Putting a Value on Comparable Corpora

Kevin Knight
USC Information Sciences Institute

joint work with

Beáta Megyesi, Christiane Schaefer (Uppsala)
Sujith Ravi (USC)

BUCC 2011

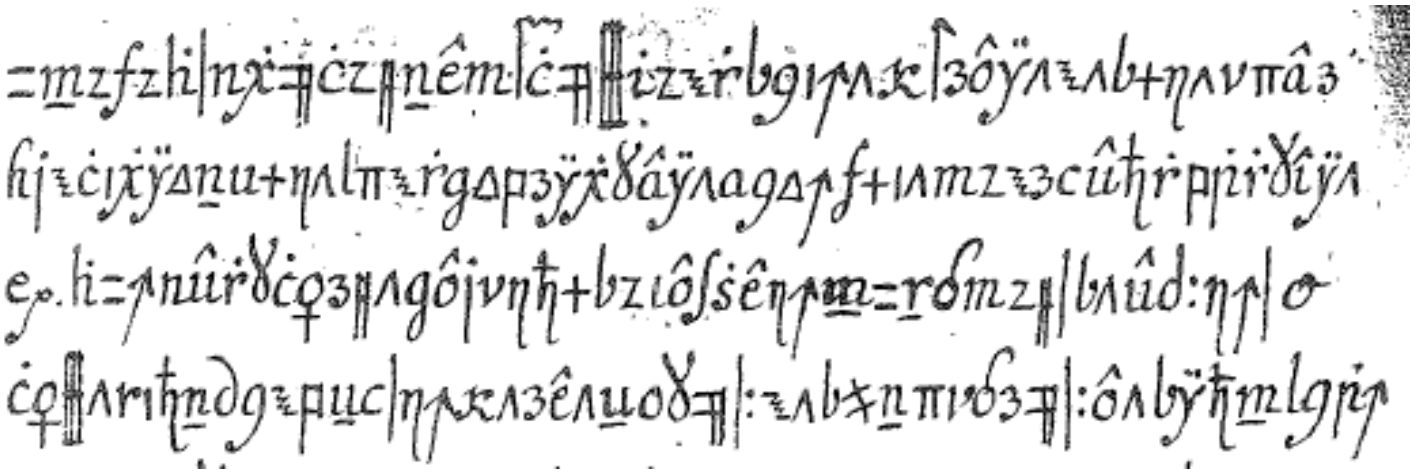
This Talk

Our hero:
Warren Weaver

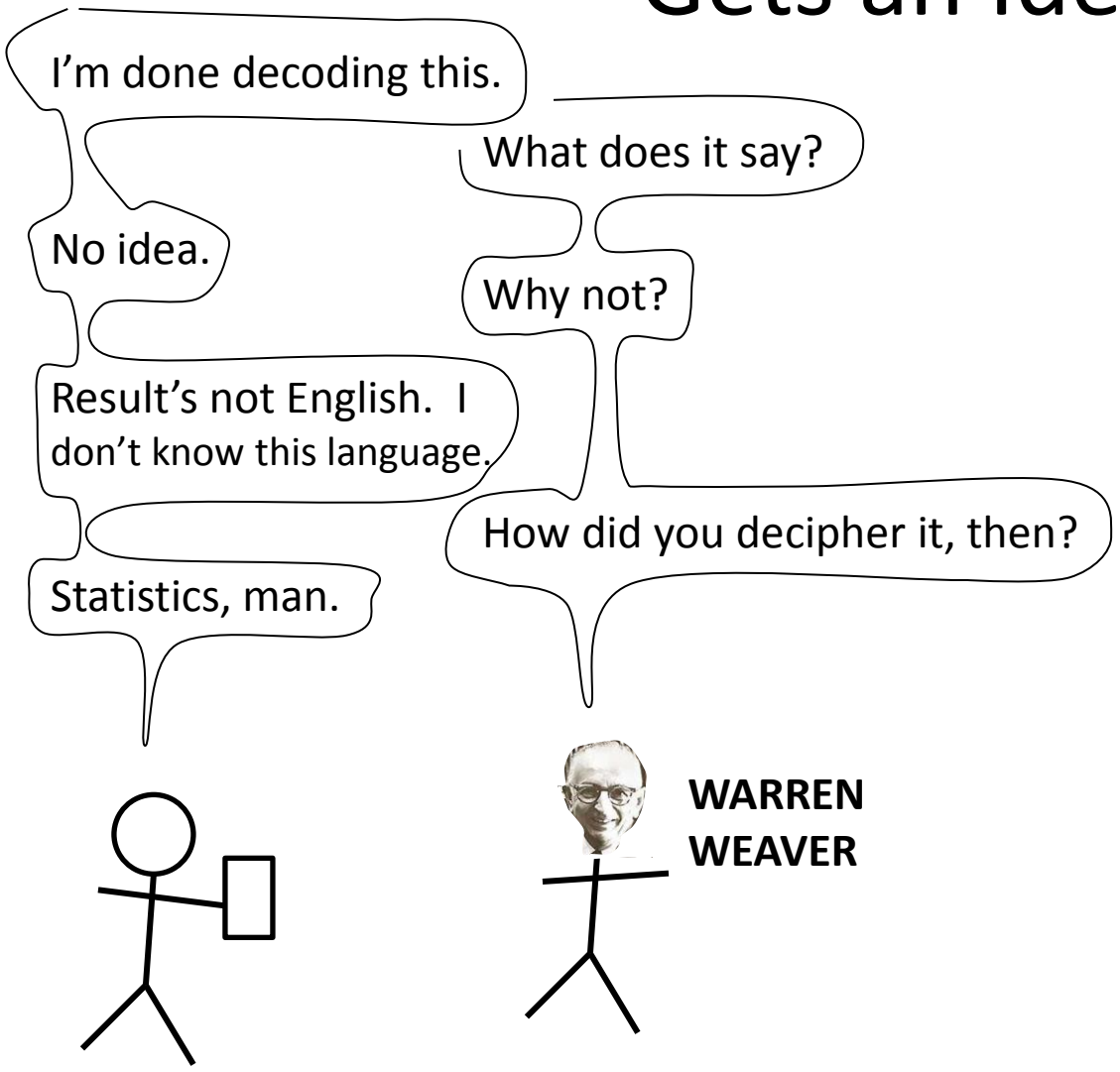
Statistical MT without parallel text

Comparable corpora & code-breaking

Decipher this previously undeciphered manuscript



It's the 1940s, and Warren Weaver Gets an Idea



Weaver's friend only knew *statistics* of Turkish, not Turkish.

Weaver's friend soon replaced by a computer.

Could a computer *translate*, only knowing statistics?

A Note on Terminology

Pre-war computer

Post-war computer

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 . . .

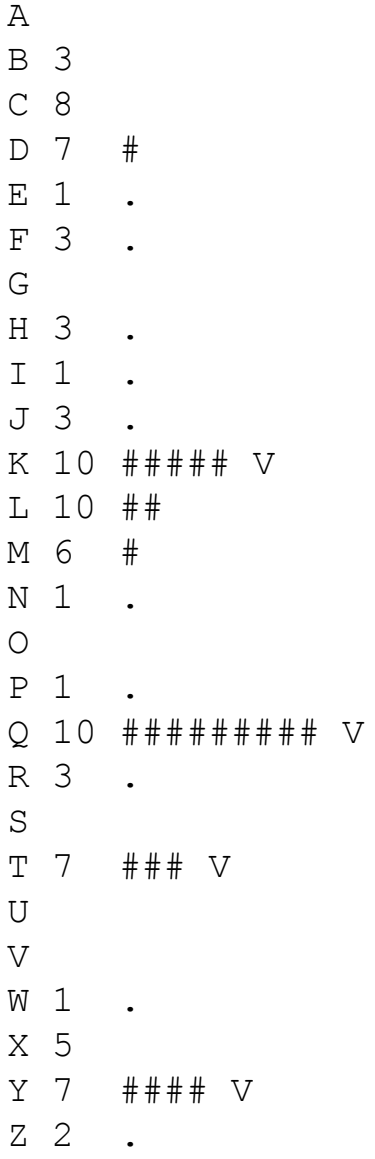
HYD FKXC, FQ MKX RLQQIQ HYDL

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  .
```

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

KDCY LQZKTLJKX CY MDBCYJQL: “TR

.o .a .e a . ee.e .o

HYD FKXC, FQ MKX RLQQIQ HYDL

a . . e .e .a

MKL DXCTW RDCDLQ JQMNKXTMB

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

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 6  #### V
Z 2  .
```

