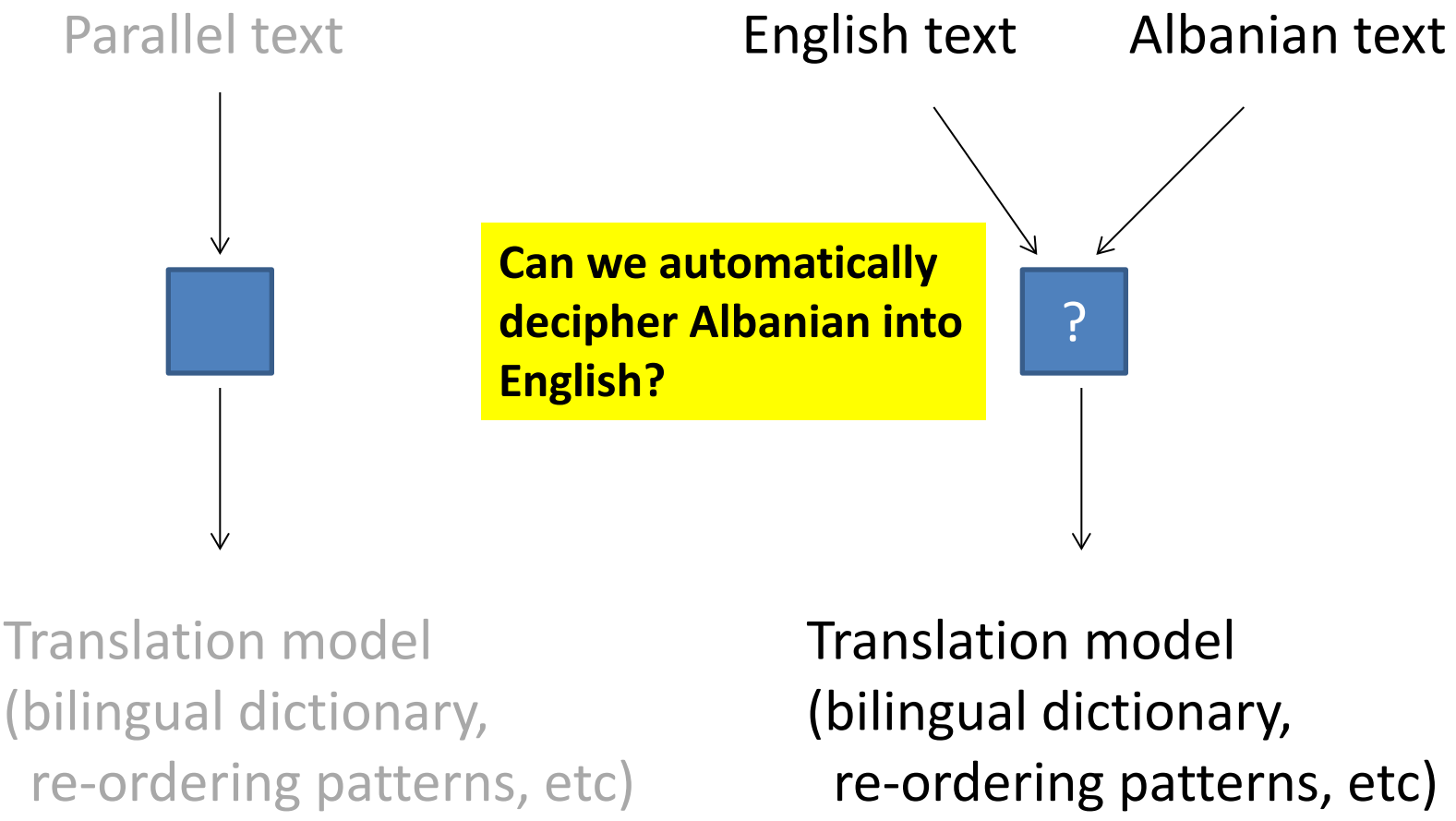


Translation model
(bilingual dictionary,
re-ordering patterns, etc)



Translation model
(bilingual dictionary,
re-ordering patterns, etc)

Learn Translation Knowledge from Non-Parallel Text?



“First NLP Task Ever” (1930s-40s)
Breaking the German Enigma Cipher

German text →  → over the radio: DFKWIFKS LWORISJD KSUEIFKR … →  → German text

```
input (intercepted ciphertext):   DFKWIFKSLWORISJDKSUEIFKR …
output (plaintext):               VASISTDASHERRCAPITANRICH …
```

You can think of this like a multi-class tagging problem.
Each letter of ciphertext can get one of 26 tags (plaintext letters A…Z).

“First NLP Task Ever” (1930s-40s)
Breaking the German Enigma Cipher

Substitution system

N → J

Substitution table **changes** with every keystroke:

NNN → JTE

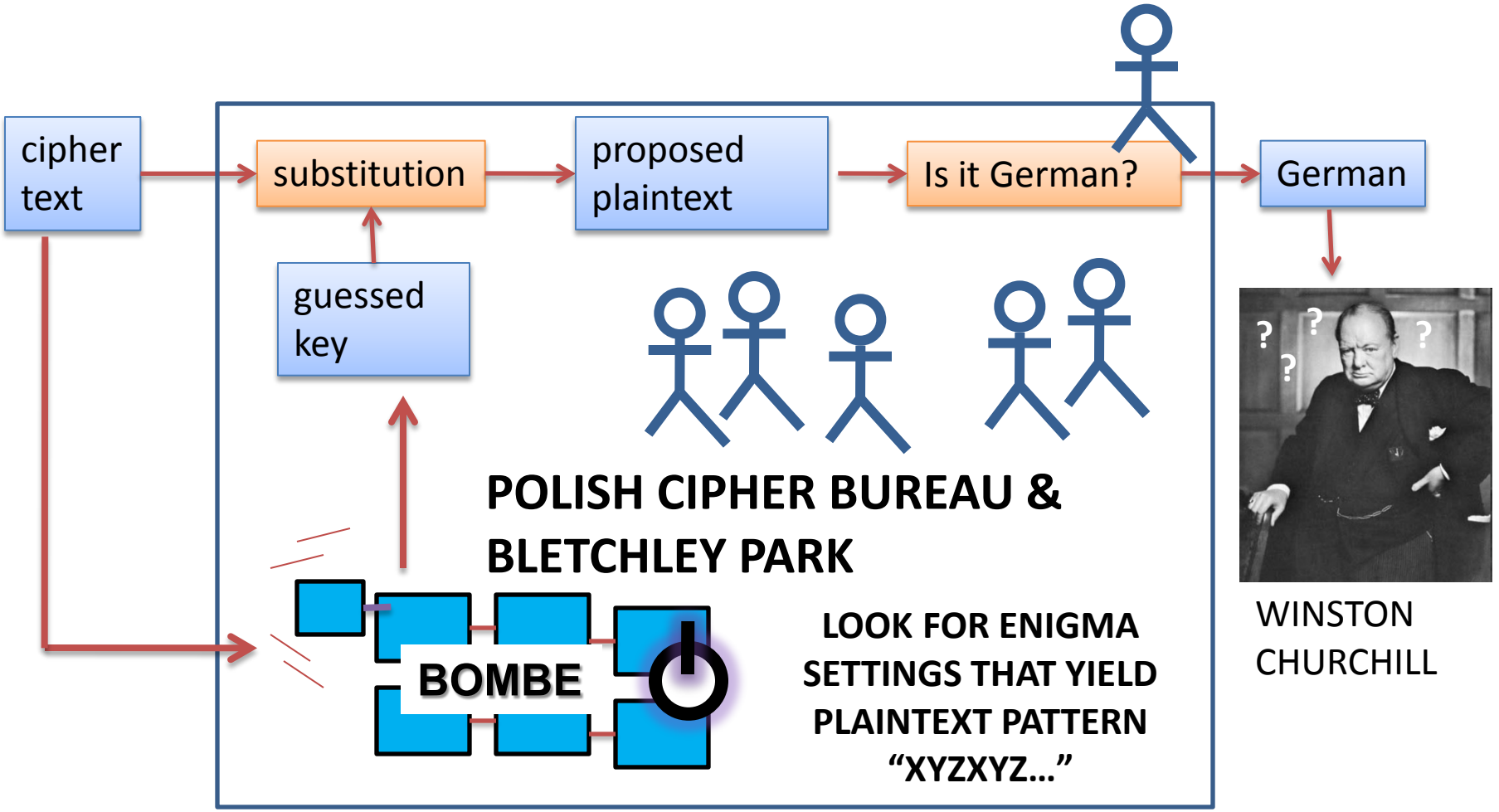
Flattens out ciphertext letter distributions.

Secret key = initial rotor ordering and settings

Reversible behavior

NNN → JTE → NNN

Breaking Enigma



Letter Substitution Cipher

- Encipherment key:

```
PLAIN:    ABCDEFGHIJKLMNOPQRSTUVWXYZ
CIPHER:   PLOKMIJNUHBYGVTFCRDXESZAQW
```

- Plaintext: **HELLO WORLD ...**
- Ciphertext: **NMYYT ZTRYK ...**

- Key itself doesn't change: "simple substitution"
- What key, if applied to the ciphertext, would yield sensible plaintext?

KDCY LQZKTLJKX CY MDBCYJQL: “TR

HYD FKXC, FQ MKX RLQQIQ HYDL

MKL DXCTW RDCDLQ JQMNKXTMB

PTBMYEQL K FKH CY LQZKTL TC.”

KDCY LQZKTLJKX CY MDBCYJQL: “TR

HYD FKXC, FQ MKX RLQQIQ HYDL

MKL DXCTW RDCDLQ JQMNKXTMB

PTBMYEQL K FKH CY LQZKTL TC.”

A	
B	3
C	8
D	7
E	1
F	3
G	
H	3
I	1
J	3
K	10
L	10
M	6
N	1
O	
P	1
Q	10
R	3
S	
T	7
U	
V	
W	1
X	5
Y	7
Z	2

. . . .

KDCY LQZKTLJKX CY MDBCYJQL: “TR

. . . . . .

HYD FKXC, FQ MKX RLQQIQ HYDL

. . . . .

MKL DXCTW RDCDLQ JQMNKXTMB

. . . . .

PTBMYEQL K FKH CY LQZKTL TC.”

A
B 3
C 8
D 7
E 1 .
F 3 .
G
H 3 .
I 1 .
J 3 .
K 10
L 10
M 6
N 1 .
O
P 1 .
Q 10
R 3 .
S
T 7
U
V
W 1 .
X 5
Y 7
Z 2 .

KDCY LQZKTLJKX CY MDBCYJQL: “TR

HYD FKXC, FQ MKX RLQQIQ HYDL

MKL DXCTW RDCDLQ JQMNKXTMB

PTBMYEQL K FKH CY LQZKTL TC.”

A		
B	3	
C	8	
D	7	#
E	1	.
F	3	.
G		
H	3	.
I	1	.
J	3	.
K	10	##### V
L	10	##
M	6	#
N	1	.
O		
P	1	.
Q	10	######## V
R	3	.
S		
T	7	### V
U		
V		
W	1	.
X	5	
Y	7	#### V
Z	2	.

a .a .a . .

KDCY LQZKTLJKX CY MDBCYJQL: “TR

. .a . a . . .

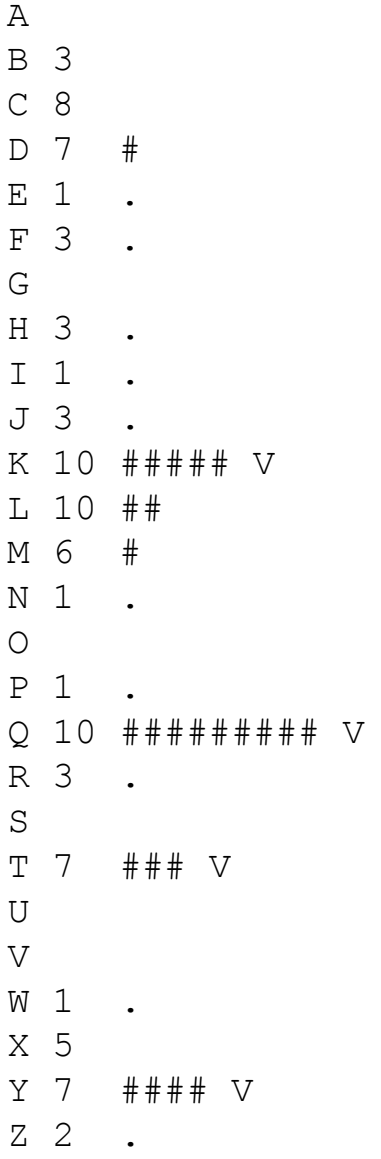
HYD FKXC, FQ MKX RLQQIQ HYDL

a . . . .a

MKL DXCTW RDCDLQ JQMNKXTMB

. . a .a. .a

PTBMYEQL K FKH CY LQZKTL TC.”



a e.a .a .e .

KDCY LQZKTLJKX CY MDBCYJQL: “TR

. .a .e a . ee.e .

HYD FKXC, FQ MKX RLQQIQ HYDL

a . . e .e .a

MKL DXCTW RDCDLQ JQMNKXTMB

. .e a .a. e.a

PTBMYEQL K FKH CY LQZKTL TC.”

didn’t create “ae”

A
B 3
C 8
D 7 #
E 1 .
F 3 .
G
H 3 .
I 1 .
J 3 .
K 10 ##### V
L 10 ##
M 6 #
N 1 .
O
P 1 .
Q 10 ######## V
R 3 .
S
T 7 ### V
U
V
W 1 .
X 5
Y 7 #### V
Z 2 .

a e.ao .a .e o.

KDCY LQZKTLJKX CY MDBCYJQL: “TR

. .a .e a . ee.e .

HYD FKXC, FQ MKX RLQQIQ HYDL

a o. . e .e .a o

MKL DXCTW RDCDLQ JQMNKXTMB

.o .e a .a. e.ao o

PTBMYEQL K FKH CY LQZKTL TC.”

don’t like “ao” – back up!

```
A
B 3
C 8
D 7  #
E 1  .
F 3  .
G
H 3  .
I 1  .
J 3  .
K 10 ##### V
L 10 ##
M 6  #
N 1  .
O
P 1  .
Q 10 ######## V
R 3  .
S
T 7  ### V
U
V
W 1  .
X 5
Y 7  #### V
Z 2  .
```

auto repairman to customer: if

KDCY LQZKTLJKX CY MDBCYJQL: “TR

you wait we can freeze your

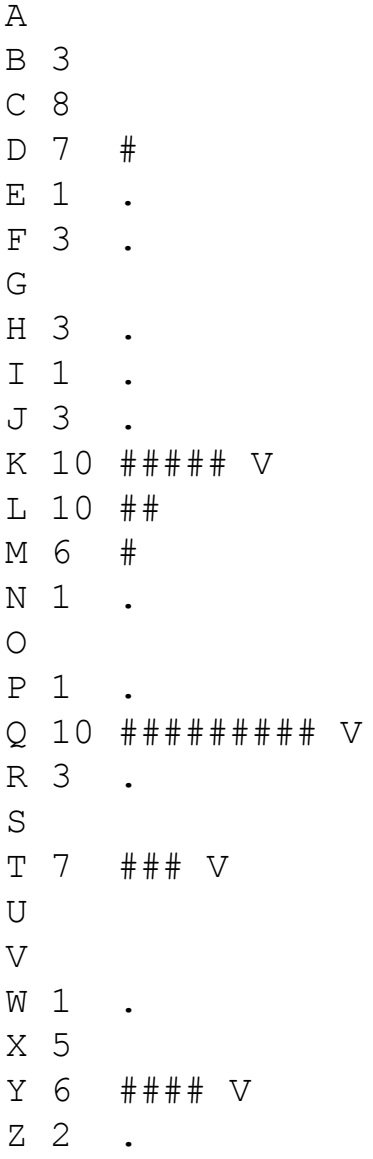
HYD FKXC, FQ MKX RLQQIQ HYDL

car until future mechanics

MKL DXCTW RDCDLQ JQMNKXTMB

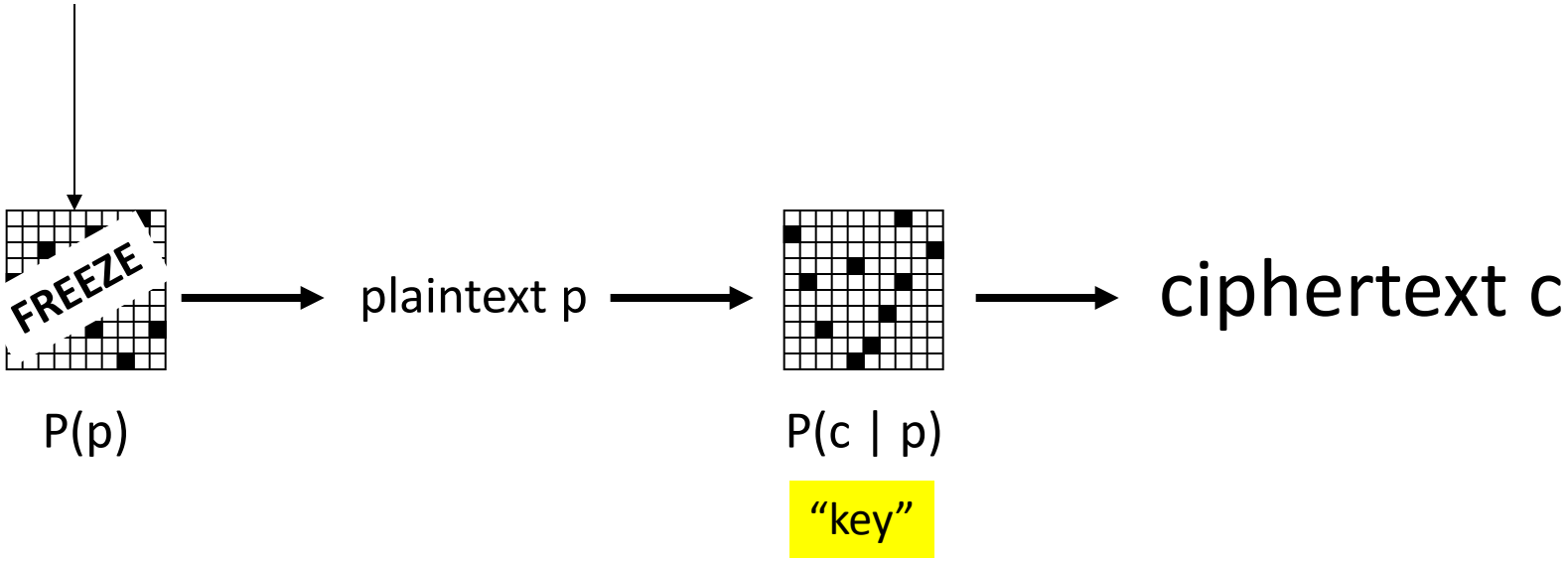
discover a way to repair it

PTBMYEQL K FKH CY LQZKTL TC.”