a o re.a r.a o o.e f

KDCY LQZKTLJKX CY MDBCYJQL: “TR

.o .a .e a freeze .o r

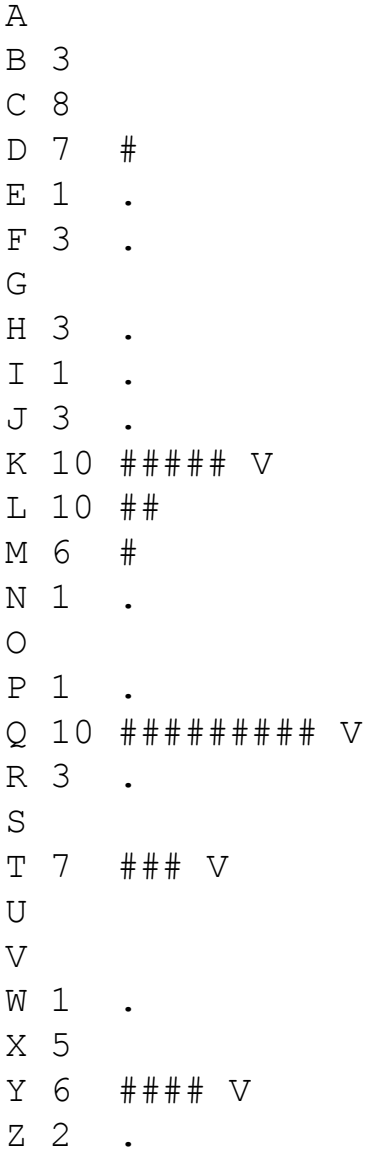
HYD FKXC, FQ MKX RLQQIQ HYDL

ar . f re .e .a

MKL DXCTW RDCDLQ JQMNKXTMB

. o.er a .a. o re.a r

PTBMYEQL K FKH CY LQZKTL TC.”



a o re.a r.a o o.e f

KDCY LQZKTLJKX CY MDBCYJQL: “TR

.o .a .e a freeze .o r

HYD FKXC, FQ MKX RLQQIQ HYDL

ar . f re .e .a

MKL DXCTW RDCDLQ JQMNKXTMB

. o.er a .a. o re.a r

PTBMYEQL K FKH CY LQZKTL TC.”

frequent cipher letters: ~~Q~~ ~~L~~ ~~K~~ C D T M ~~Y~~ X
frequent English letters: ~~e~~ t ~~o~~ ~~a~~ n i ~~r~~ s h

```
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 6  #### V
Z 2  .
```

a no re.air.a no no.e if

KDCY LQZKTLJKX CY MDBCYJQL: “TR

.o .a n .e a freeze .o r

HYD FKXC, FQ MKX RLQQIQ HYDL

ar ni. f n re .e .a i

MKL DXCTW RDCDLQ JQMNKXTMB

.i o.er a .a. no re.air in

PTBMYEQL K FKH CY LQZKTL TC.”

frequent cipher letters: ~~Q~~ ~~L~~ ~~K~~ C D T M ~~Y~~ X
frequent English letters: ~~e~~ t ~~o~~ ~~a~~ n i ~~r~~ s h

```
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 6  #### V
Z 2  .
```

a to re.air.a to to.e if

KDCY LQZKTLJKX CY MDBCYJQL: “TR

.o .a t .e a freeze .o r

HYD FKXC, FQ MKX RLQQIQ HYDL

ar ti. f t re .e .a i

MKL DXCTW RDCDLQ JQMNKXTMB

.i o.er a .a. to re.air it

PTBMYEQL K FKH CY LQZKTL TC.”

frequent cipher letters: ~~Q~~ ~~L~~ ~~K~~ ~~C~~ D ~~T~~ M ~~Y~~ X
frequent English letters: ~~e~~ ~~t~~ ~~o~~ ~~a~~ n ~~i~~ ~~r~~ s h

```
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 6  #### V
Z 2  .
```

a to repair.a to to.e if

KDCY LQZKTLJKX CY MDBCYJQL: “TR

.o .a t .e a freeze .o r

HYD FKXC, FQ MKX RLQQIQ HYDL

ar ti. f t re .e .a i

MKL DXCTW RDCDLQ JQMNKXTMB

.i o.er a .a. to repair it

PTBMYEQL K FKH CY LQZKTL TC.”

frequent cipher letters: ~~Q~~ ~~L~~ ~~K~~ ~~C~~ D ~~T~~ M ~~Y~~ X
frequent English letters: ~~e~~ ~~t~~ ~~o~~ ~~a~~ n ~~i~~ ~~r~~ s h

```
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 6  #### V
Z 2  .
```

auto repairman to customer: if

KDCY LQZKTLJKX CY MDBCYJQL: “TR

you wait we can freeze your

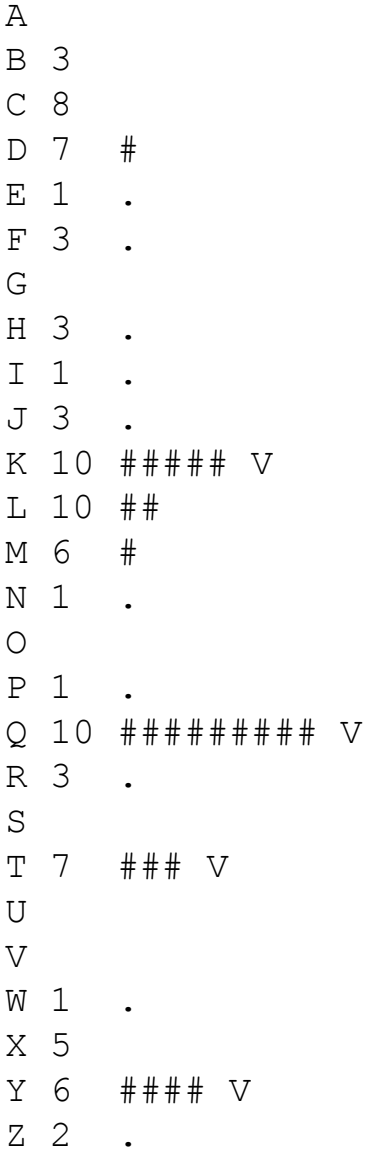
HYD FKXC, FQ MKX RLQQIQ HYDL

car until future mechanics

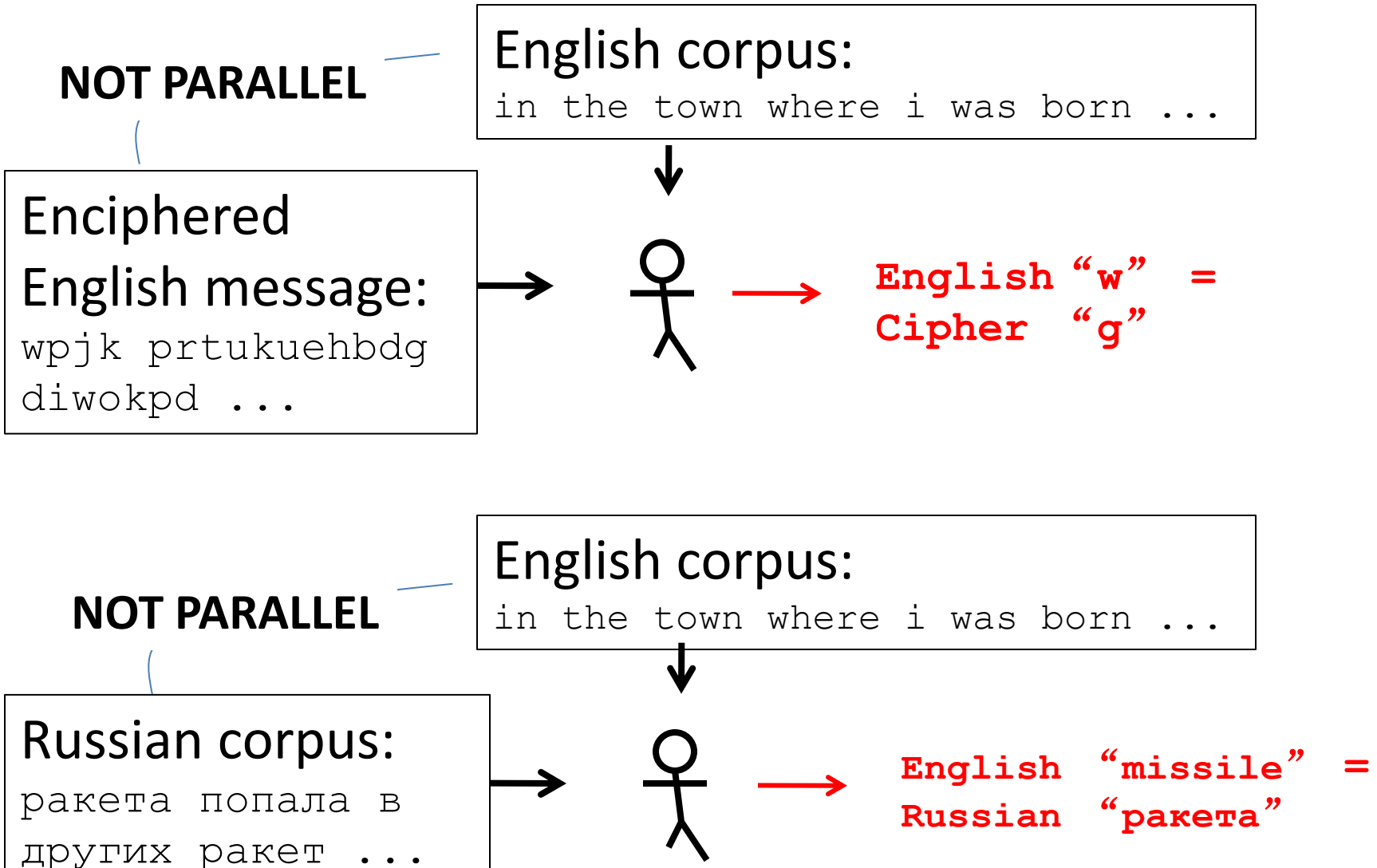
MKL DXCTW RDCDLQ JQMNKXTMB

discover a way to repair it

PTBMYEQL K FKH CY LQZKTL TC.”



Comparable, Non-Parallel Corpora

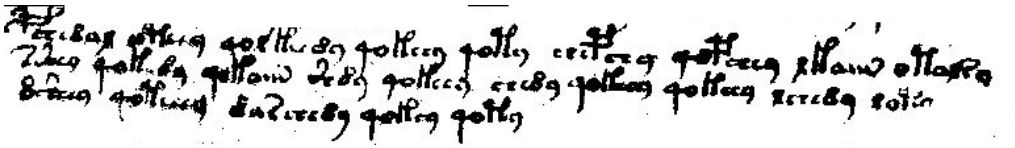


Letter Based Ciphers

- Techniques undoubtedly worked out during 1950s-1960s at NSA and counterparts
  - Still classified!
- Active amateur community (e.g., ACA)
- Related to archaeological decipherment
- Some work in CL community:
  - [Knight & Yamada 1999, Knight et al 2006]
  - [Sproat, 2007] **workshop**
  - [Ravi & Knight 2008, 2009abc, 2011]
  - [Sproat & Knight, 2009] **tutorial**
  - [Corlett & Penn 2010, Snyder et al 2010] **ACL session**
  - [Reddy & Knight, 2011]

Some Unsolved Ciphers

- Voynich Manuscript (1400)

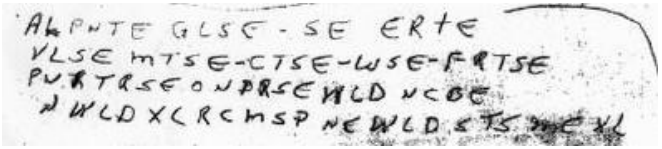


for more, see:
[Reddy & Knight, ACL LaTeCH 2011]

- Zodiac 408 Serial Killer (1967)

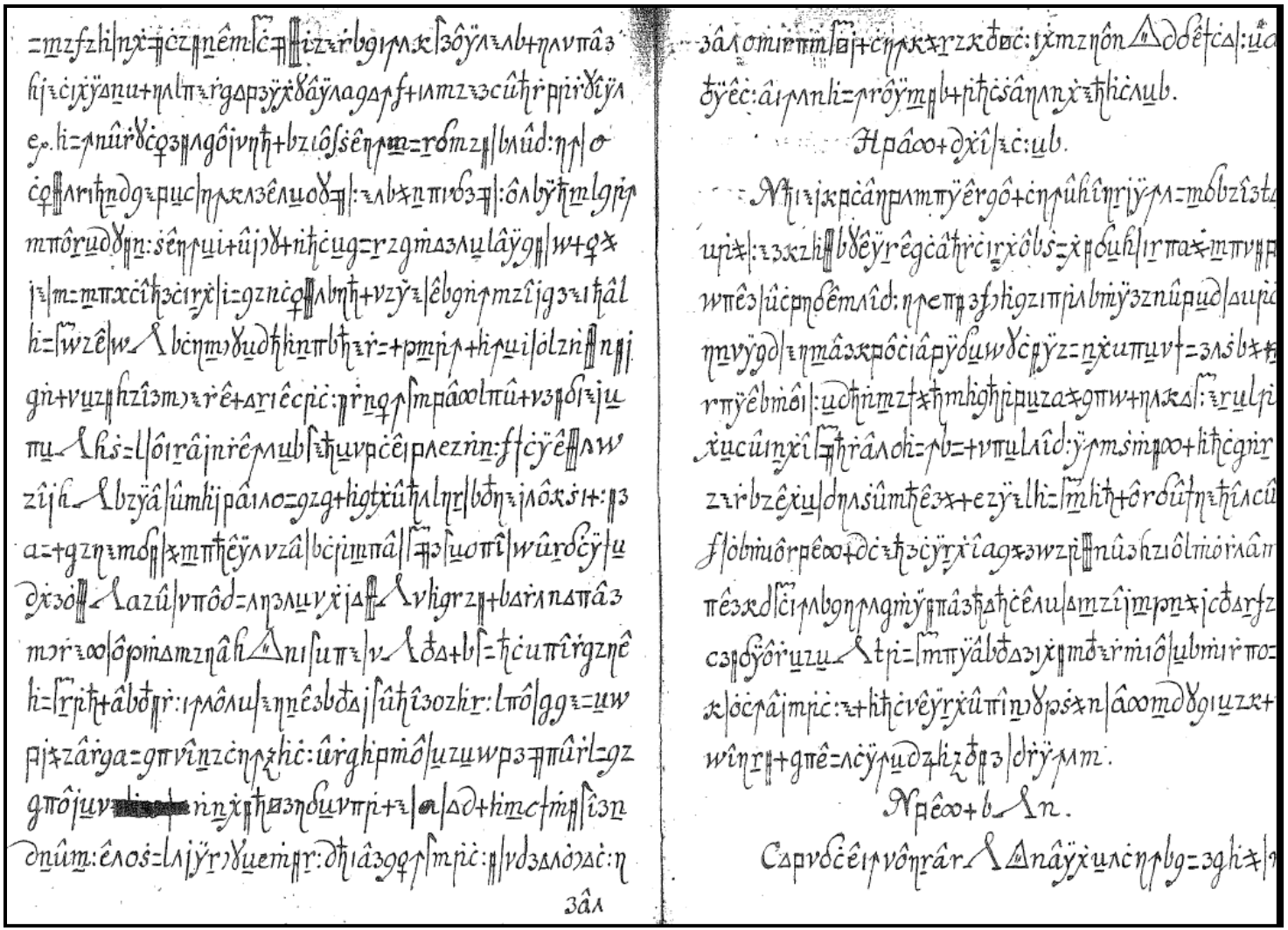
for more, see:
[Ravi & Knight, ACL 2011]