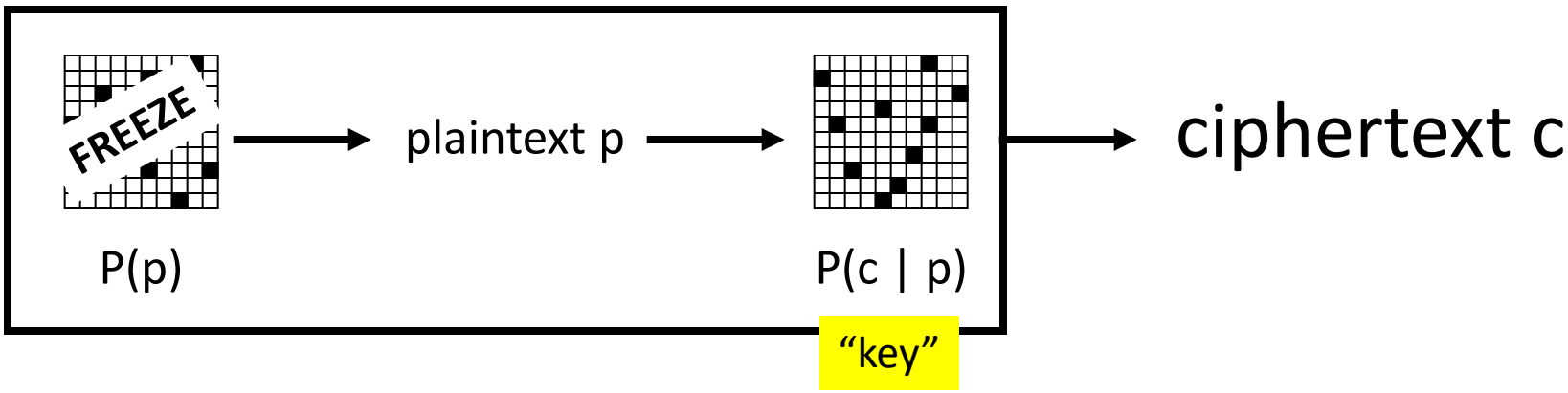


Letter Substitution Cipher

plaintext samples,
unrelated to ciphertext



Letter Substitution Cipher

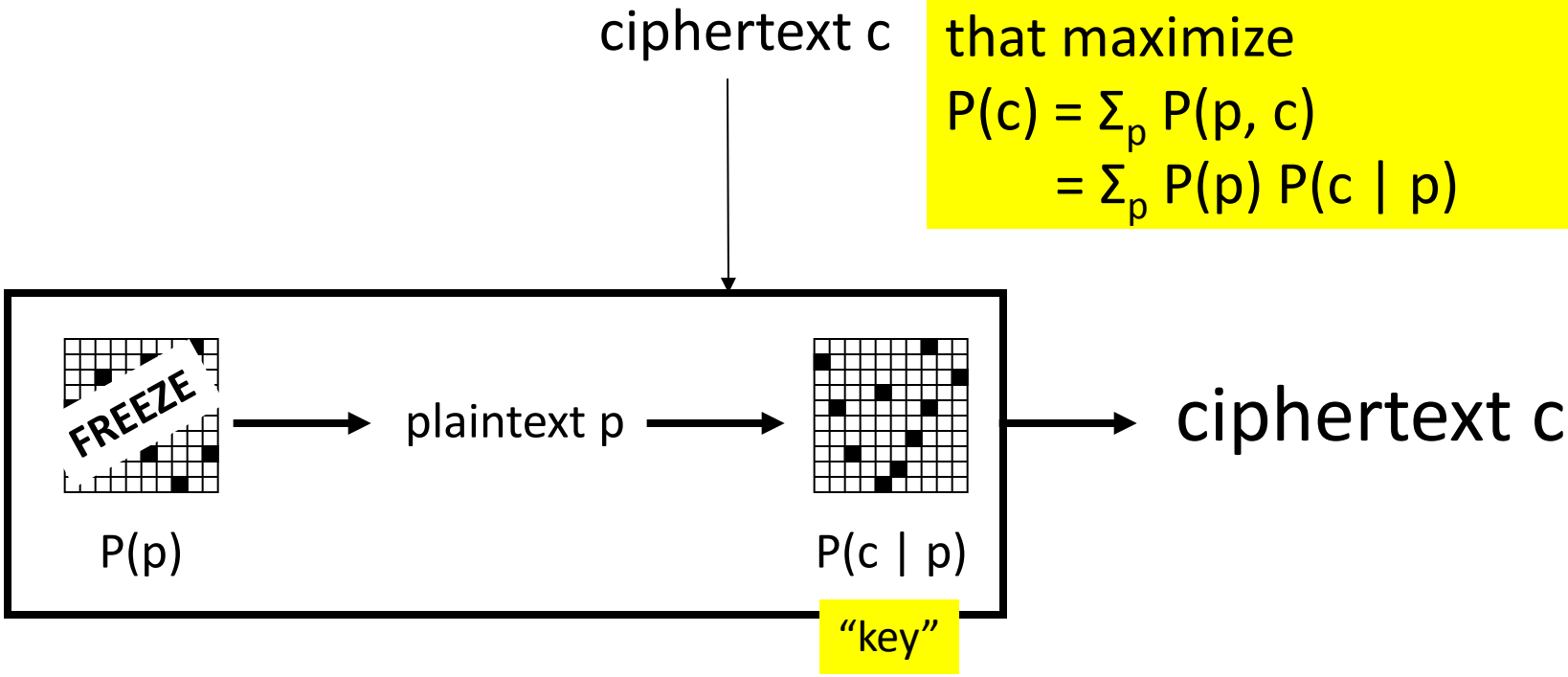


Letter Substitution Cipher

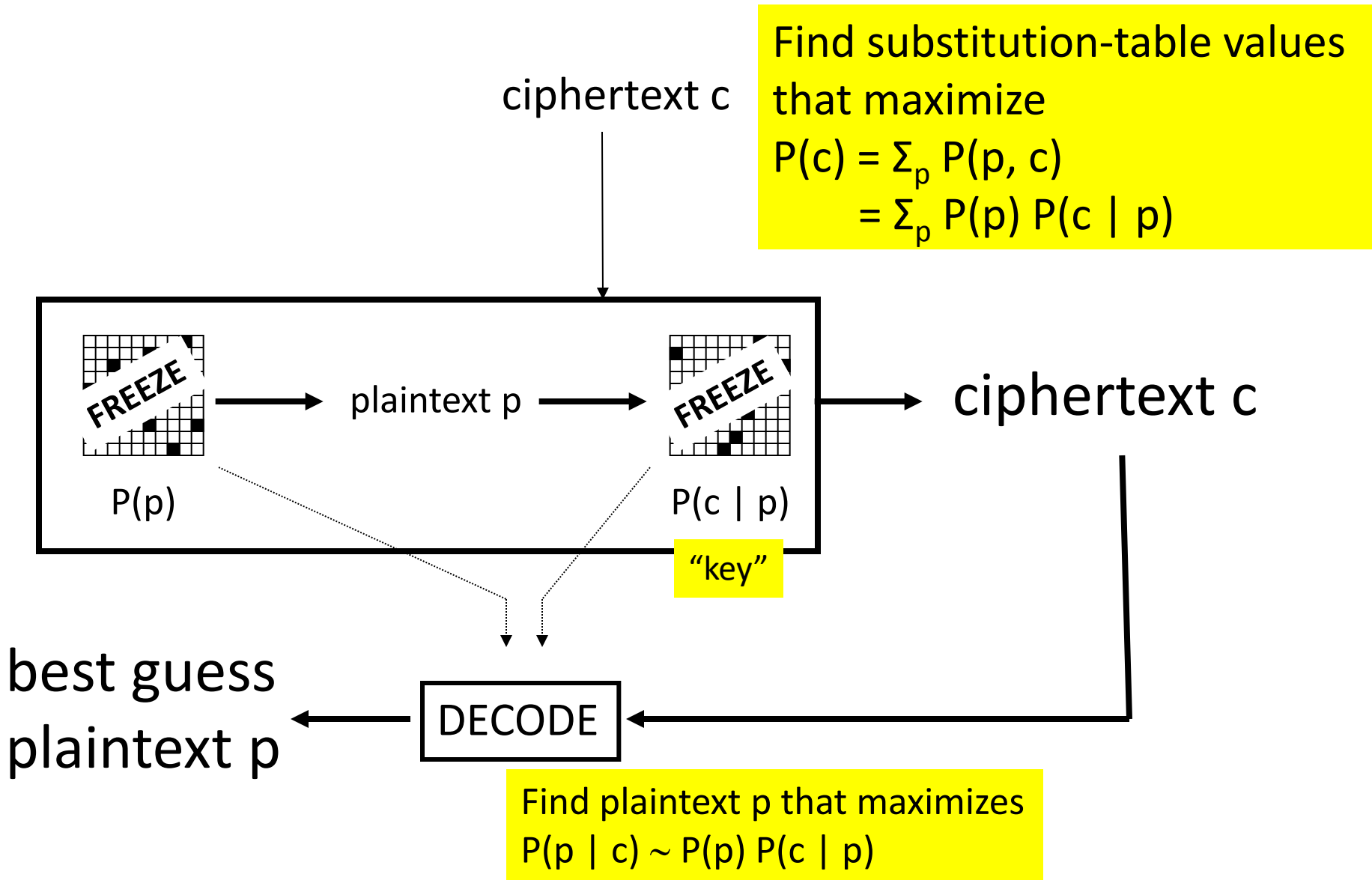
Find substitution-table values that maximize

$P(c) = \Sigma_p P(p, c)$

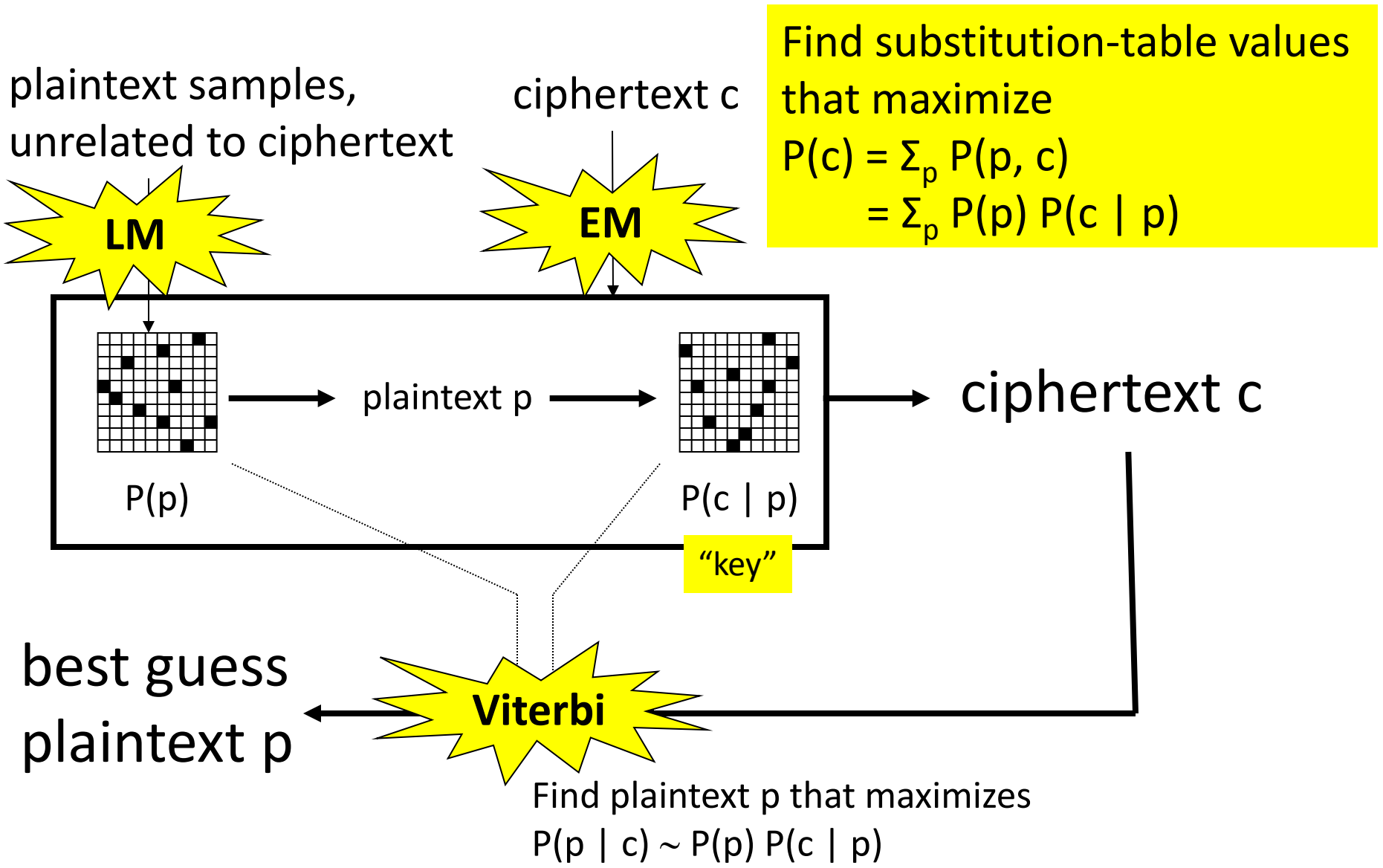
$= \Sigma_p P(p) P(c \mid p)$



Letter Substitution Cipher

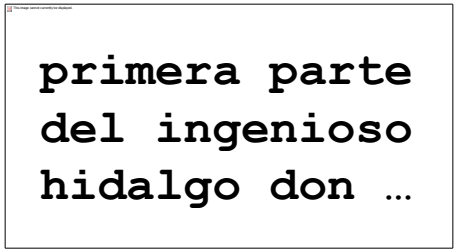


Letter Substitution Cipher



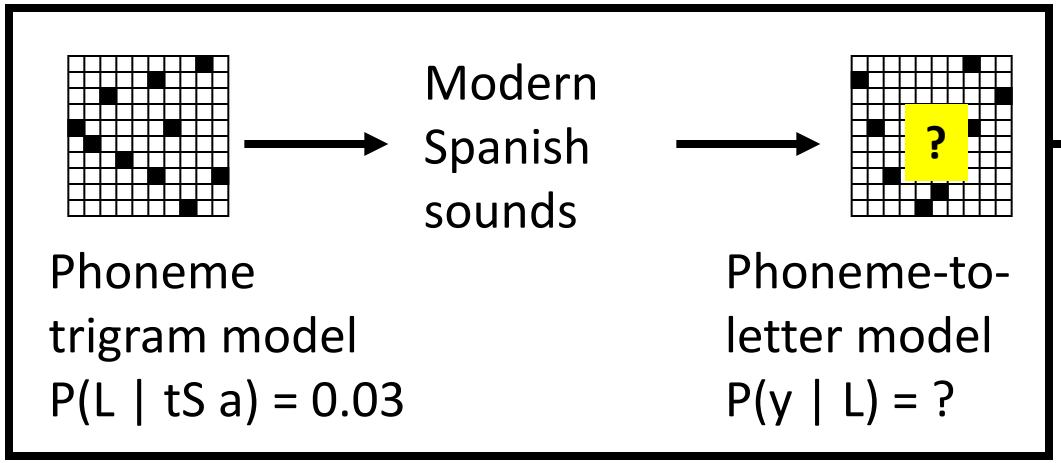
Phonetic Decipherment

ciphertext



[Knight & Yamada 99]

Phonetic Decipherment



ciphertext

primera parte del ingenioso hidalgo don …

(Don Quixote)

26 sounds:
B, D, G, J (can<u>y</u>on),
L (<u>y</u>arn), T (<u>th</u>in), a,
b, d, e, f, g, i, k, l,
m, n, o, p , r,
rr (trilled), s,
t, tS, u, x (<u>h</u>at)

↔ ? ↔

32 letters:
ñ, á, é, í, ó, ú,
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

EM approach = 93% accurate phonetic decipherment

[Knight & Yamada 99]

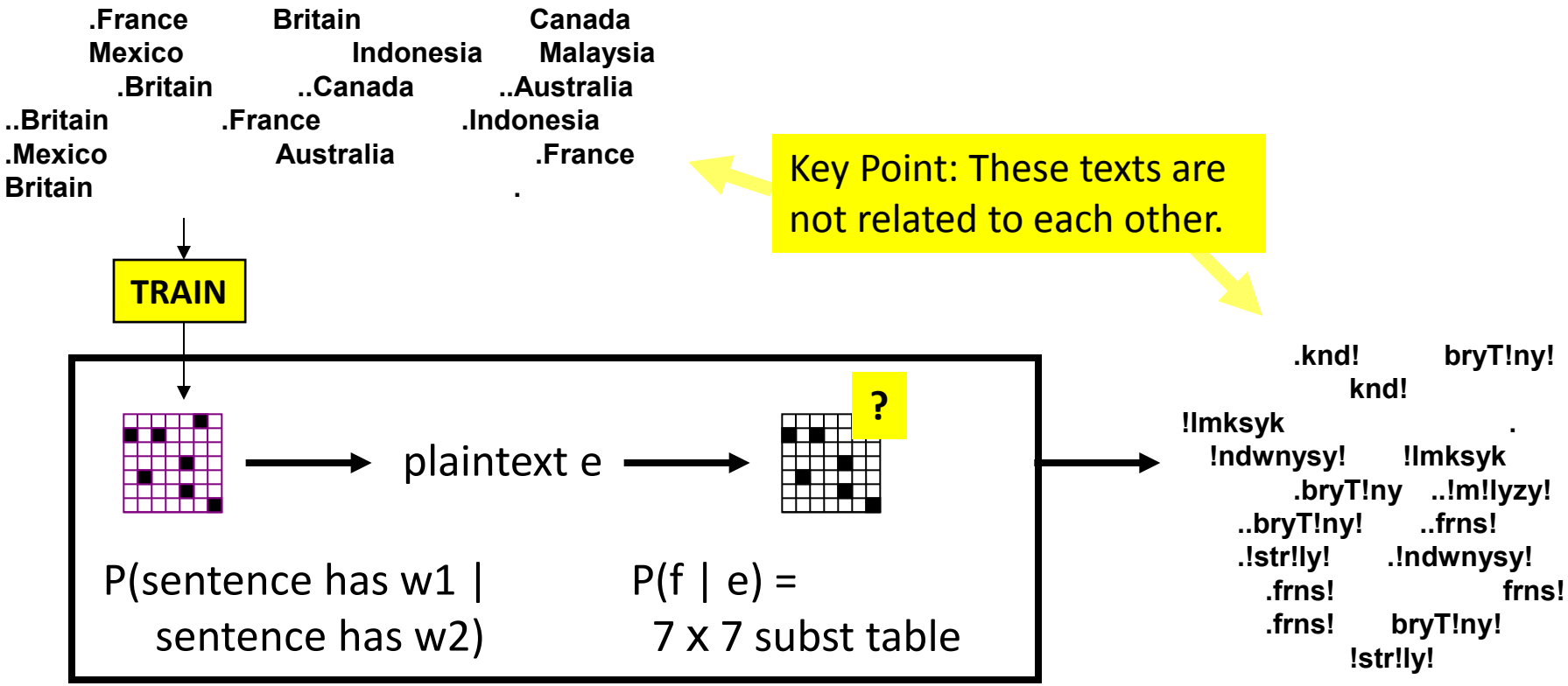
What if Spoken Language Behind Script is Unknown?