- FBI cipher (1999, released 2011)

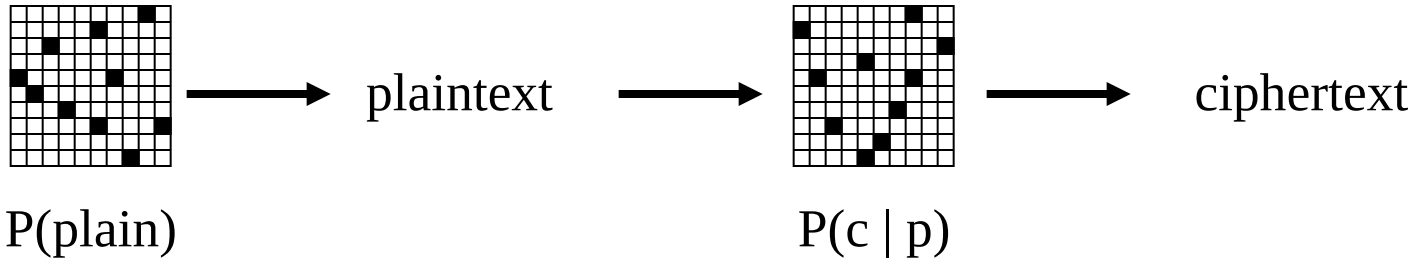


- Kryptos (1990, new clue in 2011)

```
                                                OBKR
UOXOGHULBSOLIFBBWFLRVQQPRNGKSSO
TWTQSJQSSEKZZWATJKLUDIAWINFBNYP
VTTMZFPKWGDKZXTJCDIGKUHUAUEKCAR
```

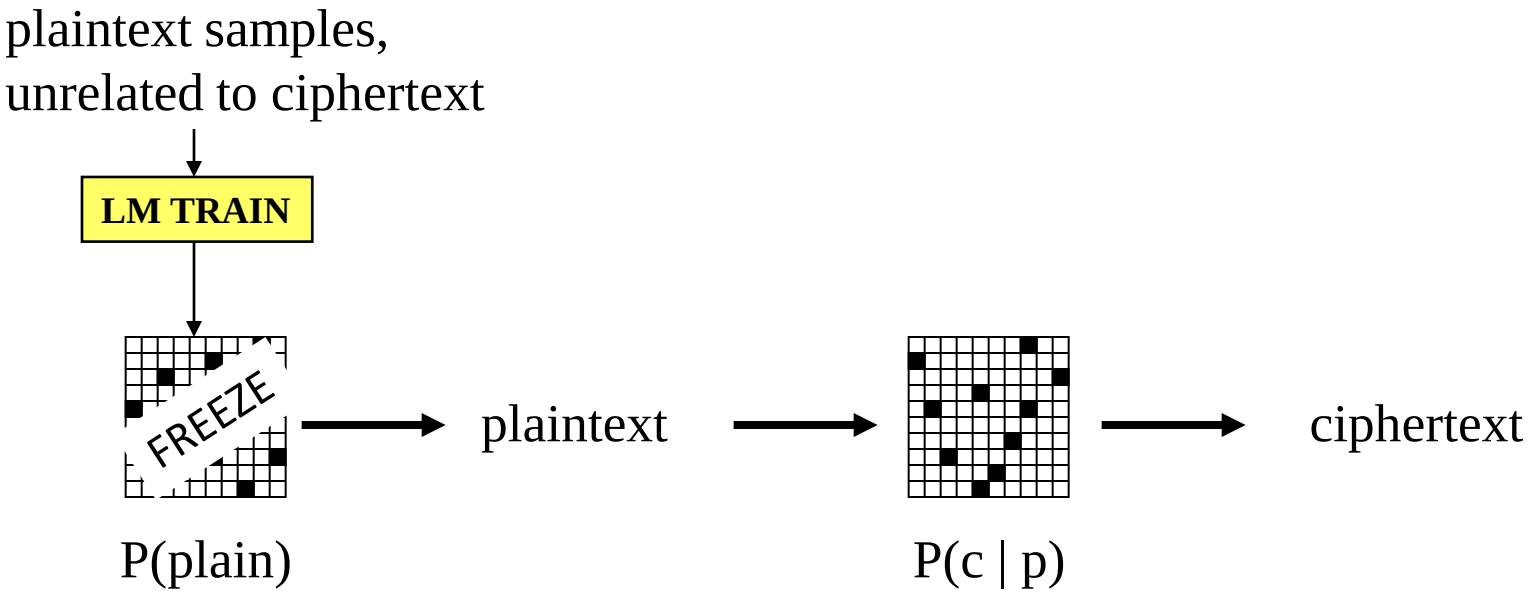


Computer Decipherment



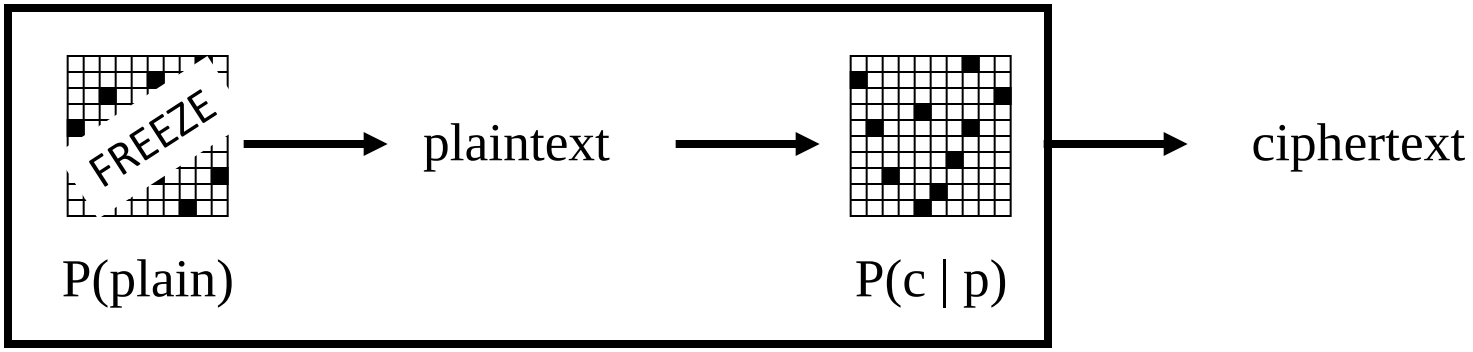
[Knight et al 2006]

Computer Decipherment



[Knight et al 2006]

Computer Decipherment



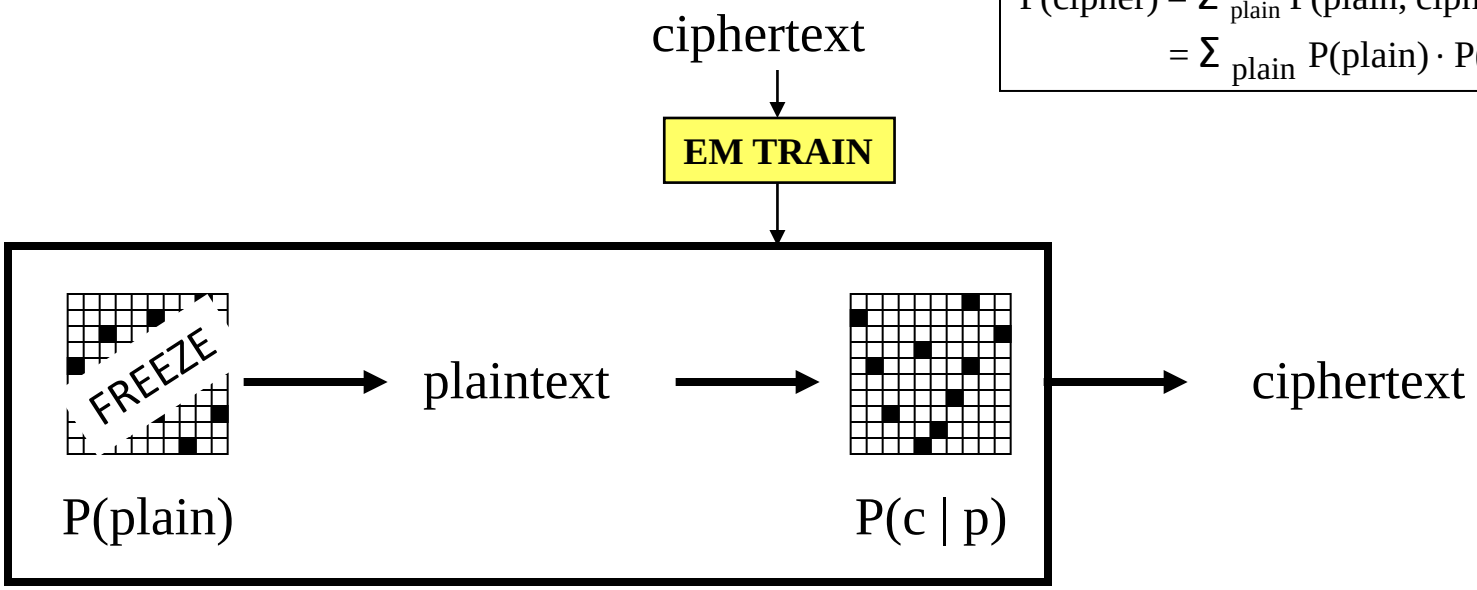
[Knight et al 2006]

Computer Decipherment

Find substitution-table values that maximize

$P(\text{cipher}) = \Sigma_{\text{plain}} P(\text{plain}, \text{cipher})$

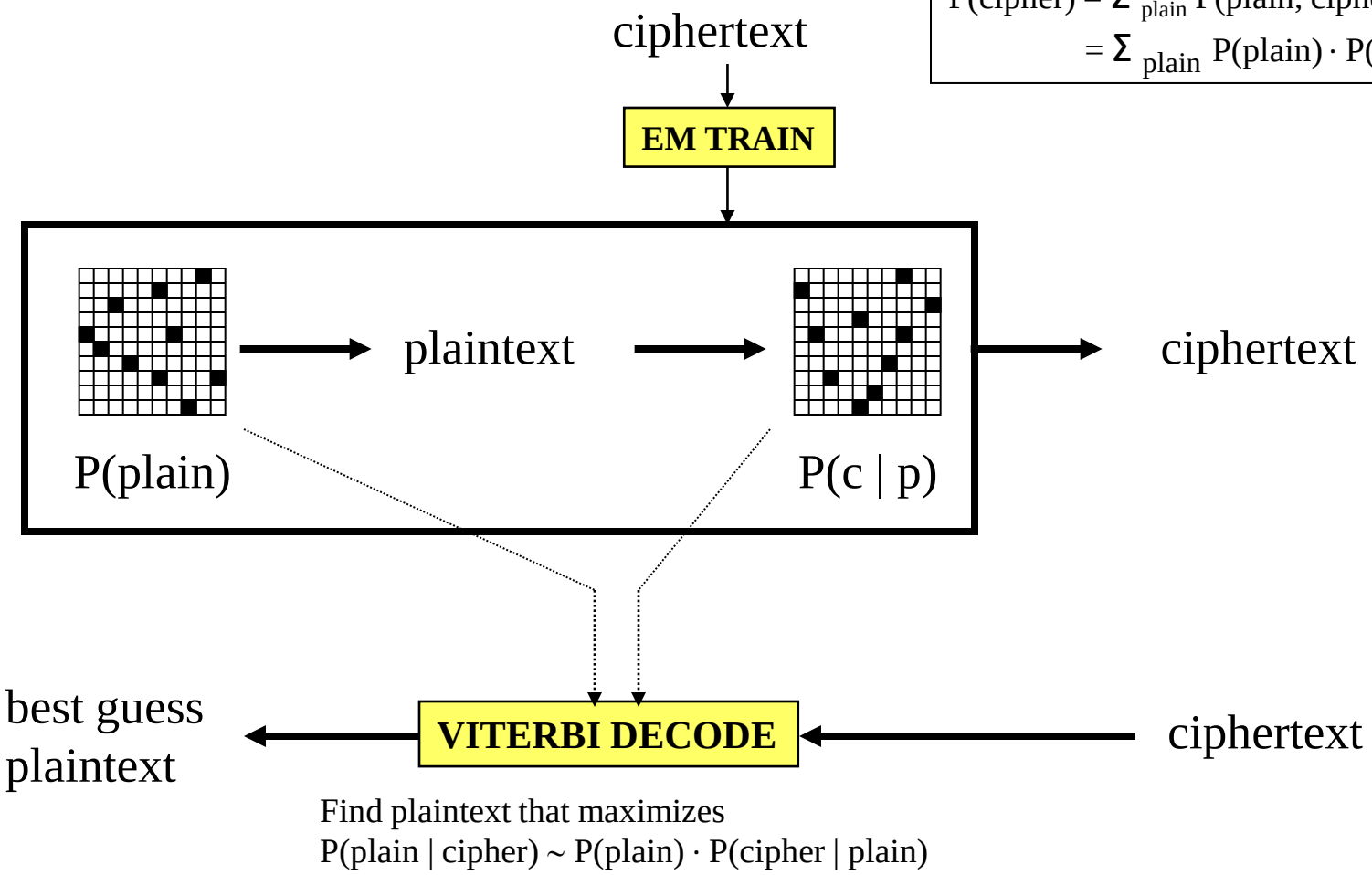
$= \Sigma_{\text{plain}} P(\text{plain}) \cdot P(\text{cipher} \mid \text{plain})$



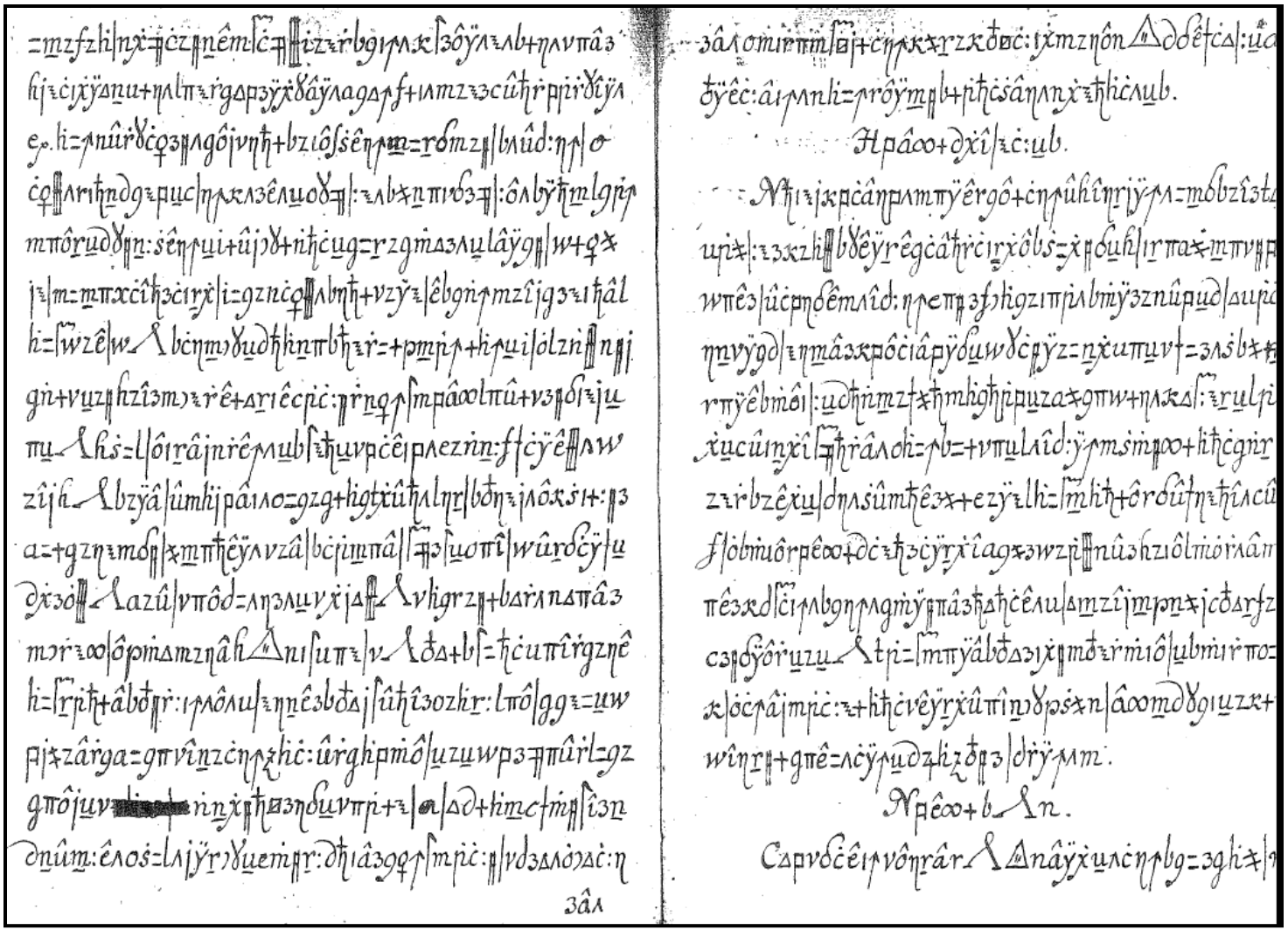
[Knight et al 2006]