- Build a universal model P(p) of human phoneme sequence production
  - human might generally say: K AH N AH R IY
  - human won't generally say: R T R K L K
- Find a P(c | p) table
  - such that there is a decoding with a good universal P(p) score

- Phoneme & syllable inventory
  - if z, then s
  - all have CV syllables; if VCC, then also VC
- Syllable sonority structure
  - dram, lomp, ? rdam, ? lopm
- Physiological preference constraints
  - tomp, tont, ? tomk, ? tonp

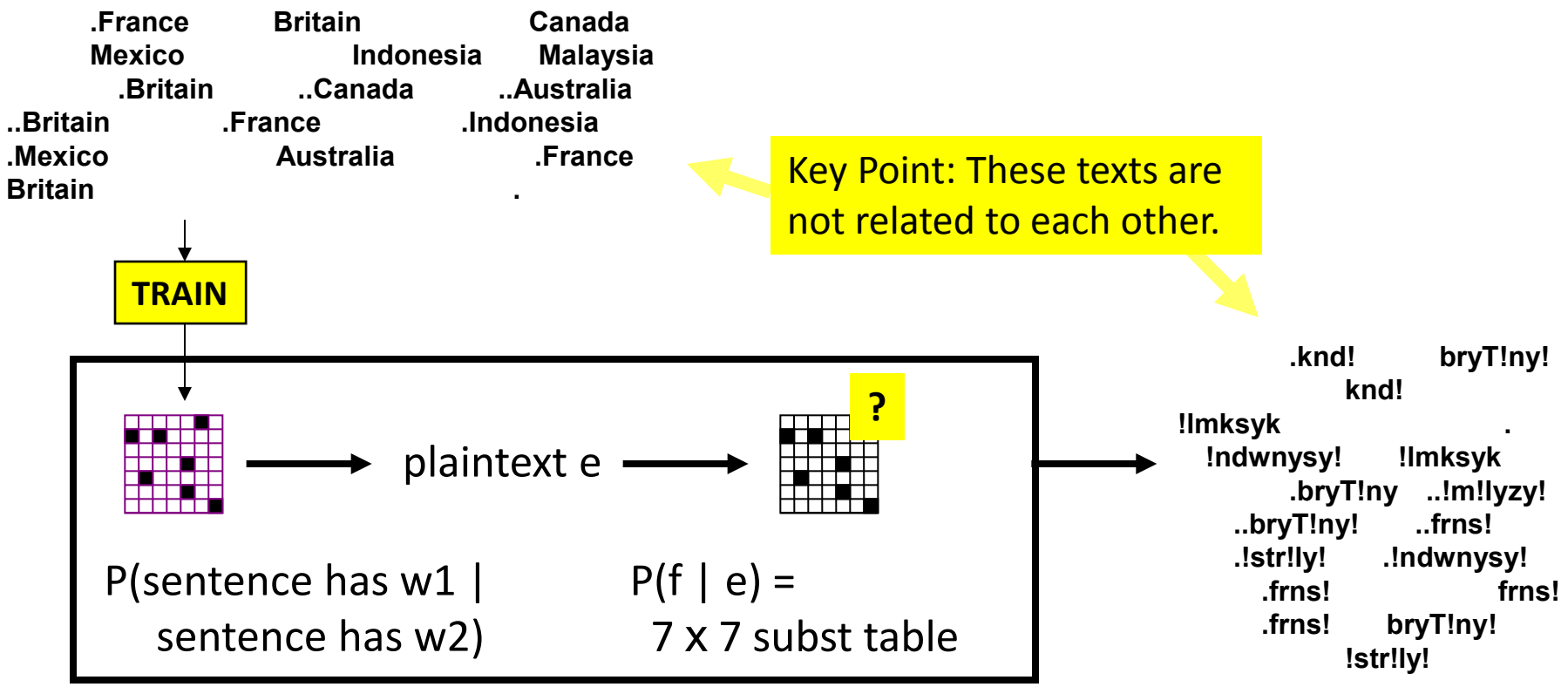
[Knight et al 06]

Word Substitution Cipher



[Knight et al 06]

Word Substitution Cipher



Australia	→	**!str!ly!**	**(0.93)**	!ndwnysy!	(0.03)	m!lyzy! (0.02)
Britain	→	**bryT!ny!**	**(0.98)**	!ndwnysy!	(0.01)	!str!ly! (0.01)
Canada	→	**knd!**	**(0.57)**	frns!	(0.33)	m!lyzy! (0.06)
France	→	**frns!**	**(1.00)**			
Indonesia	→	**!ndwnysy!**	**(1.00)**			
Malaysia	→	**m!lyzy!**	**(0.93)**	lmksyk	(0.07)	
Mexico	→	**!lmksyk**	**(0.91)**	m!lyzy!	(0.07)	

[Knight et al 06]

Time Expressions

!l@!m
!lywm
!lth!ny&
!l@!m !lm!Dy
Sfr
@!m
th!ny&
@!m 1992
@!m 1993
ywm
!l!sbw@ !lm!Dy
fy !ldqyq&
!lsn& !lj!ry&
!lsn&
!lsh=hr !lm!Dy
!lsh=hr !lj!ry
snw!t
sn&
=hdh! !l@!m
s!@&
!l@Sr
@!m 1991

@!m 1990
w!lth!ny&
fy !lywm
mn !lsh=hr !lj!ry
!lqrn
!'y!m
@!m!aN
!ls!@&
17 shb!T 1994
thl!th snw!t
dqyq&
=hdh=h !lsn&
ywmyn
mn !l@!m !lm!Dy
!lsn& !lmqbl&
fy !lsn&
kl ywm
fy !l@!m !lm!Dy

!l@Swr
=hdh! !lsh=hr
fy ywm
nys!n
!sbw@
=hdh=h !l!'y!m
qbl !'y!m
fy !l@Sr
mn !lsn&
!lsnw!t
b@d ywm
!l!y!m
13 nys!n 1994
!lth!ny& @sh!&
thl!th& !y!m
qbl !sbw@yn
fy !lywm !lt!ly
sh@b!n
tmwz
3 dhw !lHj& 1414
fy shb!T !lm!Dy
qbl ywmyn

Time Expressions

<n><n>* **???** 19<n><n>

9 Hzyr!n 1942
8 tshryn !l!wl 1990
7 k!nwn !l!wl 1993
6 !'y!r 1993
6 !~Adh!r 1991
5 shb!T 1950
4 Hzyr!n 1989
30 !~Adh!r 1944
29 !y!r 1945
29 !~Adh!r 1993
28 k!nwn !l!'wl 1994
27 tmwz 1993
26 tmwz 1953
26 shb!T 1993
26 k!nwn !l!wl 1994
25 !ylwl 1926
24 !~Adh!r 1993
22 !ylwl 1957
22 tshryn !l!wl 1948
22 tmwz 1952
21 !y!r 1994
21 k!nwn !l!wl 1988
21 Hzyr!n 1967
20 !'y!r 1990
20 tshryn !'wl 1983
20 tshryn !l!'wl 1921
1 !y!r 1994
17 Hzyr!n 1972
16 !ylwl 1919
16 Hzyr!n 1984
16 !~Ab 1929

Time Expressions

<n> Hzyr!n <n>

13	4 Hzyr!n 1967	2	fy 30 Hzyr!n 1995
12	fy 12 Hzyr!n 1993	2	fy 18 Hzyr!n 1994
7	5 Hzyr!n 1967	2	fy 14 Hzyr!n 1993
6	fy 30 Hzyr!n 1989	2	fy 14 Hzyr!n 1991
6	30 Hzyr!n 1989	2	fy 12 Hzyr!n 1990
4	fy 30 Hzyr!n 1994	2	7 Hzyr!n 1994
4	fy 30 Hzyr!n 1993	2	6 Hzyr!n 1941
3	fy 19 Hzyr!n 1967	2	26 Hzyr!n 1994
2	ywm 30 Hzyr!n 1989	2	21 Hzyr!n 1994
2	w 6 Hzyr!n 1994	2	1 Hzyr!n 1994
2	qbl 5 Hzyr!n 1967	2	19 Hzyr!n 1965
2	fy 9 Hzyr!n 1967	2	18 Hzyr!n 1994
2	fy 7 Hzyr!n 1981	2	18 Hzyr!n 1940
2	fy 6 Hzyr!n 1994	2	12 Hzyr!n 1993
2	fy 5 Hzyr!n 1967	2	11 Hzyr!n 1994

Time Expr

<n> Hzyr!n <n>

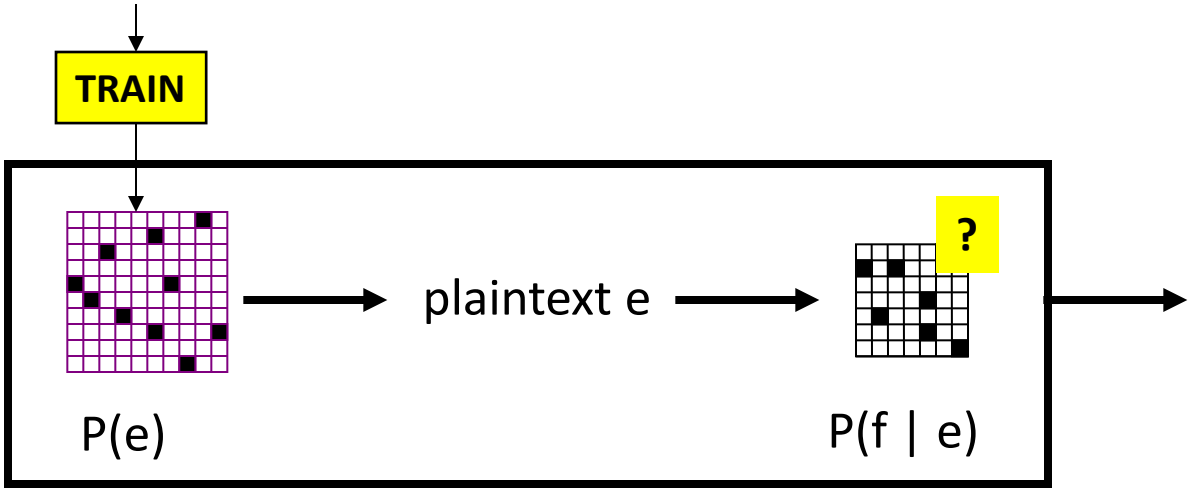
13	4 Hzyr!n 1967
12	fy 12 Hzyr!n 1993
7	5 Hzyr!n 1967
6	fy 30 Hzyr!n 1989
6	30 Hzyr!n 1989
4	fy 30 Hzyr!n 1994
4	fy 30 Hzyr!n 1993
3	fy 19 Hzyr!n 1967
2	ywm 30 Hzyr!n 1989
2	w 6 Hzyr!n 1994
2	qbl 5 Hzyr!n 1967
2	fy 9 Hzyr!n 1967
2	fy 7 Hzyr!n 1981
2	fy 6 Hzyr!n 1994
2	fy 5 Hzyr!n 1967

Search query	Documents
January 4, 1967	8040
February 4, 1967	9270
March 4, 1967	10700
April 4, 1967	21800
May 4, 1967	14000
June 4, 1967	39300
July 4, 1967	12600
August 4, 1967	7970
September 4, 1967	7390
October 4, 1967	8800
November 4, 1967	6560
December 4, 1967	9770

Foreign Language as a Cipher

BAGHDAD, Iraq (CNN) -- Six bombings killed at least 54 Iraqis and wounded 96 others Wednesday, including 20 civilians who died as they lined up to join the Iraqi army in Hawija when a suicide bomber detonated explosives hidden under his clothing, Iraqi officials said.
That attack in the town about 130 miles (209 kilometers) north of Baghdad also wounded 30 Iraqis, said Iraqi army Lt. Col. Khalil al-Zawbai.
A car bombing in Saddam Hussein's ancestral homeland of Tikrit also killed 30 Iraqis and wounded another 40, Iraqi officials said. The Tikrit explosion

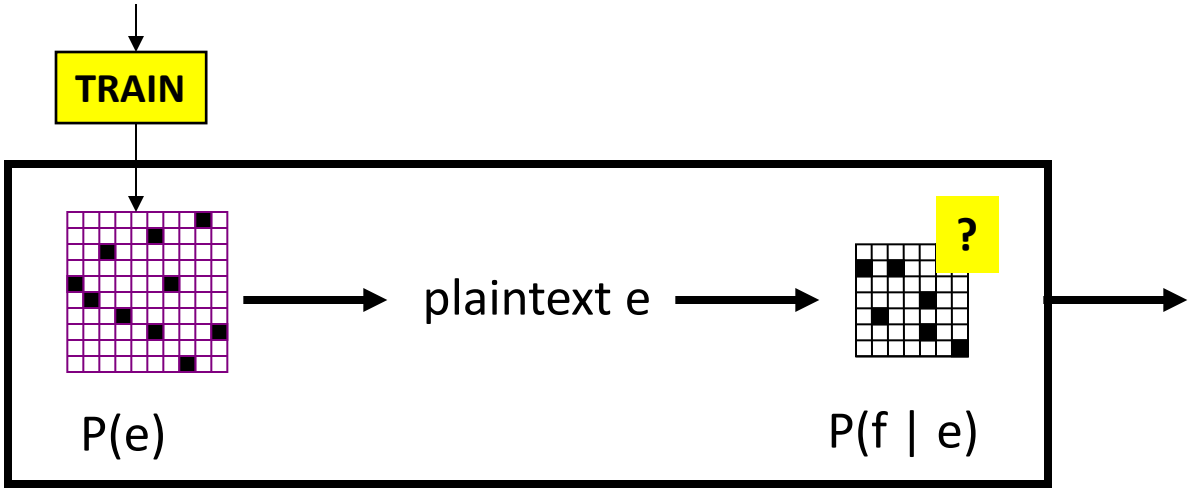
Key Point: These texts are not related to each other.



رفض رئيس السلطة الفلسطينية محمود عباس مجددا تصريحات وزير الخارجية الإسرائيلي سيلفان شالوم التي قال فيها إنه يتعين على إسرائيل إعادة النظر في انسحابها من غزة، المقرر أن يتم الصيف المقبل إذا فازت حركة المقاومة الإسلامية حماس في الانتخابات التشريعية.
وقال عباس في مؤتمر صحفي على هامش مشاركته في القمة العربية-اللاتينية الأولى إنه يتعين على إسرائيل احترام خيار الشعب الفلسطيني حتى لو فازت حماس بالانتخابات، وأضاف "إذا نجحت حماس أو فتح سيكون هذا خيار الشعب الفلسطيني، وعلى الجميع قبول هذا الخيار بكل ترحاب".
من جانبه شجب رئيس الحكومة الفلسطينية أحمد قريع الطابع الأحادي الجانب للانسحاب الإسرائيلي من غزة، وأكد أن إسرائيل تريد مغادرة هذه الأراضي لتعزيز سيطرتها على الضفة الغربية.
وقال قريع في كلمة له خلال مؤتمر نظمته وزارة الأوقاف في رام الله "سينسحبون من غزة ولكننا لا نعرف ما هو شكل هذا الانسحاب وماذا سيتركون، وما هو مصير المعابر والحدود, وكل ذلك غامض لأنه قرار أحادي الجان

Foreign Language as a Cipher

BAGHDAD, Iraq (CNN) -- Six bombings killed at least 54 Iraqis and wounded 96 others Wednesday, including 20 civilians who died as they lined up to join the Iraqi army in Hawija when a suicide bomber detonated explosives hidden under his clothing, Iraqi officials said.
That attack in the town about 130 miles (209 kilometers) north of Baghdad also wounded 30 Iraqis, said Iraqi army Lt. Col. Khalil al-Zawbai.
A car bombing in Saddam Hussein's ancestral homeland of Tikrit also killed 30 Iraqis and wounded another 40, Iraqi officials said. The Tikrit explosion



رفض رئيس السلطة الفلسطينية محمود عباس مجددا تصريحات وزير الخارجية الإسرائيلي سيلفان شالوم التي قال فيها إنه يتعين على إسرائيل إعادة النظر في انسحابها من غزة، المقرر أن يتم الصيف المقبل إذا فازت حركة المقاومة الإسلامية حماس في الانتخابات التشريعية.
وقال عباس في مؤتمر صحفي على هامش مشاركته في القمة العربية-اللاتينية الأولى إنه يتعين على إسرائيل احترام خيار الشعب الفلسطيني حتى لو فازت حماس بالانتخابات، وأضاف "إذا نجحت حماس أو فتح سيكون هذا خيار الشعب الفلسطيني، وعلى الجميع قبول هذا الخيار بكل ترحاب".
من جانبه شجب رئيس الحكومة الفلسطينية أحمد قريع الطابع الأحادي الجانب للانسحاب الإسرائيلي من غزة، وأكد أن إسرائيل تريد مغادرة هذه الأراضي لتعزيز سيطرتها على الضفة الغربية.
وقال قريع في كلمة له خلال مؤتمر نظمته وزارة الأوقاف في رام الله "سينسحبون من غزة ولكننا لا نعرف ما هو شكل هذا الانسحاب وماذا سيتركون، وما هو مصير المعابر والحدود, وكل ذلك غامض لأنه قرار أحادي الجان

Exploiting Giga-Scale Non-Parallel Text

(Dou & Knight subm.)