Computer Decipherment

Find substitution-table values that maximize
$P(\text{cipher}) = \Sigma_{\text{plain}} P(\text{plain}, \text{cipher})$
$= \Sigma_{\text{plain}} P(\text{plain}) \cdot P(\text{cipher} \mid \text{plain})$



[Knight et al 2006]



Simultaneous ID and Decipherment

- Cipher:

  FZFZ VMJVLH DV RPMOVG VMOZM JZDZULZM KPV DV VMNLAUH NHG OHIH NZLAGH VM NPZGOH DV IANV MP SABH KPV DZ KPAVLV

- What language is this a code for?
- What is the decoding?

Method: For various plaintext languages, run EM to get

$$P(\text{cipher}) = \sum_{\text{plain}} P(\text{plain}, \text{cipher})$$

$$= \sum_{\text{plain}} P(\text{plain}) \cdot P(\text{cipher} \mid \text{plain})$$

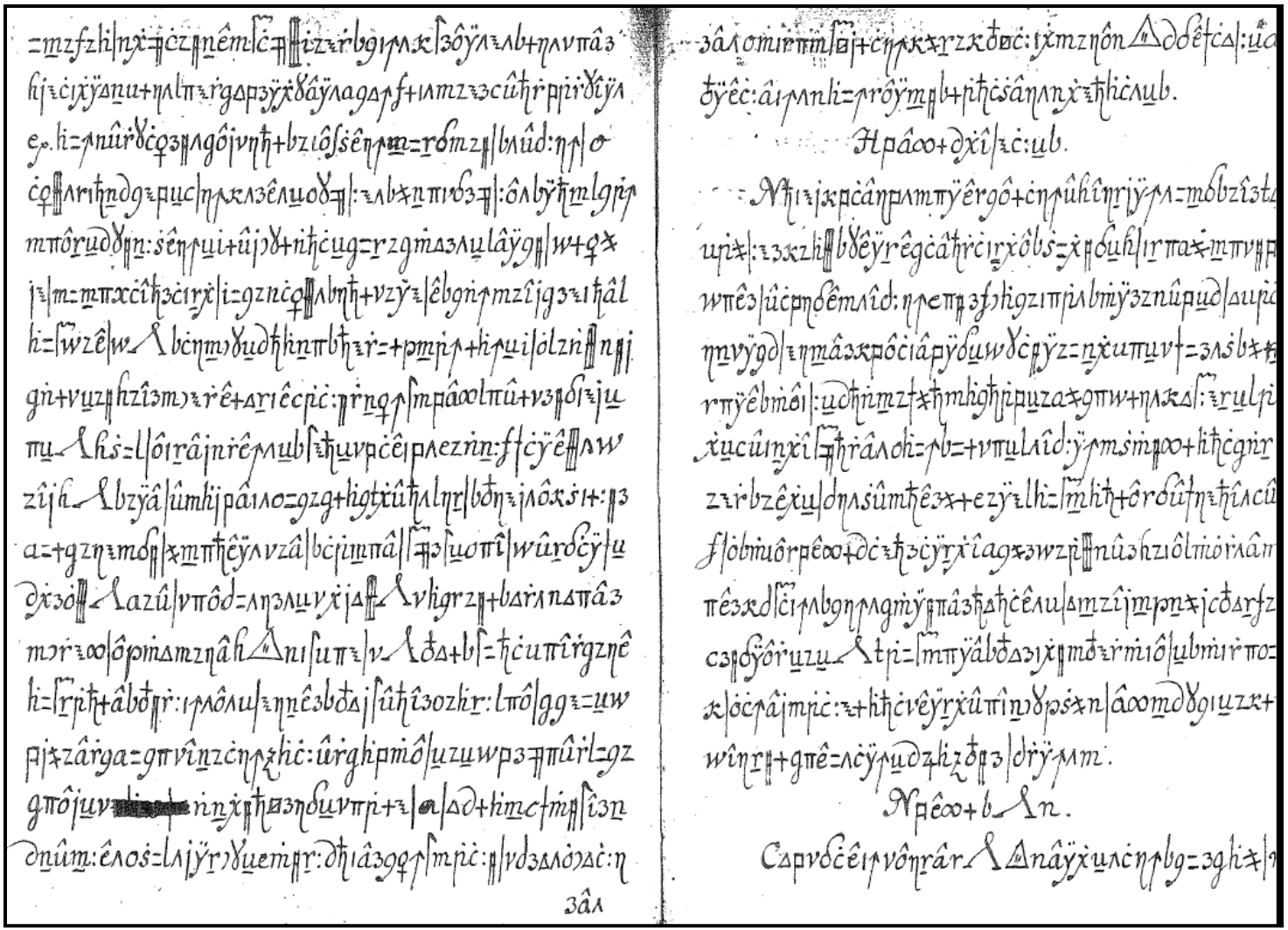
different P(plain) for each plaintext language