Accuracy of learned bilingual dictionary

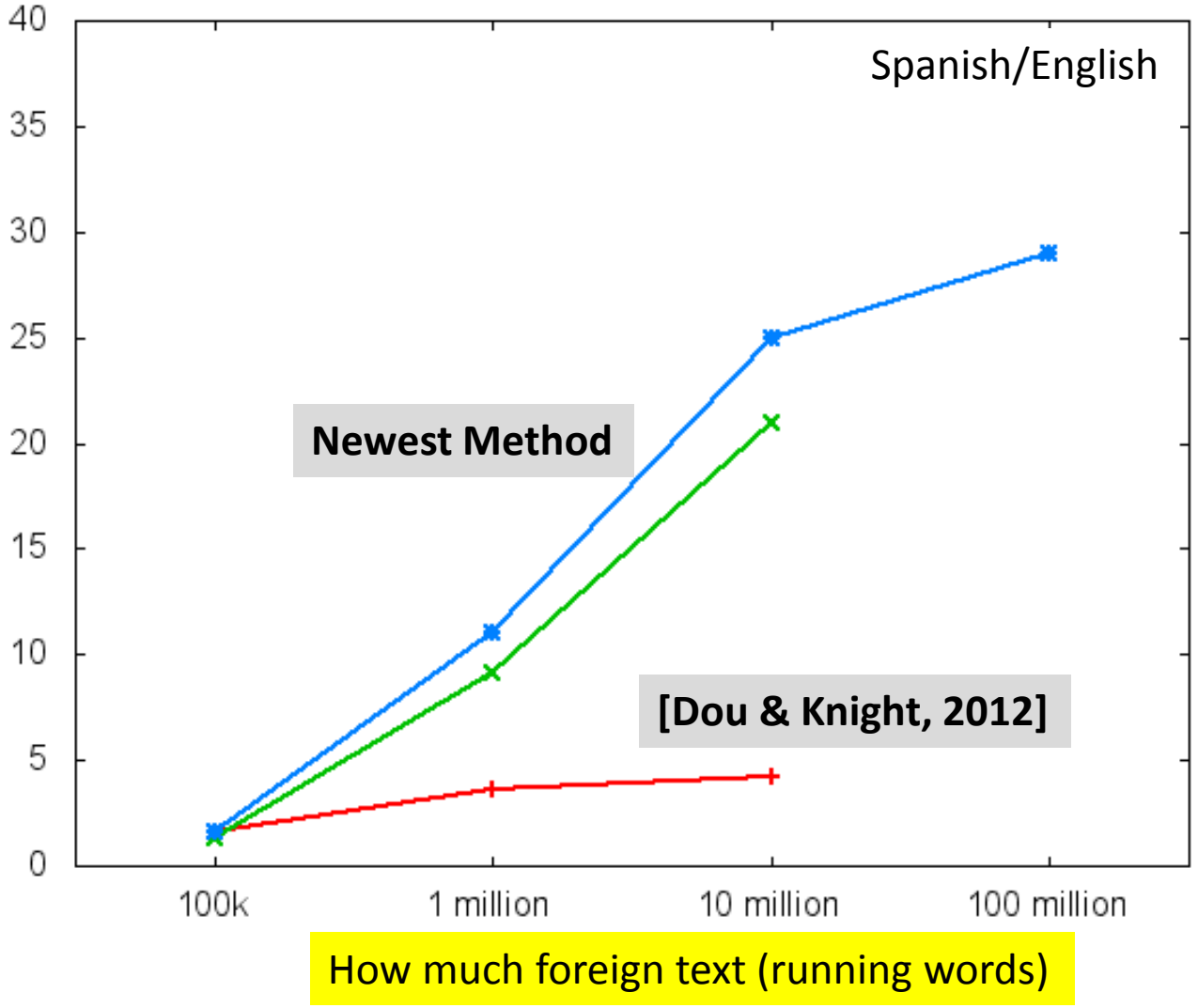
not translations of each other

English text

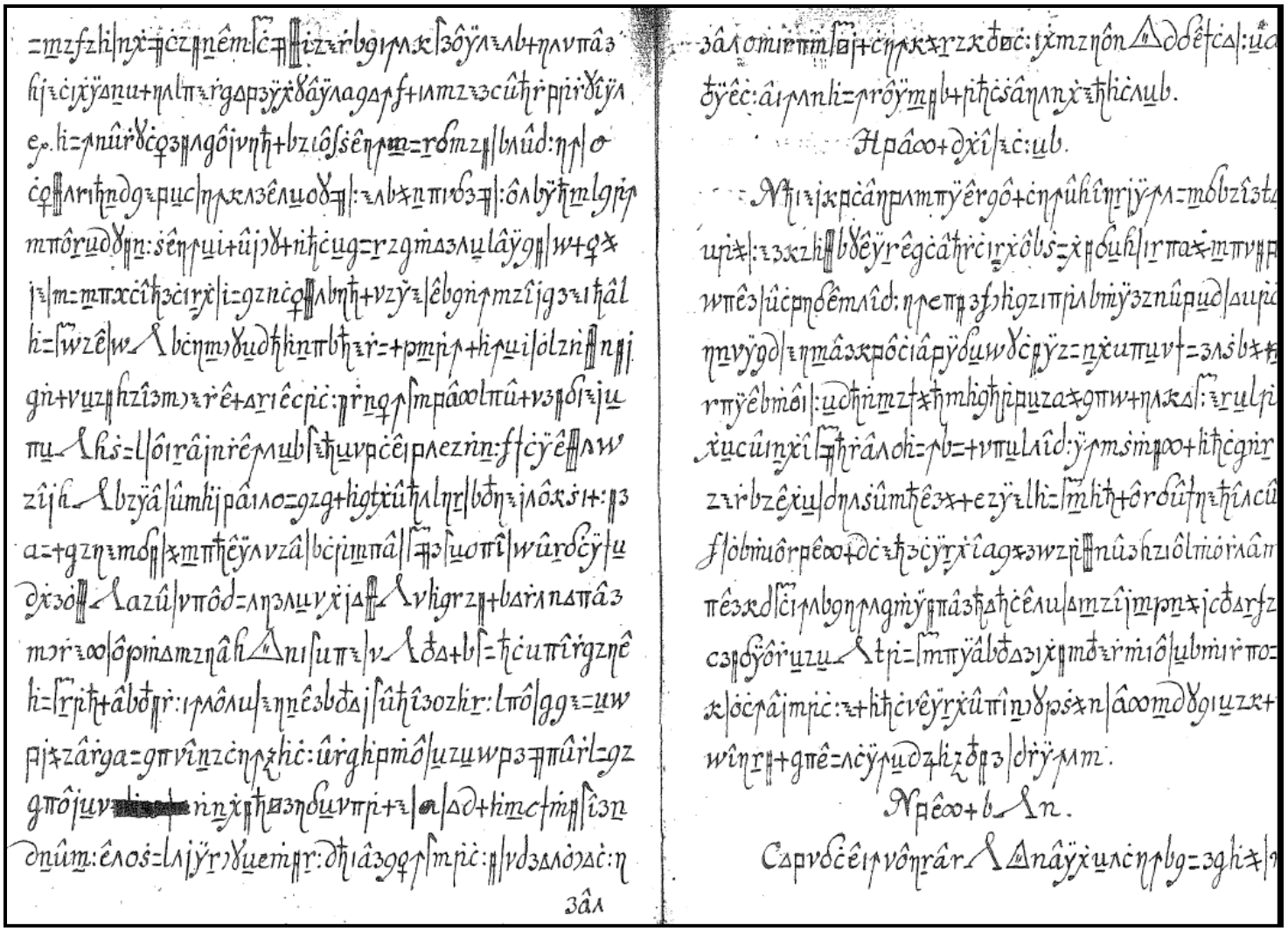
Foreign text

deciphering engine

bilingual word-for-word dictionary



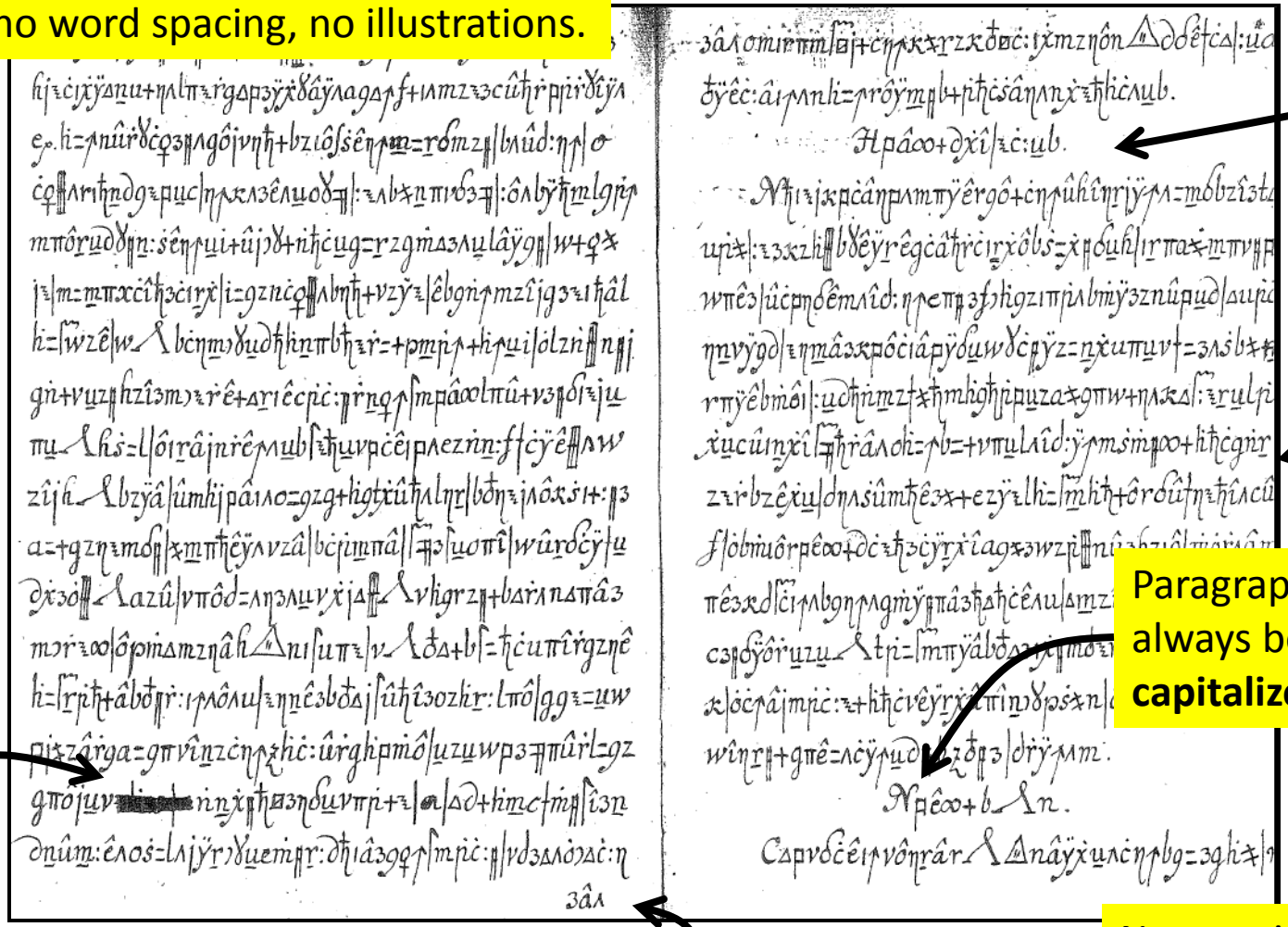
Copiale Cipher



[Knight, Megyesi, Schaefer 11]

Copiale Cipher

105 pages, 75000 letter tokens, no word spacing, no illustrations.



Section headers

Lines ≈ equal length

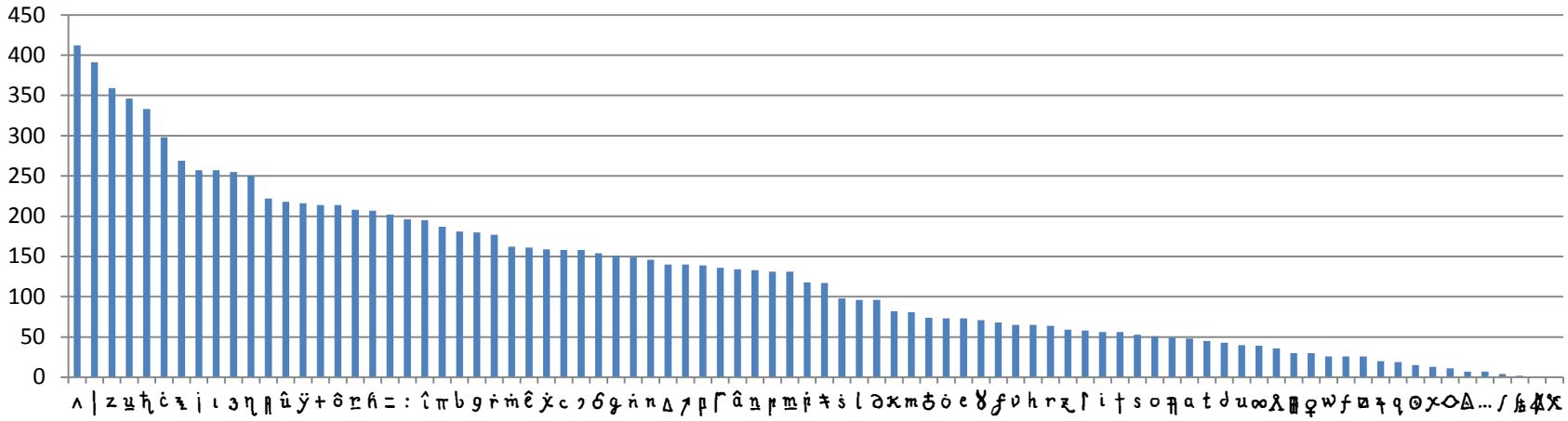
Paragraphs and section titles always begin with **capitalized Roman letters**.

Non-enciphered inscriptions: **Copiales 3** and **Philipp 1866**

Some scratch-outs, rare

Preview text fragments (“catchwords”)

Letter Frequencies



digraphs:		trigraphs:	
ɔ ɧ	99	ɔ ɧ ʌ	47
ċ :	66	ċ : u̲	23
ɧ ʌ	49	ɳ ɔ ɧ	22
: u̲	48	ÿ ɔ ɧ	18
z ʁ	44	ɦ ċ |	17

tendencies:

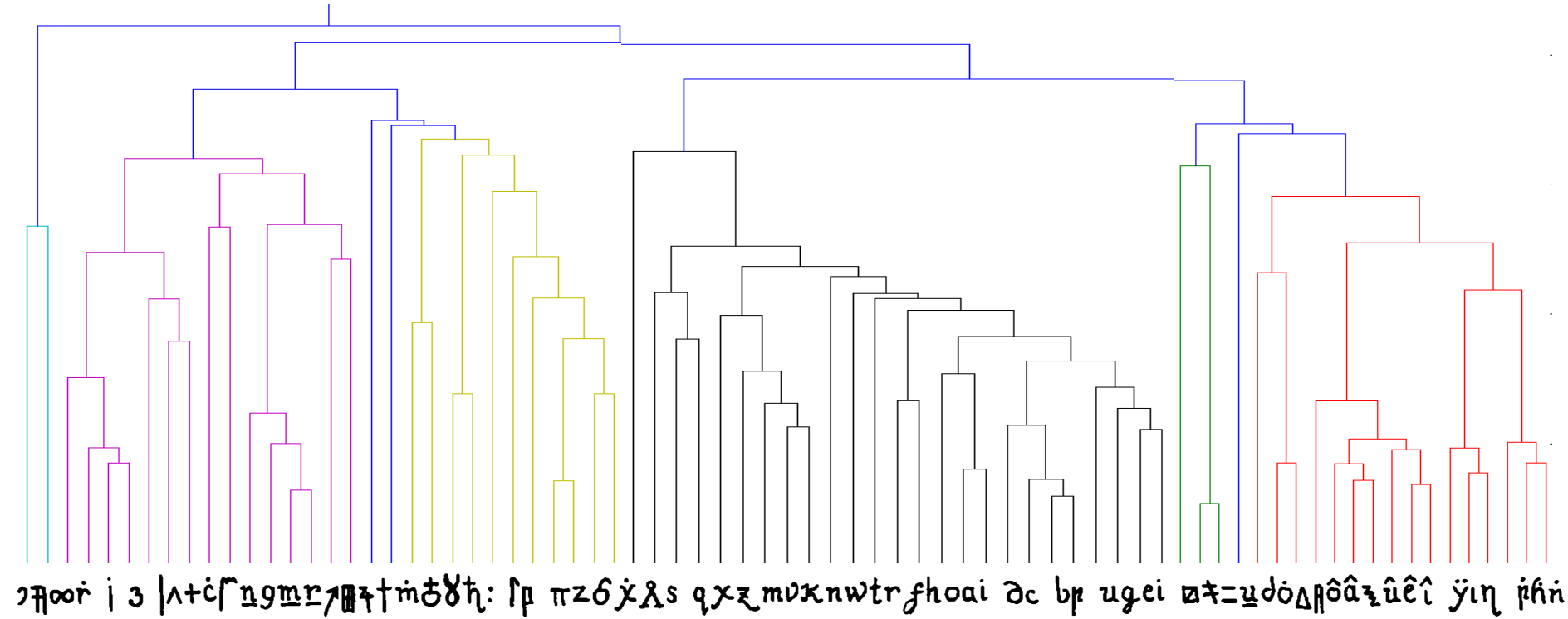
â, ê, î, ô, û followed by ʒ and ¡

â, ê, î, ô, û preceded by z and π

Clustering of Cipher Letters

letters grouped if they have similar contexts (L/R neighbors)

Scipy software

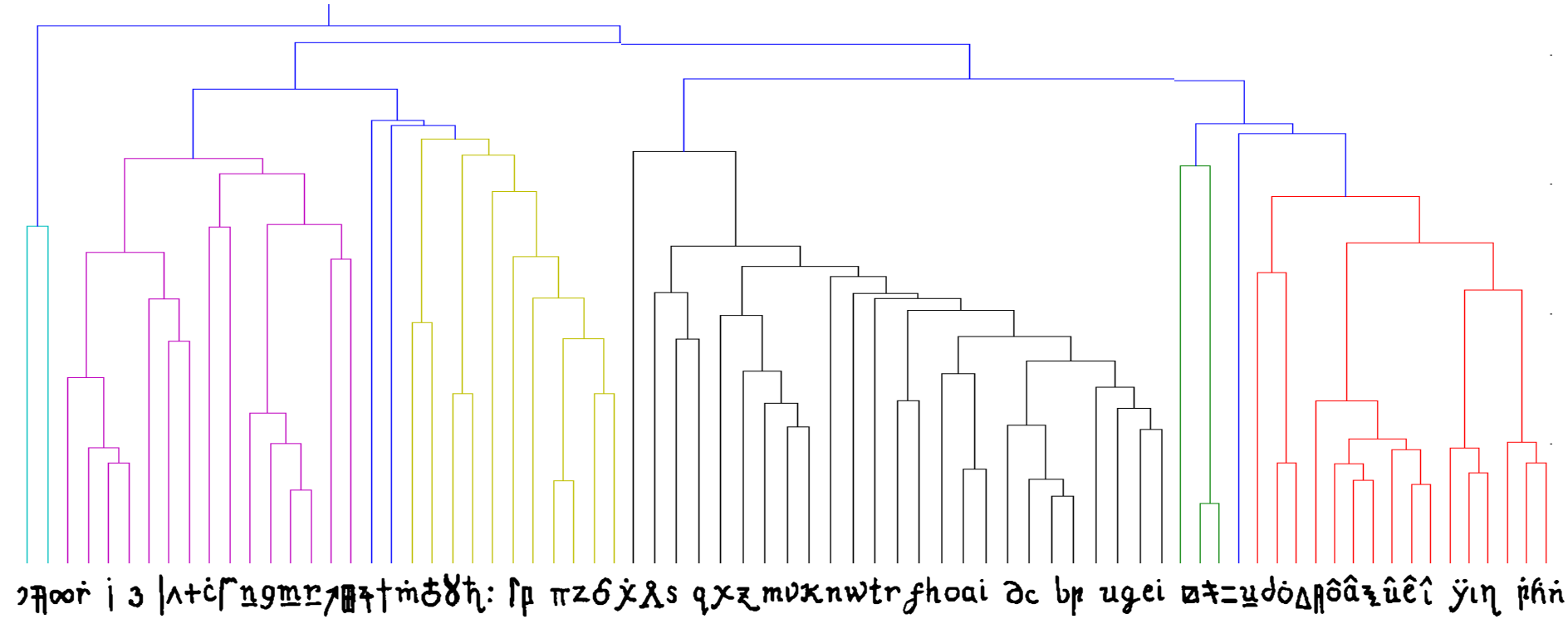


thanks Jon Graehl

Clustering of Cipher Letters

letters grouped if they have similar contexts (L/R neighbors)

Scipy software



underlined letters

unaccented Roman letters

circumflexed vowels

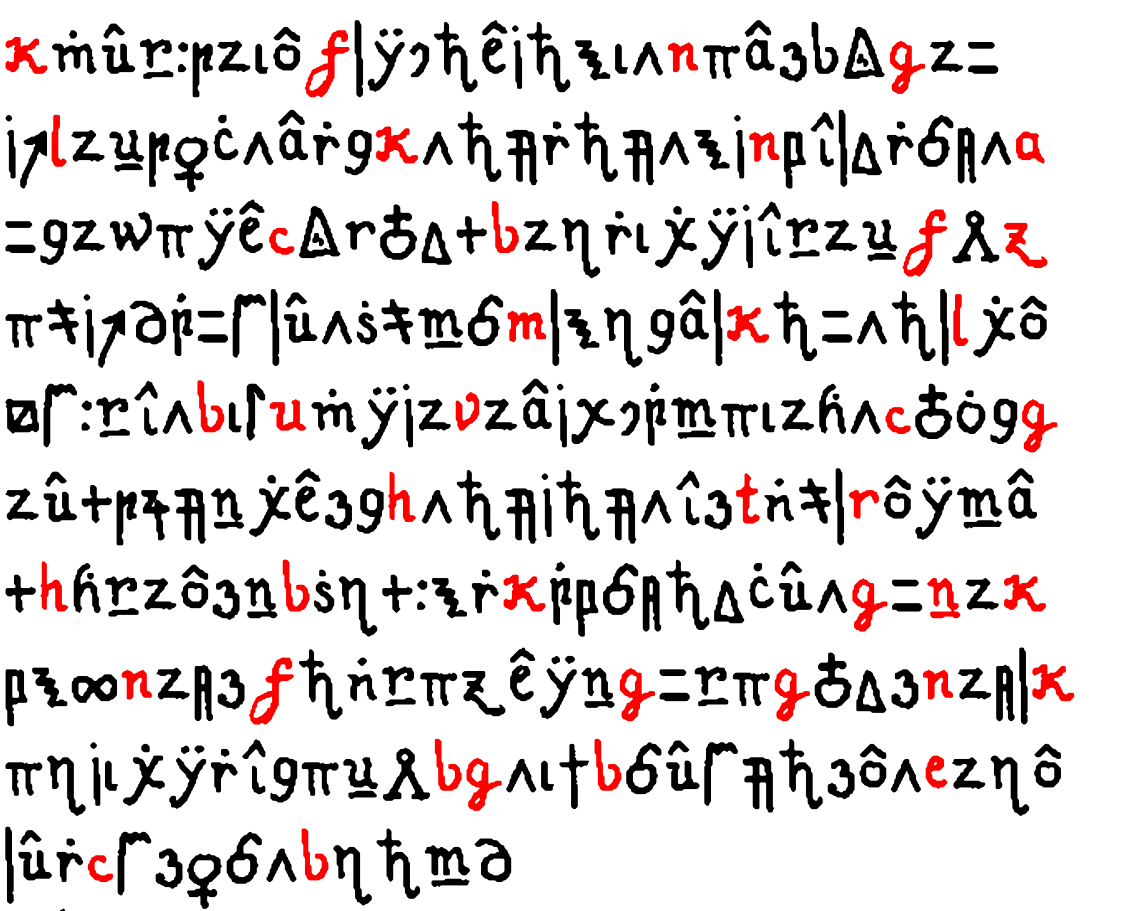
thanks Jon Graehl

First Decipherment Approach

unaccented Roman letters that cluster:

a b c d e f g h i
k l m n o p q r s
t u v w x y z

most common letter = 12%
least common = very small



k f n g l k n a c b f z m k
l b u v c g h t r h b k g n k n
f g g n k b g b e c b ...

Decipher against
80 plaintext languages.

Second Decipherment Approach

Homophonic cipher, e.g.:

A = 𝔄 𝔦 𝔩 𝔶 𝔯

B = 𝔲

C = 𝔬 𝔫

D = 𝔄

E = 𝔵 𝔣 𝔇 𝔨 𝔣 𝔦 𝔷 𝔷

F = 𝔭

G = 𝔶

etc.

𝔨𝔪𝔲𝔫:𝔭𝔷𝔦𝔬𝔣|𝔶𝔬𝔥𝔢|𝔥𝔷𝔦𝔞𝔫𝔭𝔞𝔷𝔟𝔇𝔤𝔷=
𝔦𝔩𝔩𝔷𝔲𝔭𝔮𝔠𝔞𝔞𝔯𝔤𝔨𝔞𝔥𝔄𝔯𝔥𝔄𝔞𝔷𝔦𝔫𝔭𝔦|𝔇𝔯𝔬𝔄𝔞𝔞
=𝔤𝔷𝔴𝔫𝔶𝔢𝔠𝔇𝔯𝔡𝔇+𝔟𝔷𝔥𝔯𝔦𝔵𝔶𝔶|𝔦𝔫𝔷𝔲𝔣𝔄𝔷
𝔫𝔨|𝔩𝔡𝔯=𝔣|𝔲𝔞𝔰𝔨𝔪𝔬𝔪|𝔷𝔥𝔤𝔞|𝔨𝔥=𝔞𝔥||𝔩𝔵𝔬
𝔴𝔣:𝔫𝔦𝔞𝔟𝔦𝔣𝔲𝔪𝔶|𝔷𝔳𝔷𝔞|𝔵𝔬𝔯𝔪𝔫𝔦𝔷𝔥𝔞𝔠𝔡𝔬𝔤𝔤
𝔷𝔲+𝔭𝔨𝔄𝔫𝔵𝔢𝔷𝔤𝔥𝔞𝔥𝔄𝔦𝔥𝔄𝔞𝔦𝔷𝔱𝔫𝔨|𝔯𝔬𝔶𝔪𝔞
+𝔥𝔥𝔫𝔷𝔬𝔷𝔫𝔟𝔰𝔥+:𝔷𝔯𝔨𝔯𝔭𝔬𝔄𝔥𝔇𝔠𝔲𝔞𝔤=𝔫𝔷𝔨
𝔭𝔷∞𝔫𝔷𝔄𝔷𝔣𝔥𝔫𝔫𝔫𝔷𝔢𝔶𝔫𝔤=𝔫𝔫𝔤𝔡𝔇𝔷𝔫𝔷𝔄|𝔨
𝔫𝔥𝔦𝔵𝔶𝔯𝔦𝔤𝔫𝔲𝔄𝔟𝔤𝔞𝔦+𝔟𝔬𝔲𝔣𝔄𝔥𝔷𝔬𝔞𝔷𝔥𝔬
|𝔲𝔯𝔠𝔣𝔷𝔭𝔬𝔞𝔟𝔥𝔥𝔪𝔡

Homophonic Cipher

Result of computer attack on Copiale, using 80 possible plaintext languages?

FAIL

But, slight numerical preference for German

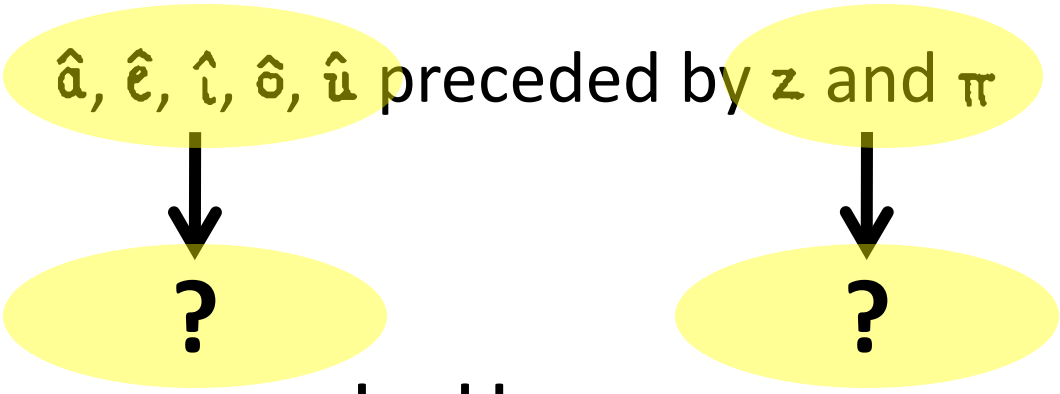
Cipher Characteristics

digraphs:		trigraphs:	
ɔ ɬ	99	ɔ ɬ ʌ	47
ċ :	66	ċ : u̲	23
ɬ ʌ	49	ɳ ɔ ɬ	22
: u̲	48	ÿ ɔ ɬ	18
z ǁ	44	ɦ ċ |	17

tendencies:

â, ê, î, ô, û followed by ɜ and ¡

â, ê, î, ô, û preceded by z and π



should appear
adjacent in German text

Make full digraph table for cipher and for German

Key Observation #1

In Copiale, ʔ almost always followed by ħ

In German, C almost always followed by H

(German CH is like English QU)

So guess: ʔ = C, ħ = H

One Thing Leads to Another

ɔħ = CH → ɔħʌ = CHT → ʌ = T ?

Each step is guesswork.

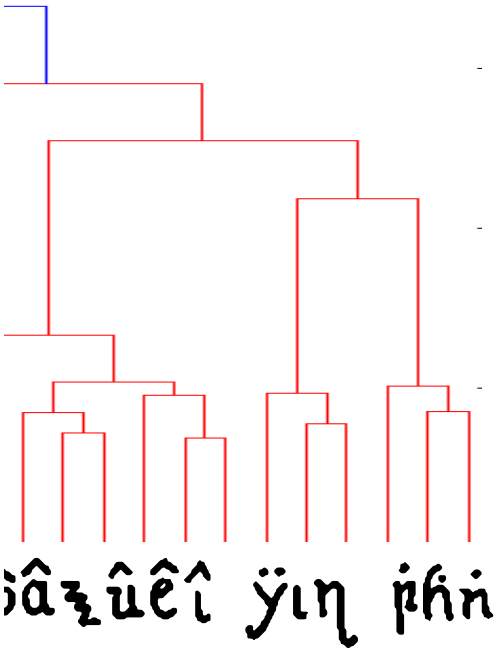
Must be willing to retract.

Weird task, not knowing German.

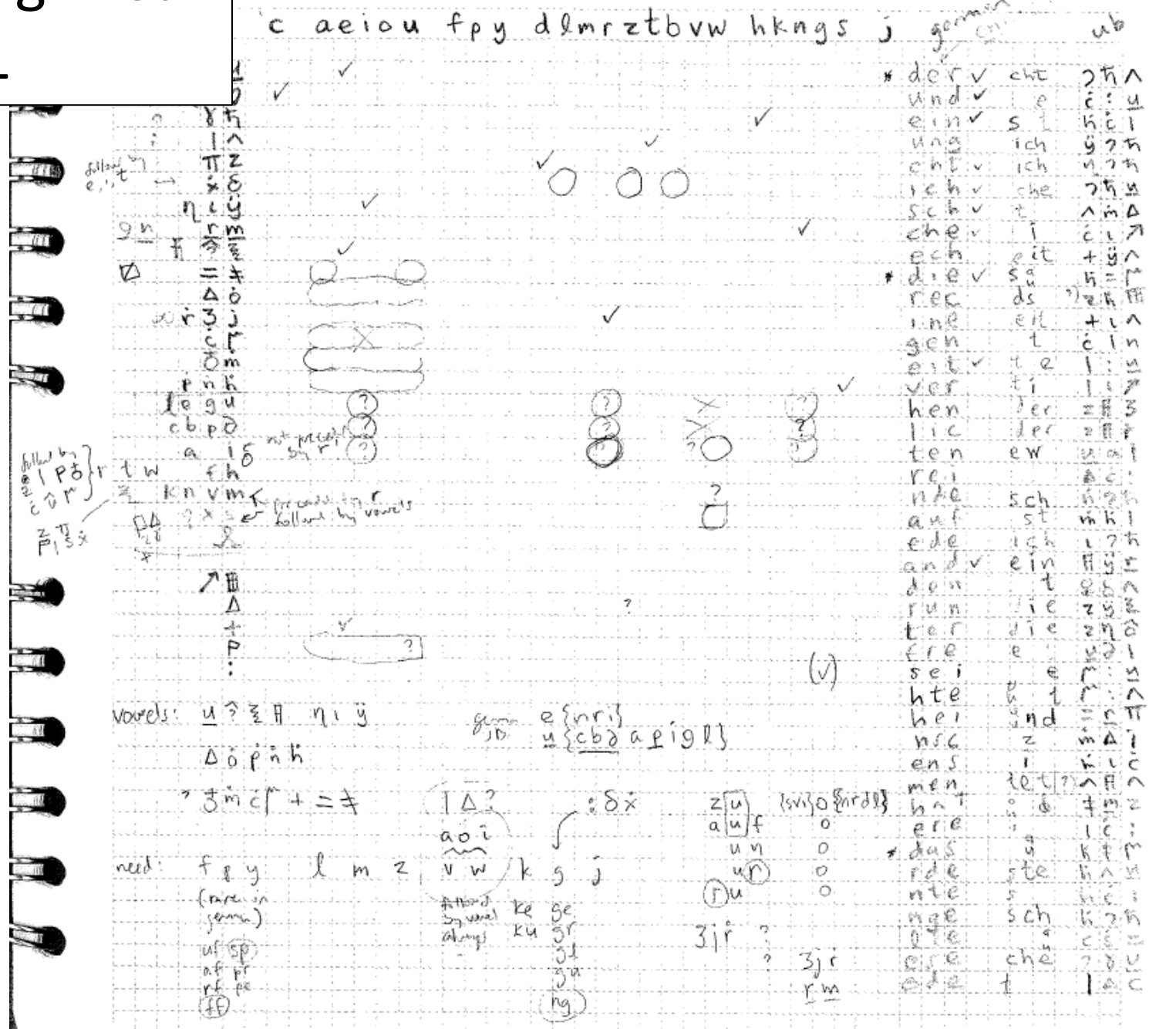
No longer care what the book says.

Cluster diagram crucial:

ÿ = I → ɩ = I , ɳ = I



Spring Break 2011

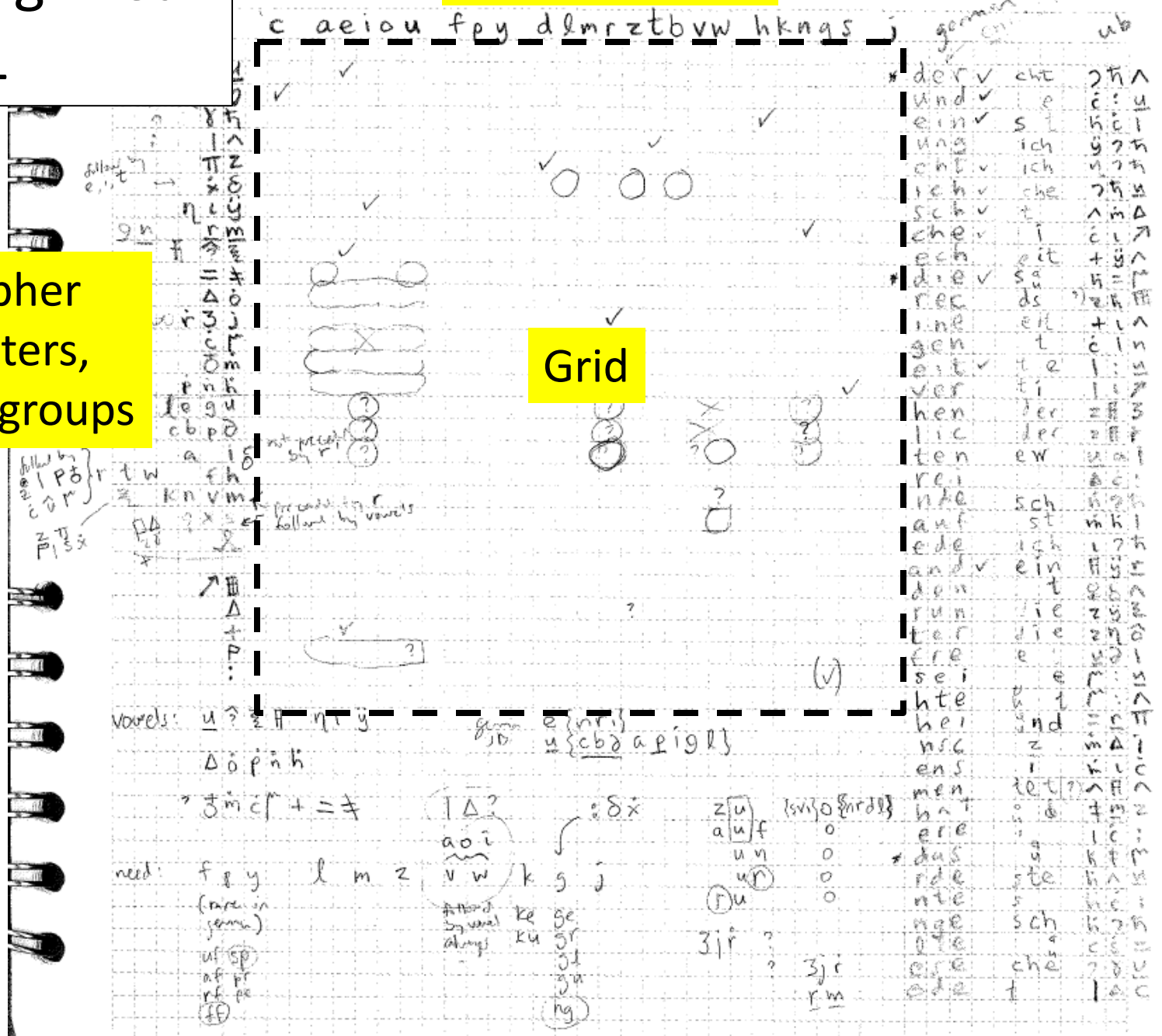


Spring Break 2011

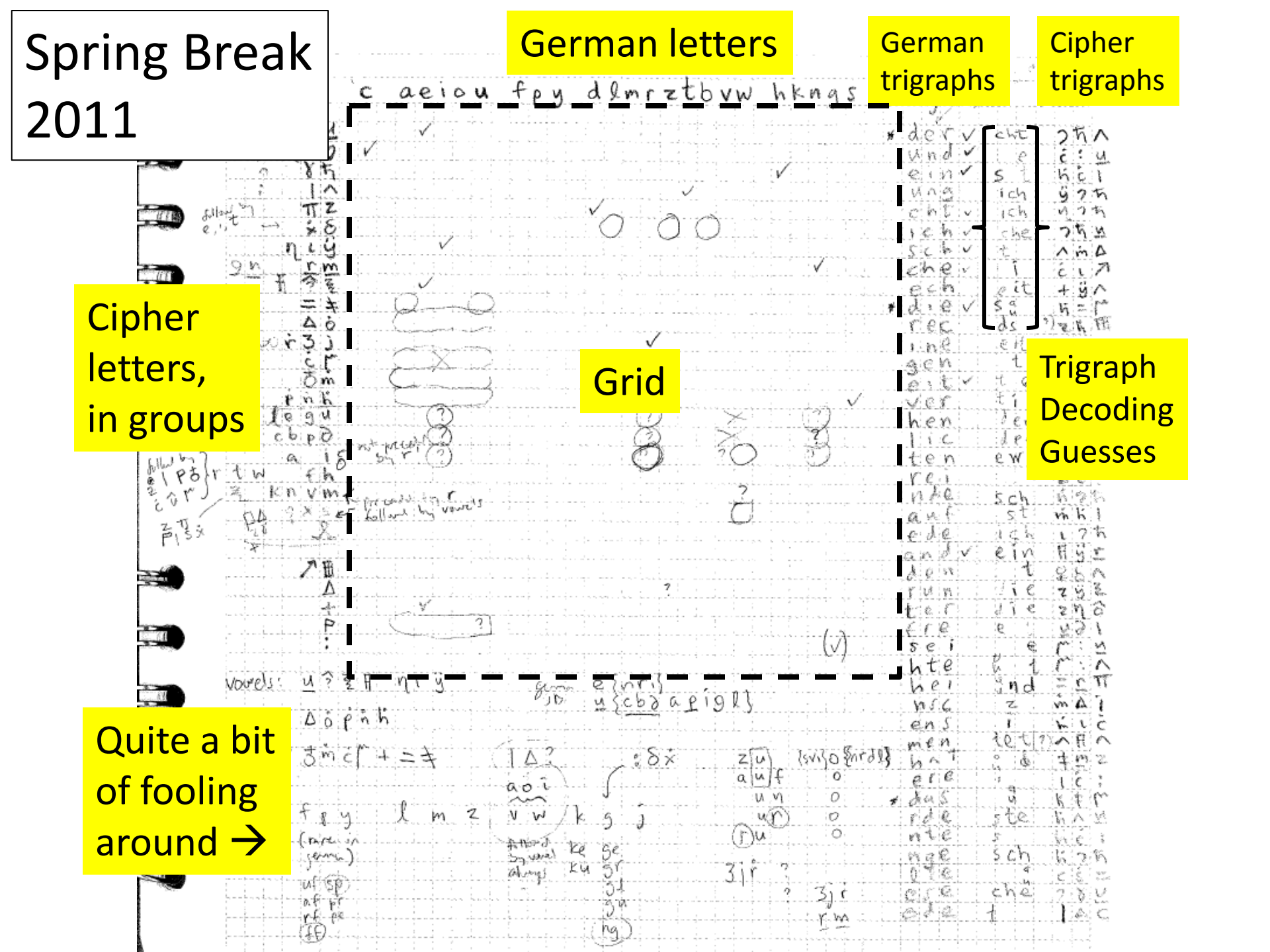
German letters

Cipher letters, in groups

Grid



Spring Break 2011



Spring Break 2011

German letters

German trigraphs

Cipher trigraphs

Cipher letters, in groups

Grid

Trigraph Decoding Guesses

Partially deciphered text

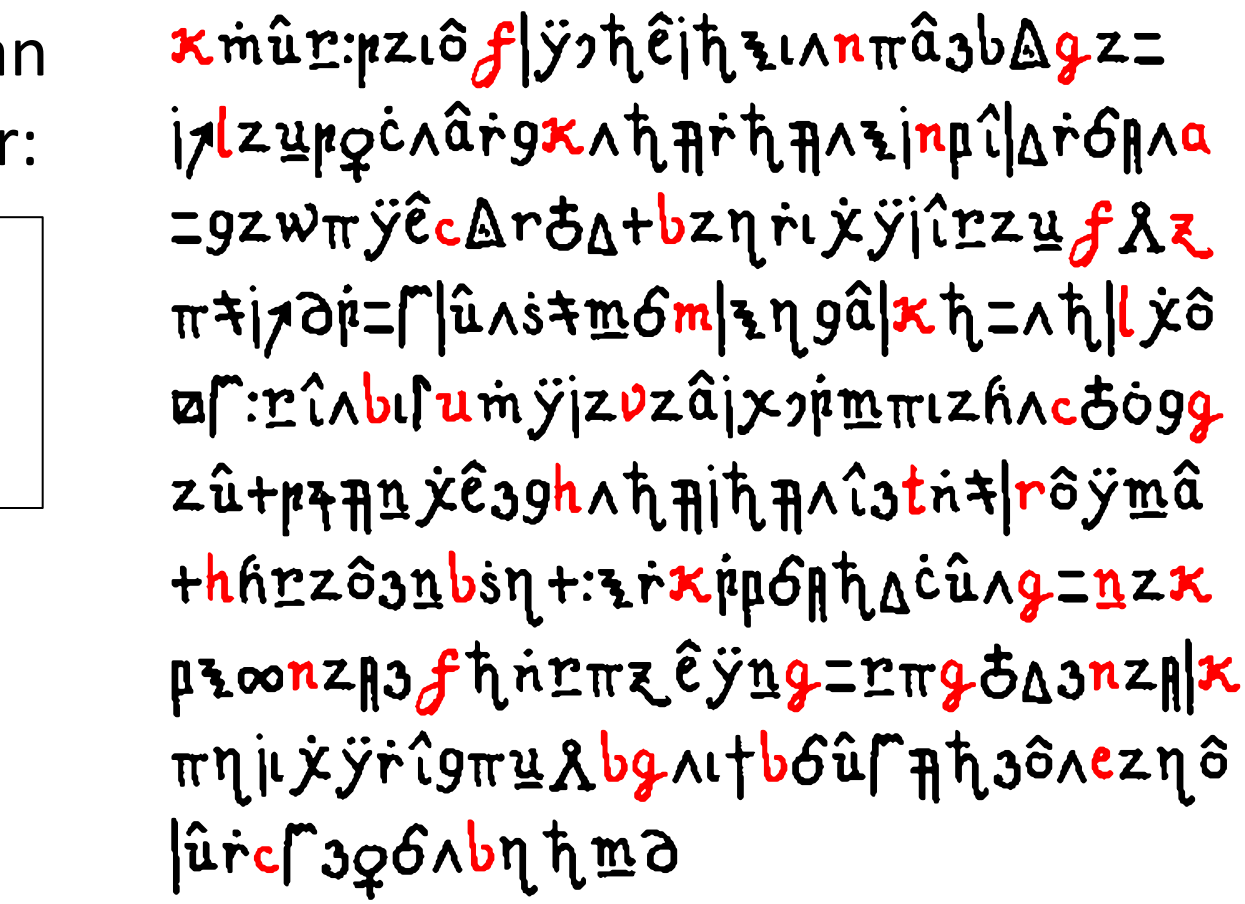
?GEHEIMER?UNTERLIST?VOR?DIE?GESELLE
?ERDER?TITUL
?CEREMONIE?DER?AUFNAHME

Quite a bit of fooling around →

Key Observation #2

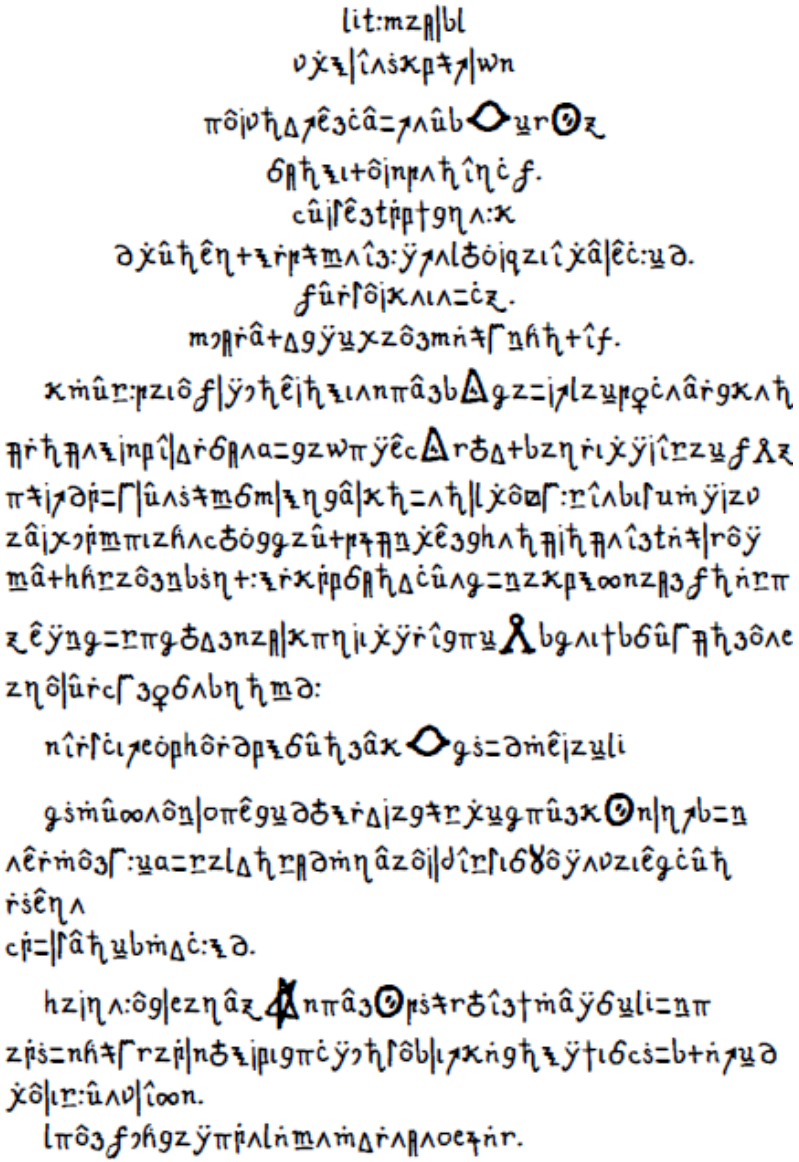
unaccented Roman letters that cluster:

a b c ə e f g h i
k l m n o p q r s
t u v w x y z



Actually, those are space bars

Copiale Decipherment



gesetz buchs
der hocherleuchte ◇ e ⊙
geheimer theil.
erster abschnitt
geheimer unterricht vor die gesellen.
erster titul.
ceremonien der aufnahme.

wenn die sicherheit der △ durch den ältern thürheter besorget und die △ vom dirigirenden 人 durch aufsetzung seines huths geöffnet ist wird der candidat von dem jüngern thürhüter aus einem andern zimmer abgeholet und bey der hand ein und vor des dirigirenden 人 tisch geführet dieser frägt ihn:

erstlich ob er begehre ◇ zu werden

zweytens denen verordnungen der ⊙ sich unterwerffen und ohne wiederspenstigkeit die lehrzeit ausstehen wolle.

drittens die 人 der ⊙ gu verschweigen und dazu auf das verbindlichste sich anheischig zu machen gesinnet sey.

der candidat antwortet ja.

Initial translation:

First lawbook

of the  e 

Secret part.

First section

Secret teachings for apprentices.

First title.

Initiation rite.

If the safety of the  is guaranteed, and the  is opened by the chief  , by putting on his hat, the candidate is fetched from another room by the younger doorman and by the hand is led in and to the table of the chief  , who asks him:

First, if he desires to become  .

Secondly, if he submits to the rules of the  and without rebelliousness suffer through the time of apprenticeship.

Thirdly, be silent about the  of the  and furthermore be willing to offer himself to volunteer in the most committed way.

The candidate answers yes.

Voynich Manuscript (VMS)

- Medieval illustrated manuscript (early 1400s)
- 235 pages, 6 sections, 38k word tokens, 35 letter types
- Undeciphered

History of Voynich Manuscript

Wilfrid Michael Voynich
book dealer

William Newbold,
Polymath, PhD UPenn

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:

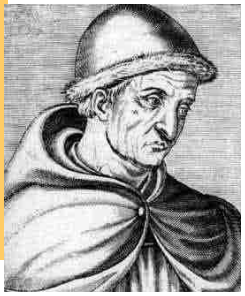
This book bequeathed to me by an intimate friend, I destined for you, **my very dear Athanasius** [**Kircher**], as soon as it came into my possession, for I was convinced that it could be read by no one except yourself. The **former owner** of this book once asked your opinion by letter Accept now this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

1630s George Baresch owns VMS sends letter to Kircher

1639 GB writes Kircher again

16xx Marci inherits VMS from GB

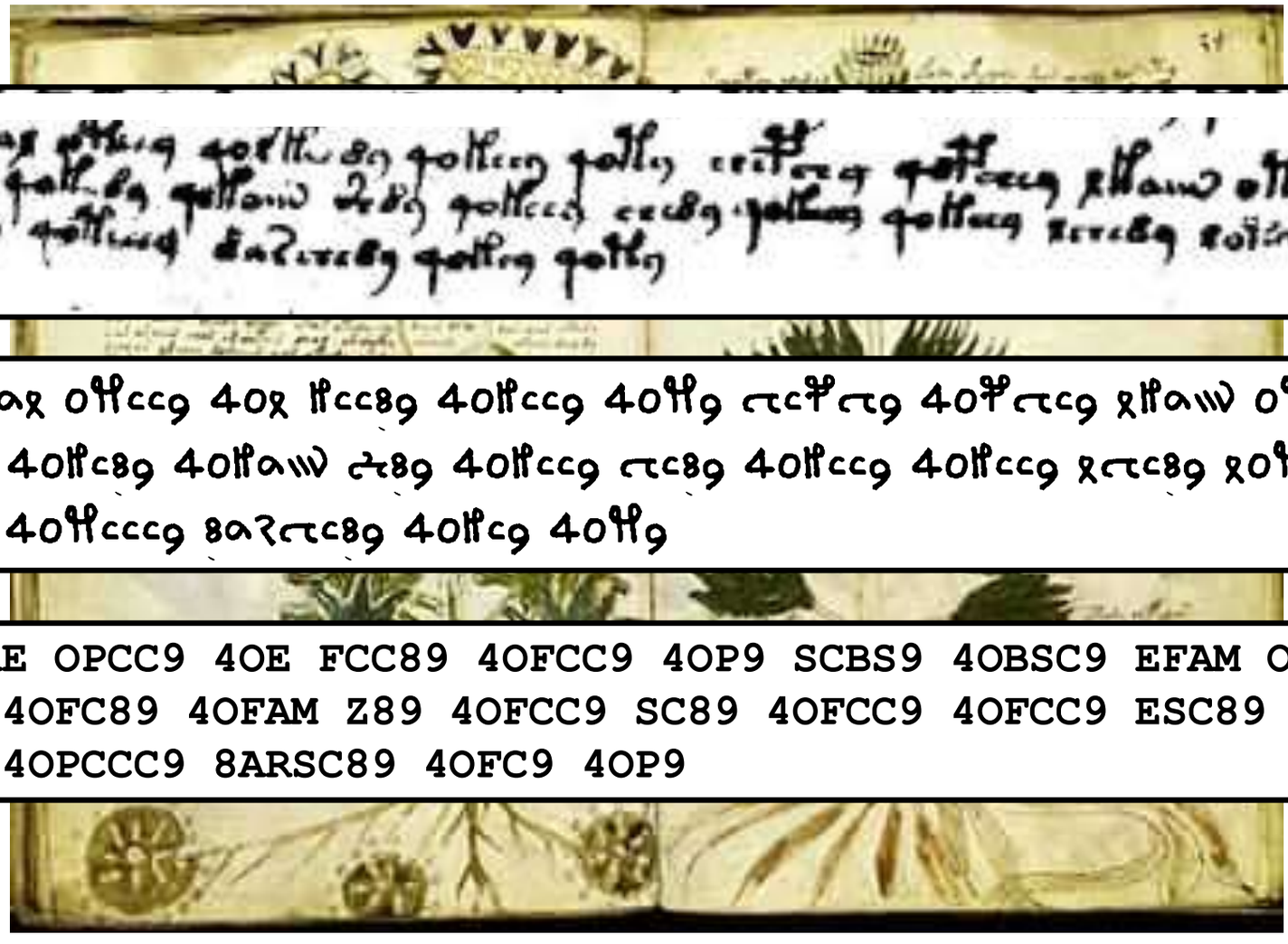
1665 Marci sends VMS to Kircher with letter

1665-80 Kircher owns VMS

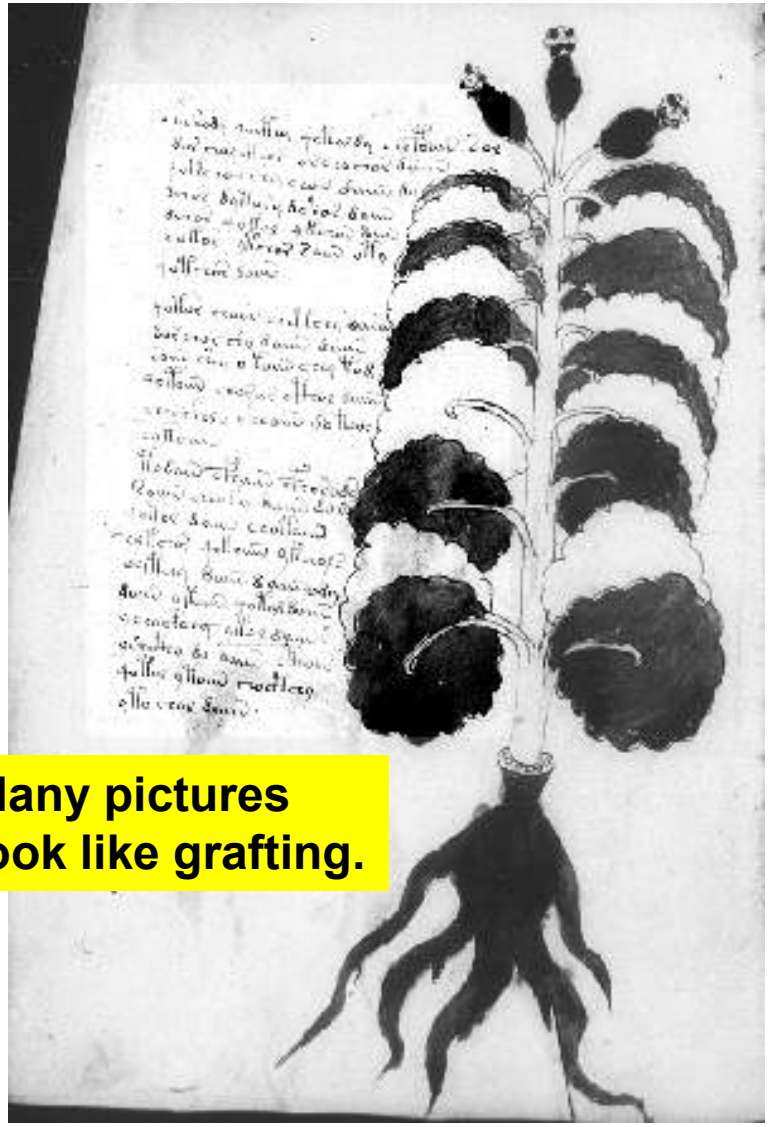
1680 Kircher dies

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

Voynich Manuscript (VMS)