Simultaneous ID and Decipherment

cipher: FZFZ VMJVLH DV RPMOVG VMOZM JZDZULZM KPV DV VMNLAUH NHG OHIH NZLAGH VM NPZGOH DV IANV MP SABH KPV DZ KPAVLV

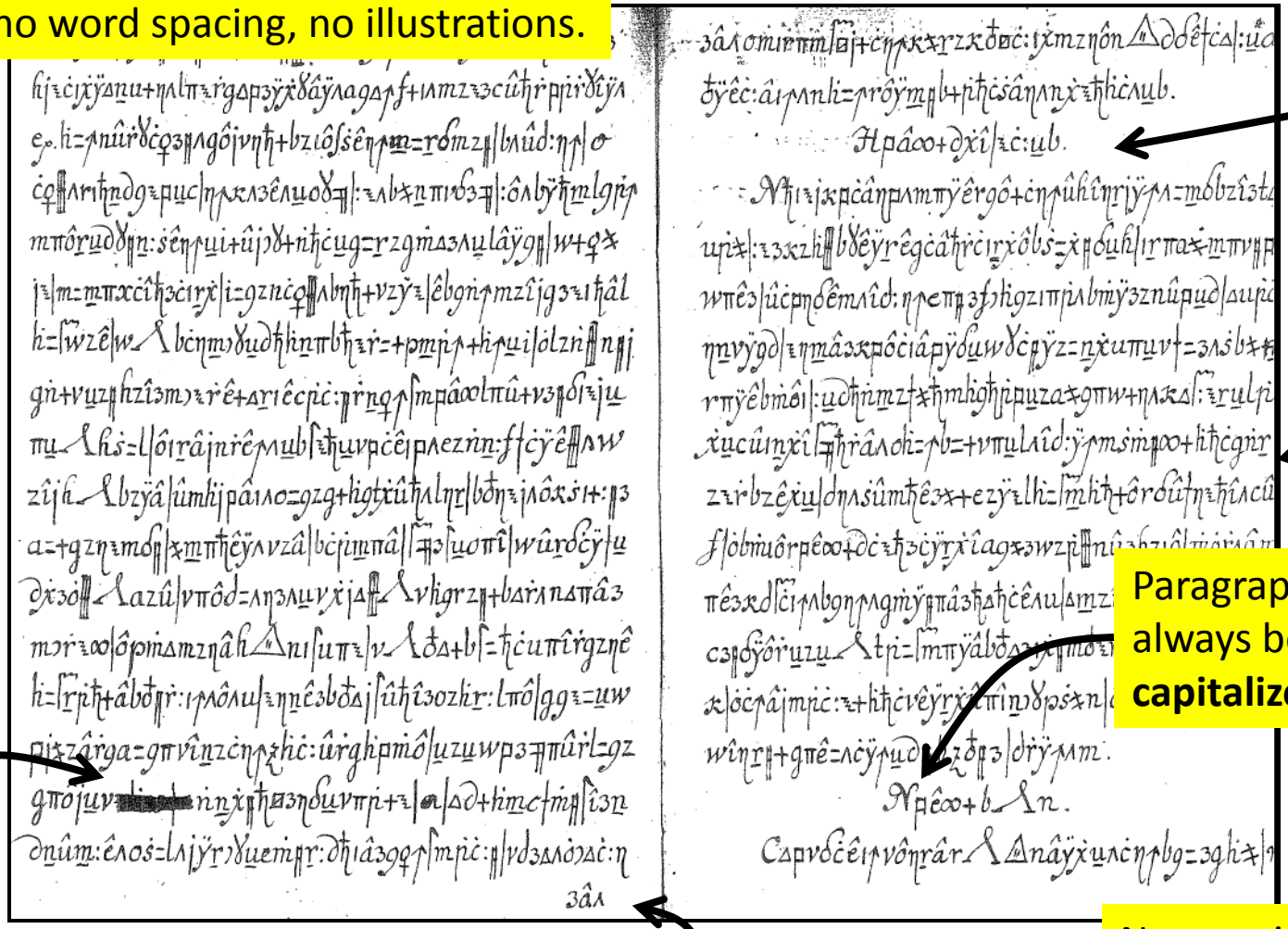
Plaintext Language	P(cipher)	1-best Decoding
English	135.6	oron thicas of behity thind inormand bet of thecums any ings prarys th pervis of pual he full bet or beucat
French	133.5	meme angart la jussan ances gelibres que la aydroit fon cont perint as duenct la vide nu hoit que le quiere
German	130.8	jede arbeit zu vorden urder begewiem von zu urlicht wen stat weicht ur poenst zu alle po scht von zu volein
Italian	131.2	dive allaro ha sultad altel libegrel qua ha esprito con como perito al puento ha fica su divo qua di supare
Portuguese	122.9	dada esfazo de juster estas famancas que de esprino por todo parilo es pualho de live ou fino que da quiaca
Spanish	**117.9**	momo espera de justen estos podotros que de escriza fan tala cogina es cuonta de libe su vida que ho cuiera
Latin	134.2	vivi habera de suatem estis fidignis que de exprogi fam haci ficomi es puilla de cone su coni que di quoere

correct: mama espero le gusten estas palabras que le escribo con todo carino es cuanto le dice su hijo que la quiere



Copiale Cipher

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



Section headers

Lines ≈ equal length

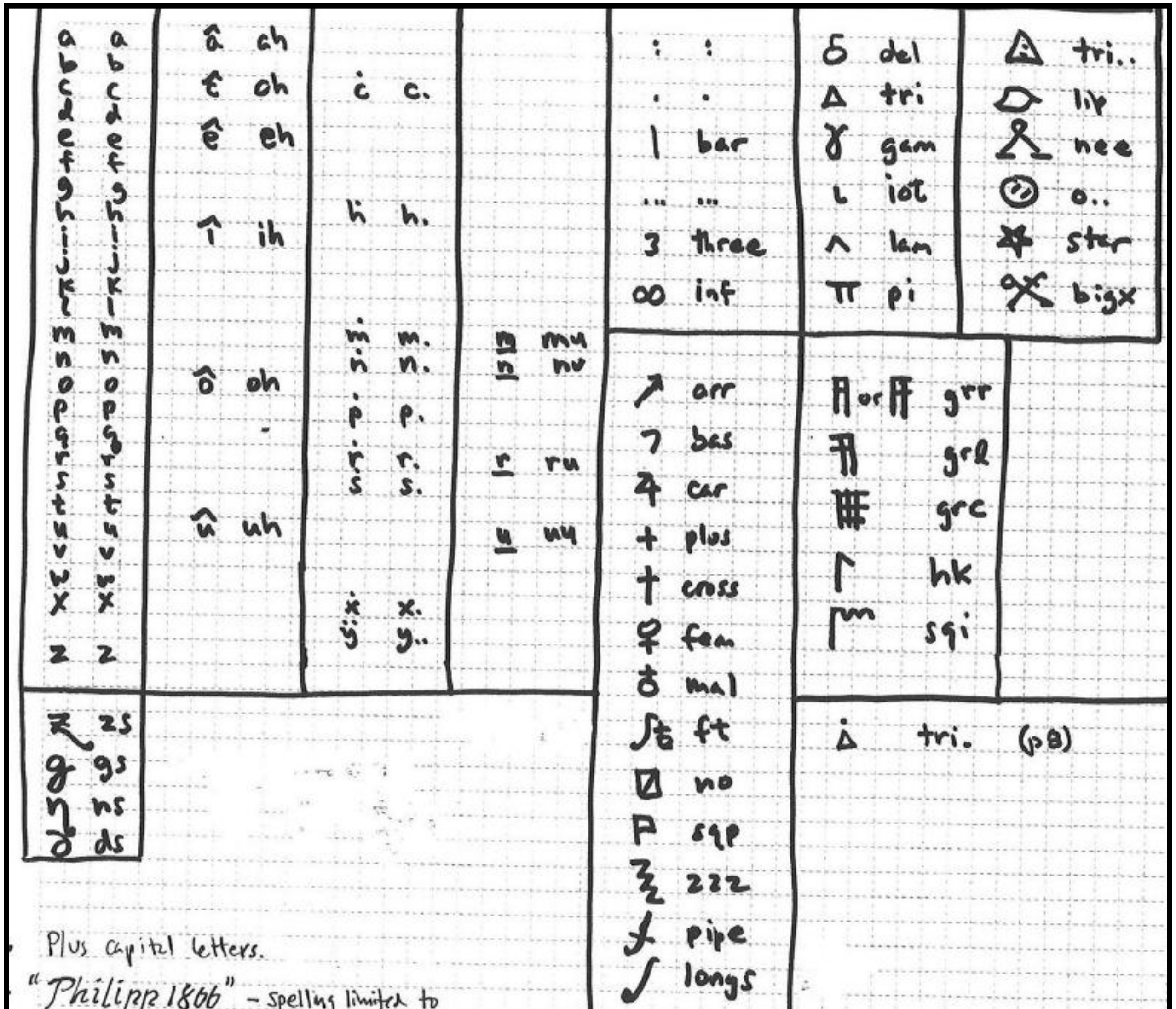
Paragraphs and section titles always begin with **capitalized Roman letters**.