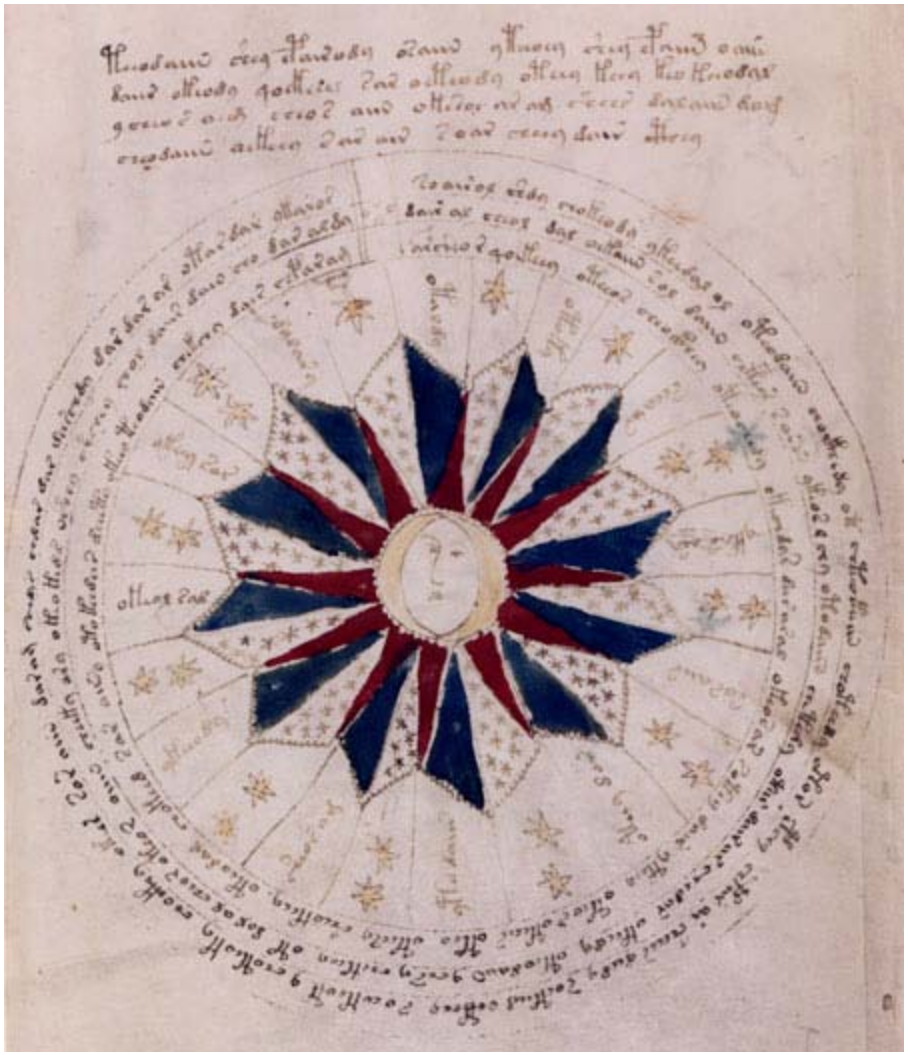
The Pictures: Herbal



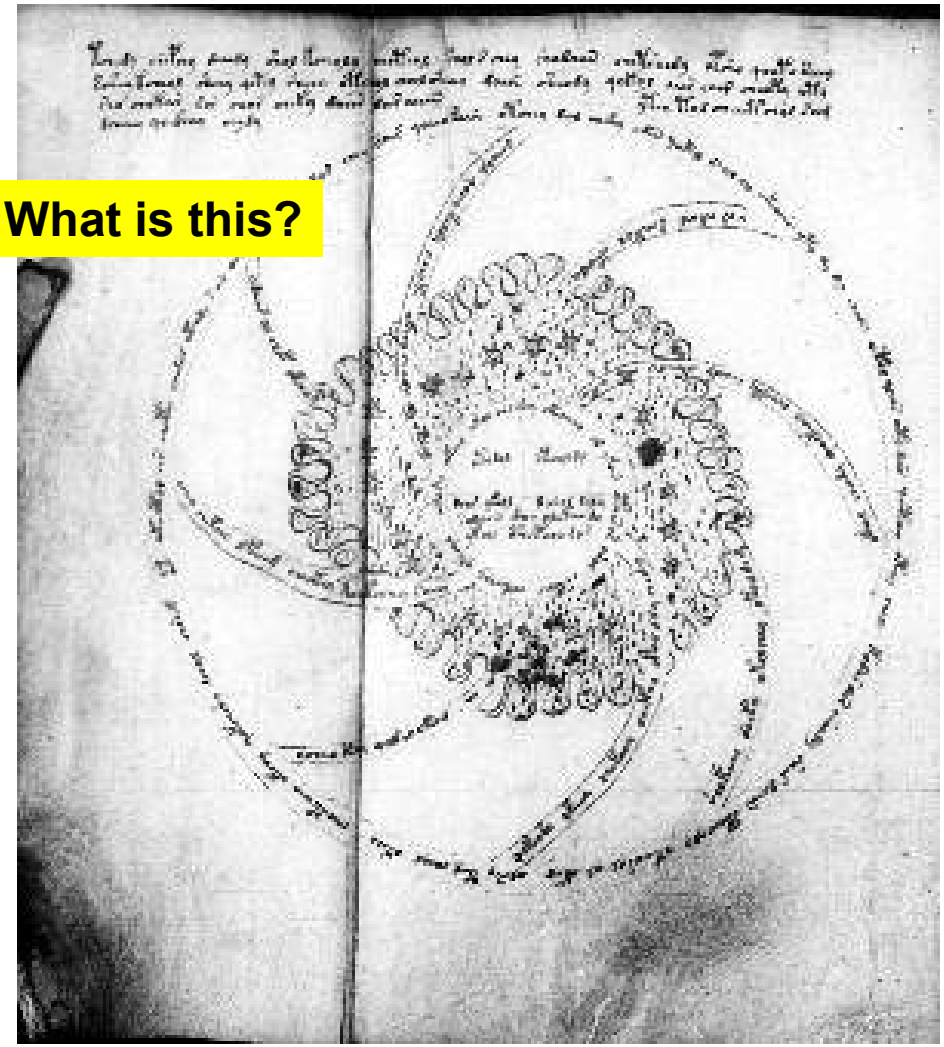
Many pictures look like grafting.

Sunflower? Would date VMS as post-1492.

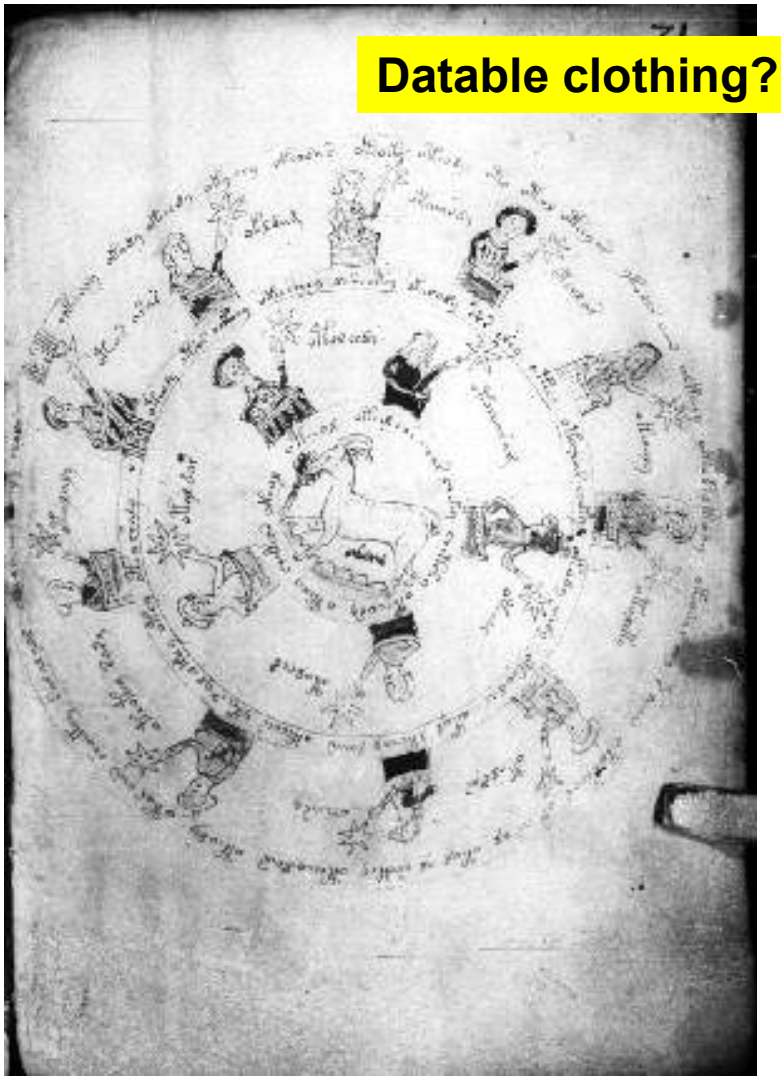
The Pictures: Astrological



The Pictures: Astrological

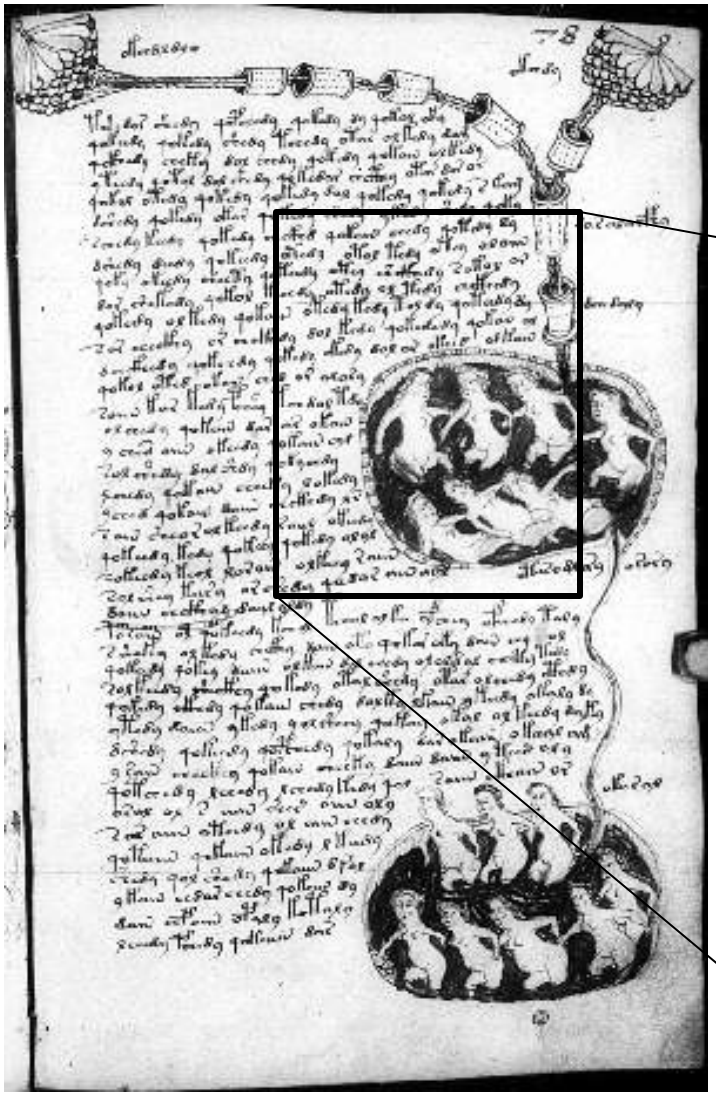


What is this?



Datable clothing?

The Pictures: Biological



Small nudes in baths

Interconnecting tubes of liquids

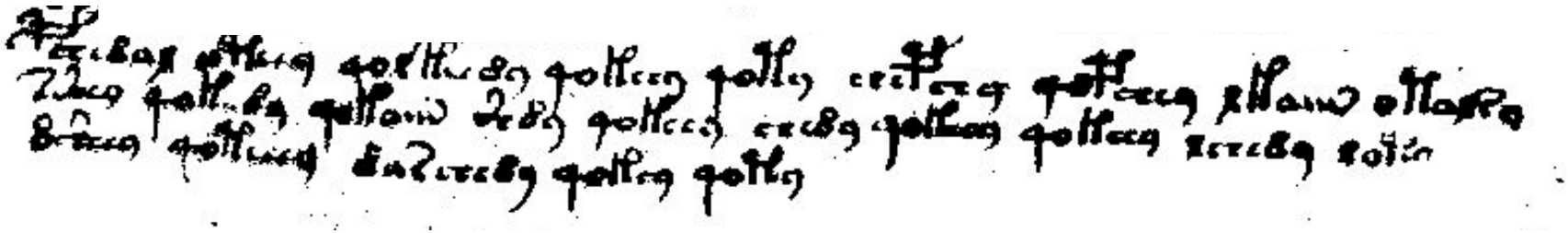
The Pictures: Pharmacological

medicine jar?

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

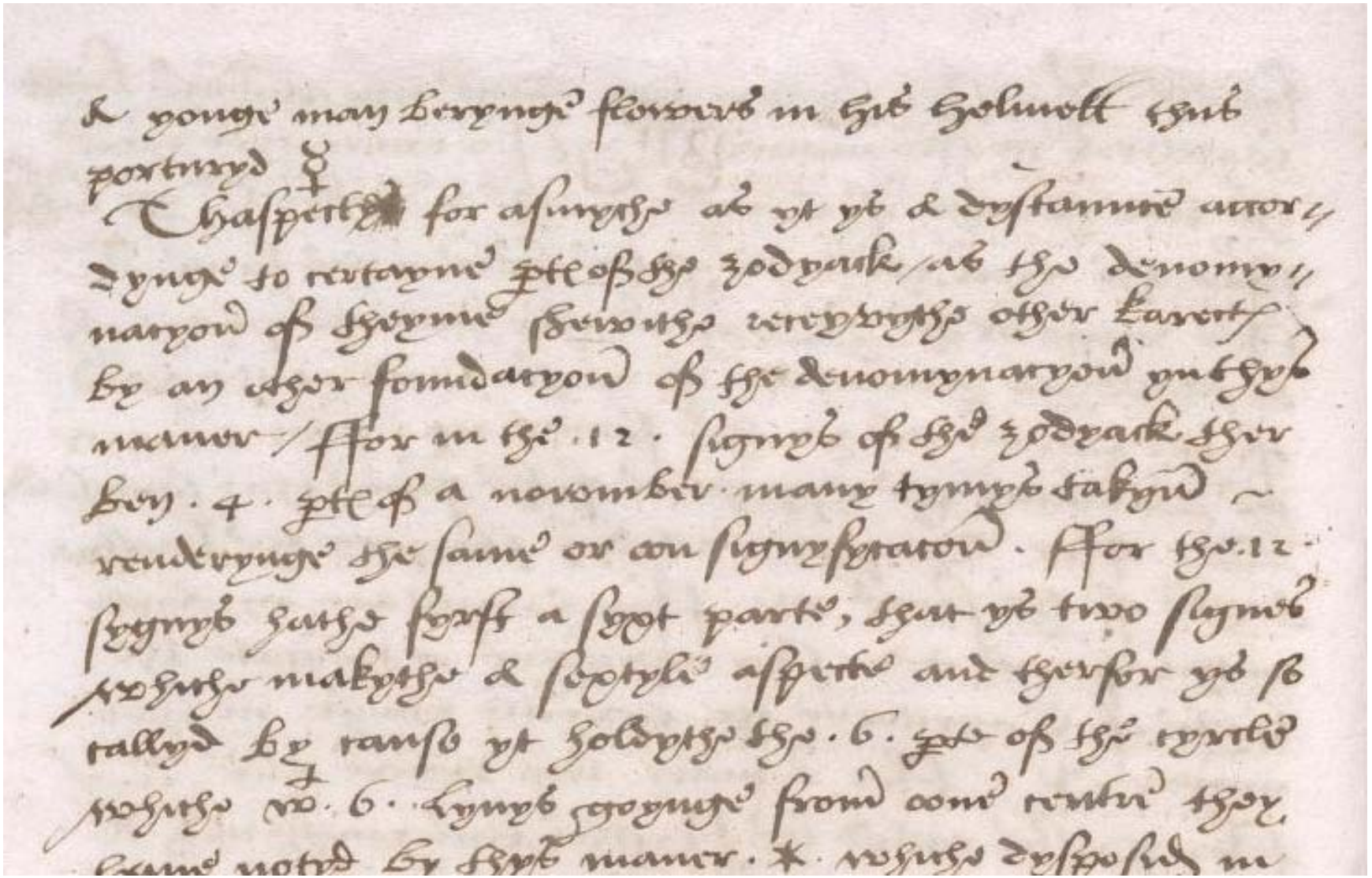
Transcription



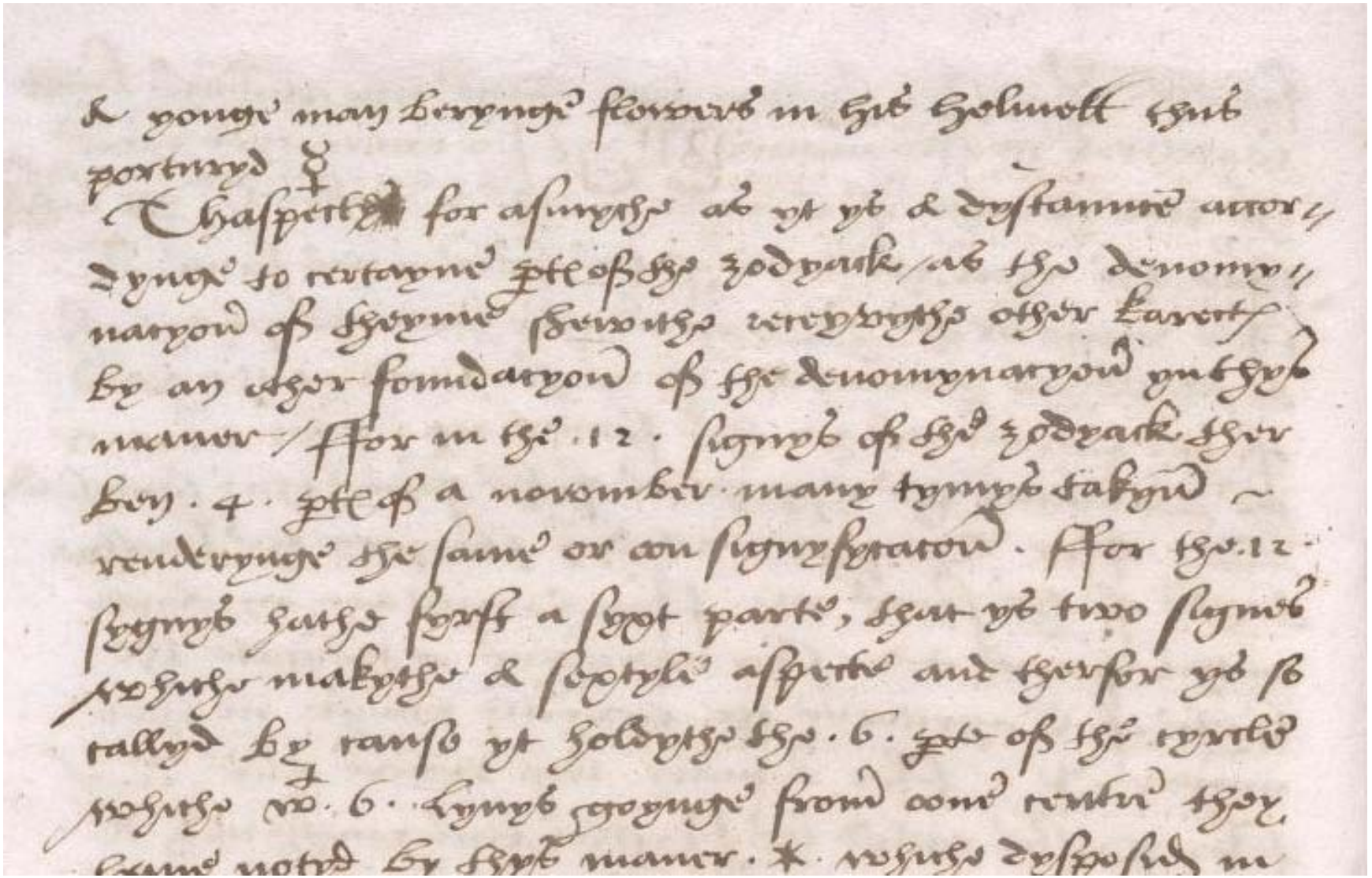
BSC8AE OPCC9 4OE FCC89 4OFCC9 4OP9 SCBS9 4OBSC9 EFAM OPAE29
2ZC9 4OFC89 4OFAM Z89 4OFCC9 SC89 4OFCC9 4OFCC9 ESC89 EOP9
8ZC9 4OPCCC9 8ARSC89 4OFC9 4OP9

last paragraph, f103r

Another medieval manuscript, just for calibration

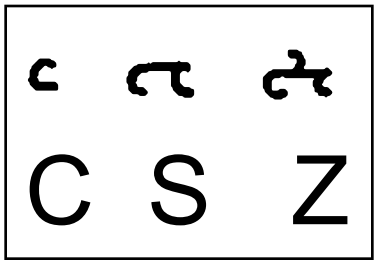


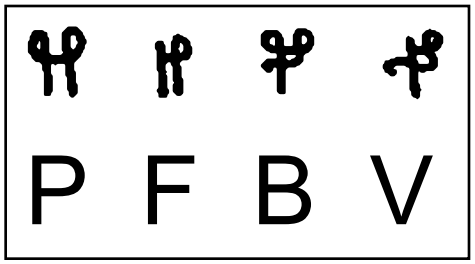
***Introduction to Astrology and Its Use in Weather Prediction, Medicine, and Agriculture*, in English. Manuscript on Paper. 1490.**

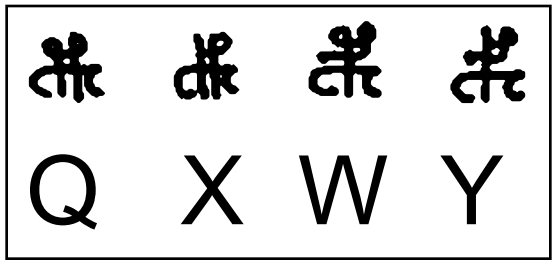


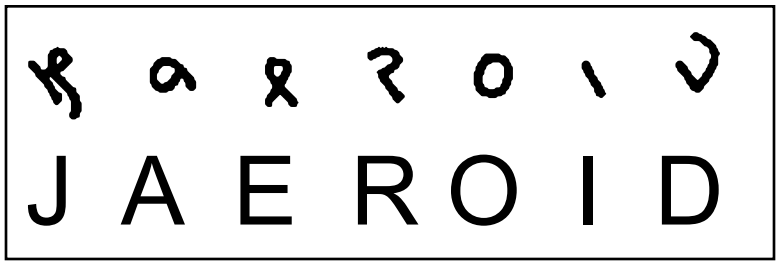
Alphabet: currier/D'Imperio

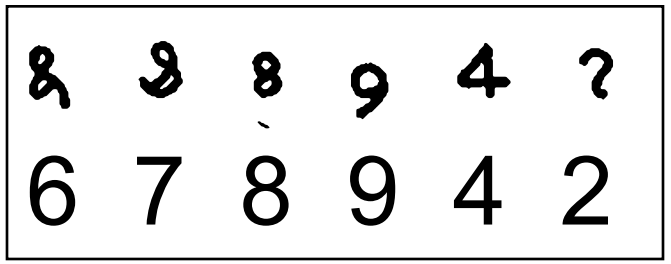
Transcription

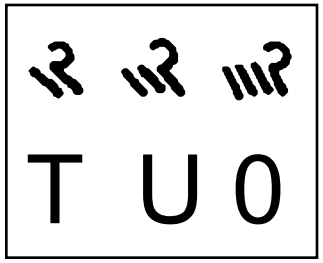


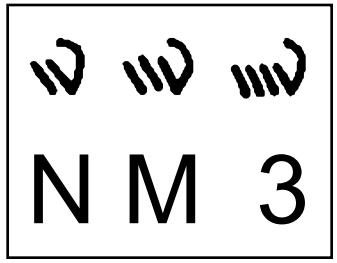




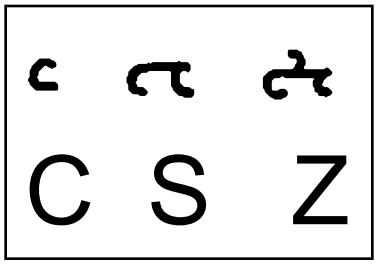


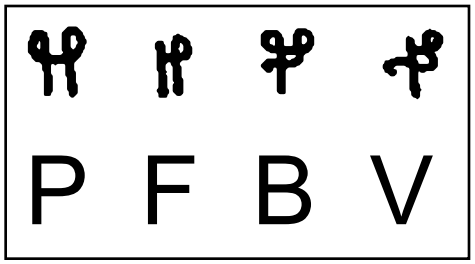


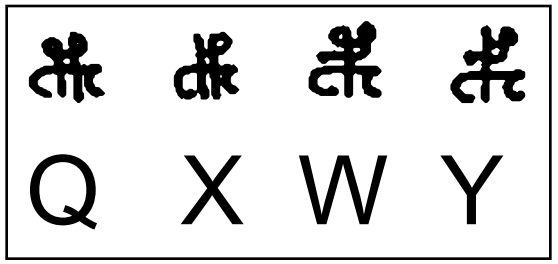


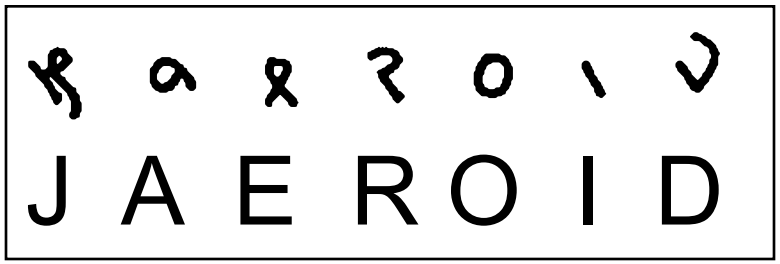


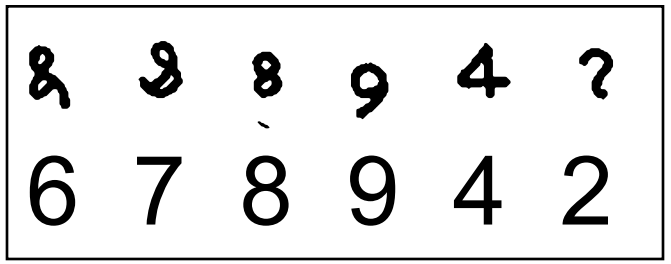
Alphabet: Currier/D'Imperio Transcription

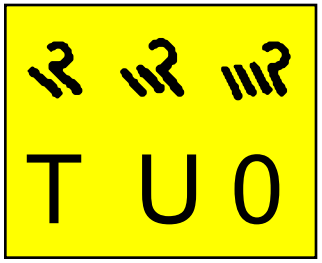










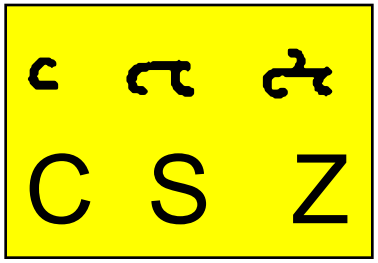


← Maybe this is really

IR IIR IIIR

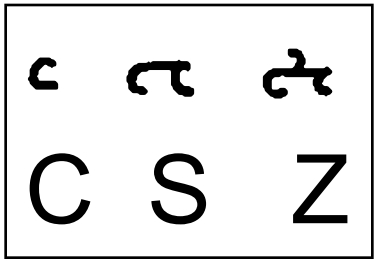
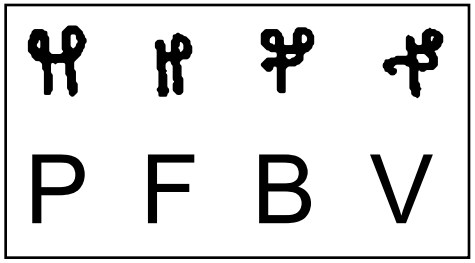
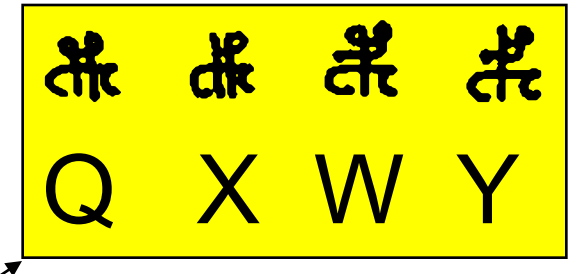
There are several transcription schemes to choose from.

Alphabet: Currier/D'Imperio Transcription



Variations of  , or separate characters?

Alphabet: Currier/D'Imperio Transcription

		
C S Z	P F B V	Q X W Y

Are these ligatures?

Is  just a fancy way of writing  ?

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

Suppose fi **never** occurred. Would that be evidence?
Suppose fi did occur, with the **same** contexts as fi (e.g., *shing)?
Suppose fi did occur, but **never** in the same context as fi ?

Another common motif: 

Letter Frequencies

count	letter	
25468	O	
20227	C	
17655	9	
14281	A	
12973	8	
11008	S	
10471	E	
10026	F	
6716	R	
5994	P	
5423	4	
4501	Z	
4076	M	

count	letter	
2886	2	
1752	N	
1413	B	
1046	J	
950	Q	
908	X	
591	T	
524	*	
431	V	
316	I	
217	W	
157	D	
156	3	

count	letter	
148	U	
96	6	
74	Y	
52	K	
31	G	
17	L	
14	H	
2	1	
1	5	
1	0	

Total
63k character tokens

Most Frequent Words

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

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

count	word	
140	OPCC9	
138	OFAE	
130	ZO	
129	OFAR	
119	ESC89	
118	OFC89	

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

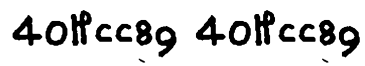
English

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

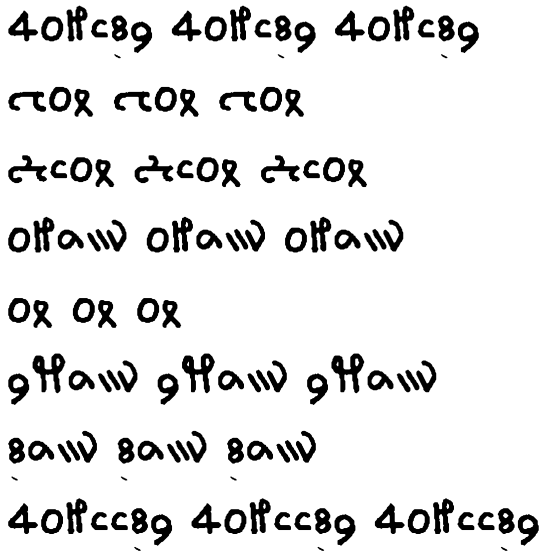
Counts on word types

Features of the Text

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



- 8 words appear tripled



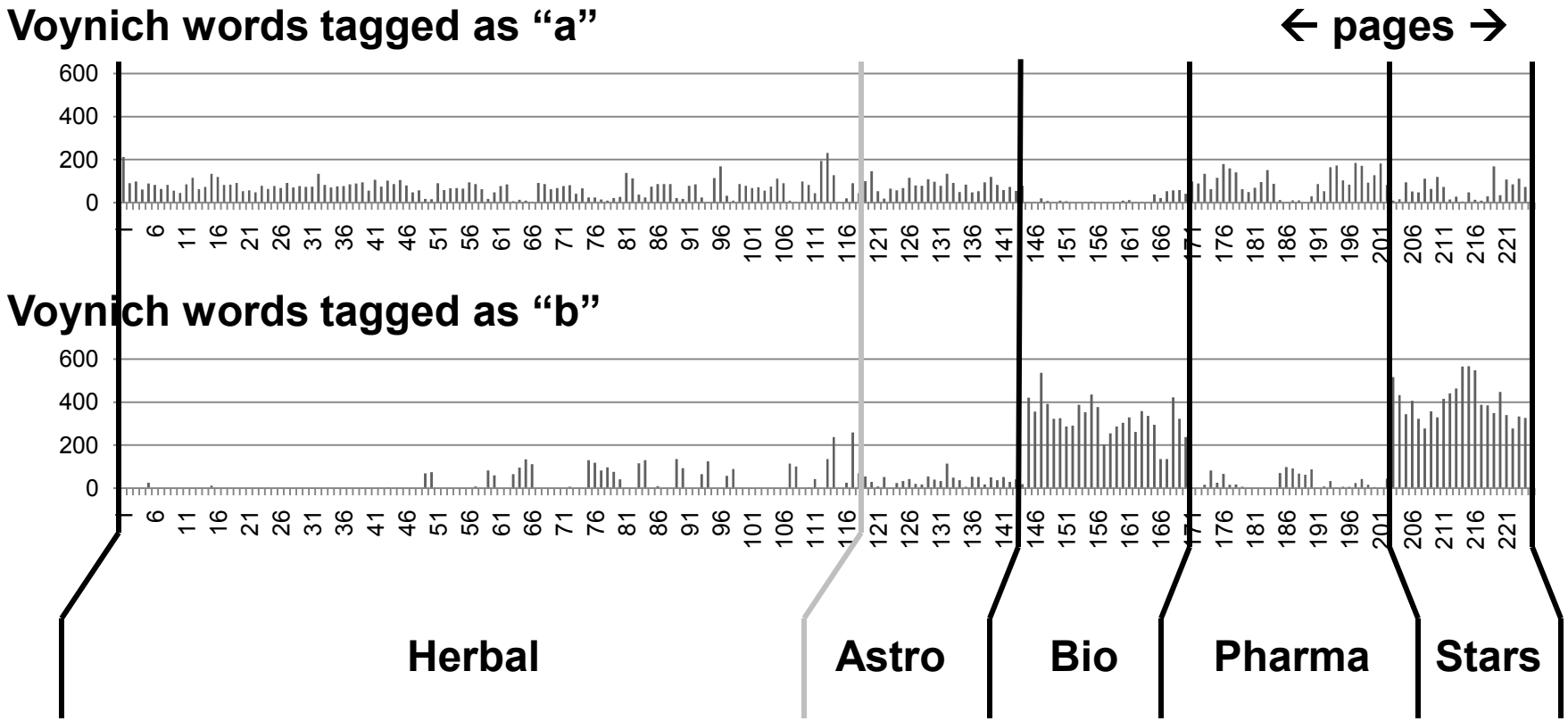
However, very few repeated word bigrams and word trigrams!

No word trigram appears more than 5 times.

Some Theories About the Text

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

Generative models



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

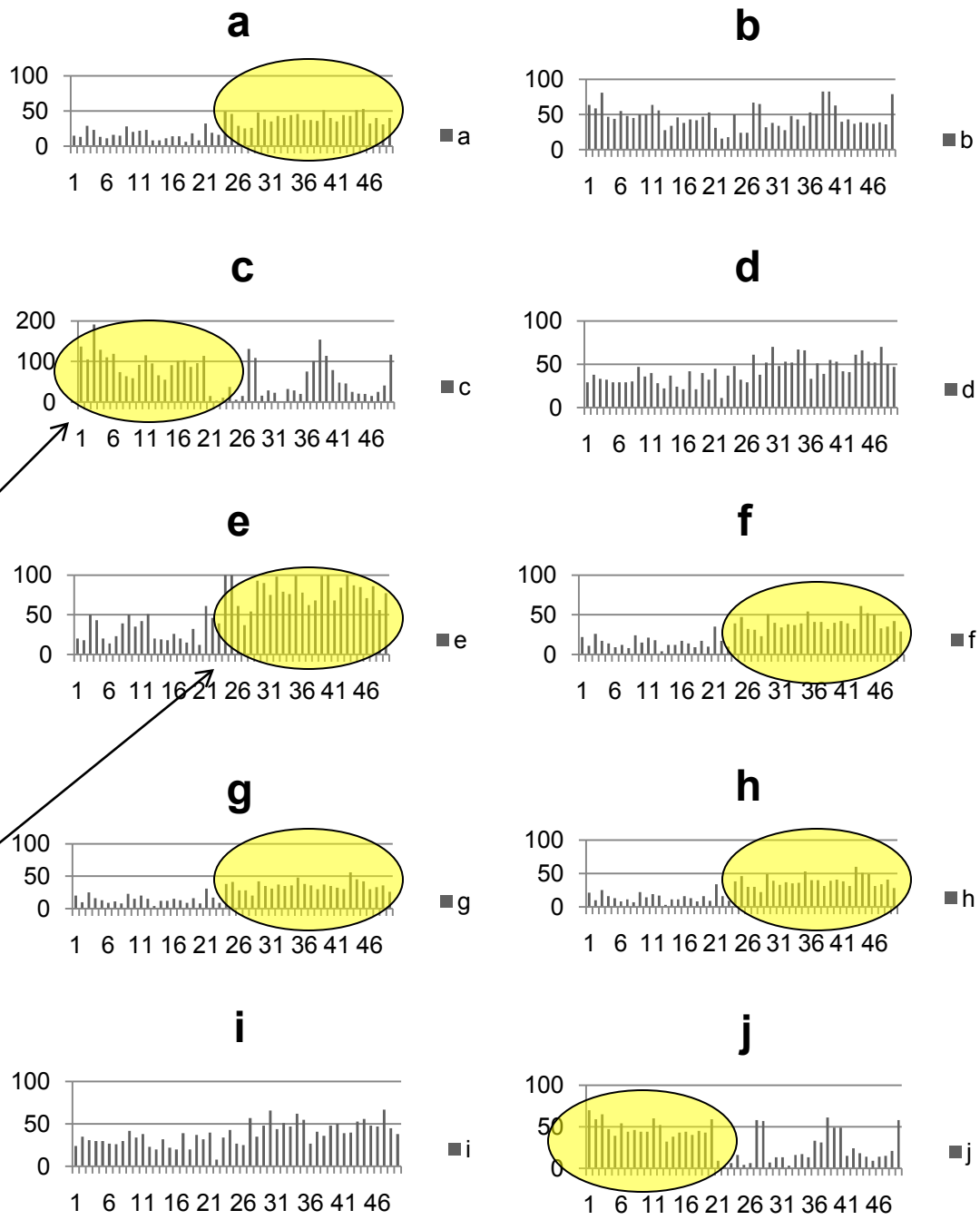
Two “languages” (in the formal sense).

Several handwriting styles, supposedly similar breakdown.

10 classes of words: Voynich-B

Tags per page.

"Bio" words vs. "stars" words



Substitution Cipher

Input	Best decipherment assuming plaintext is Spanish
cevzren cnegr qry 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

quss hates s qum hatis …

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

Undeciphered writing systems

Indus Valley Script (3300BC)

Linear A (1900BC)

Rongorongo (1800s?)

Phaistos Disc (1700BC?)