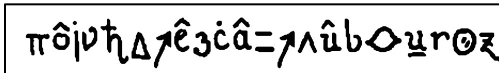
Non-enciphered inscriptions: **Copiales 3** and **Philipp 1866**

Some scratch-outs, rare

Preview text fragments (“catchwords”)

Transcription Scheme





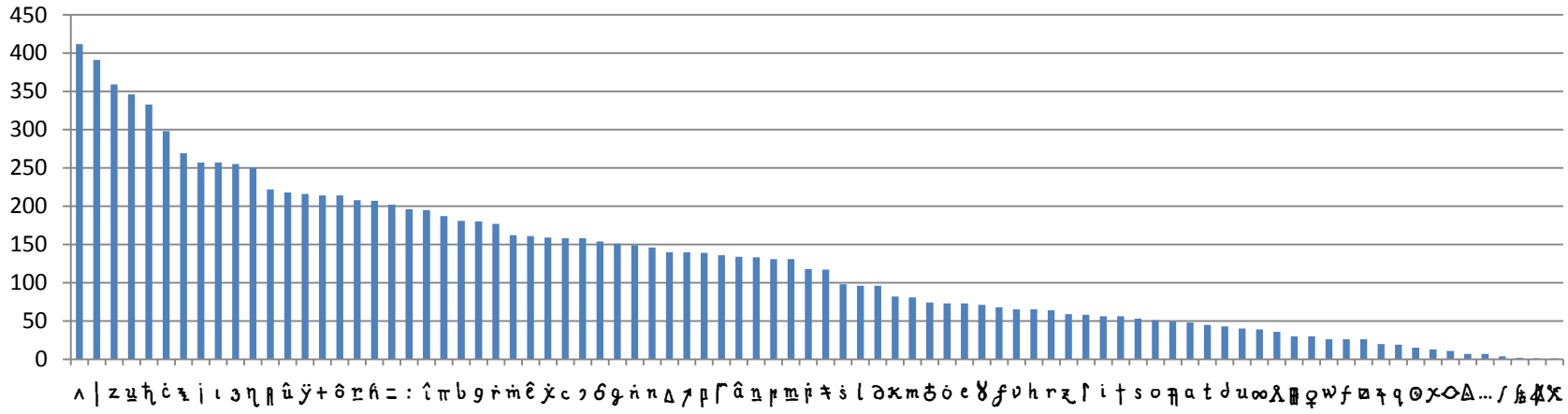
typed as:

pi oh j v hd tri arr eh three c. ah ni arr lam uh b lip uu r o.. zs

~90 cipher letters (plus capitals)

Initially typed 16 pp (10,840 letters).

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 ꞑ	44	ɦ ċ |	17

tendencies:

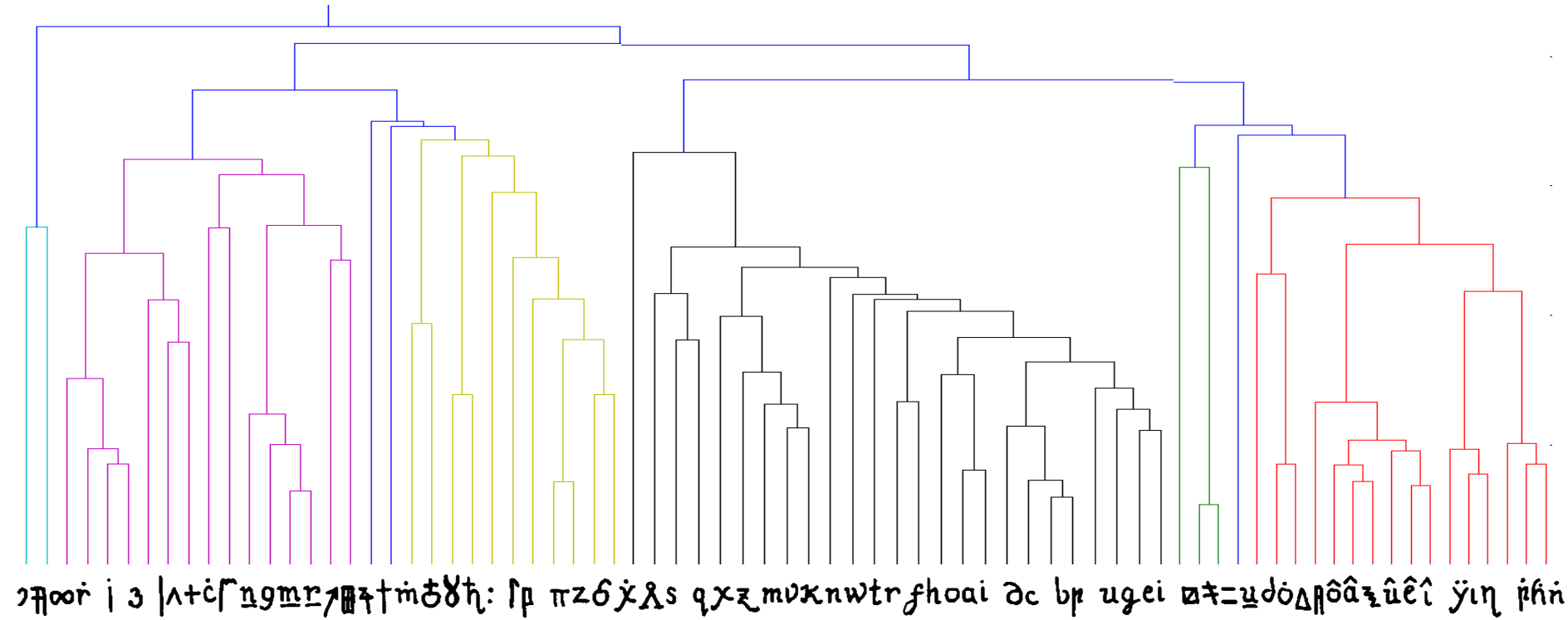
â, ê, î, ô, û followed by ʒ and ¡

â, ê, î, ô, û preceded by z and ɯ

Clustering of Cipher Letters

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

Scipy software

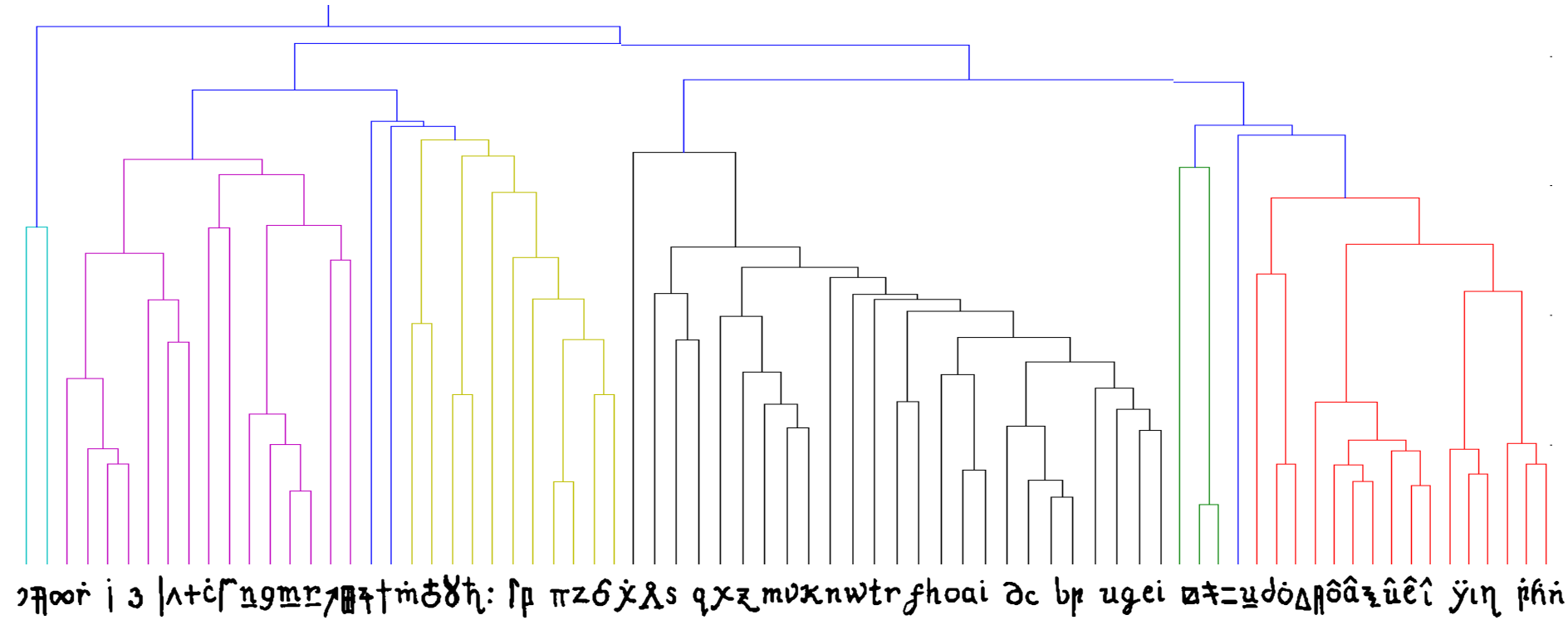


thanks Jon Graehl

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

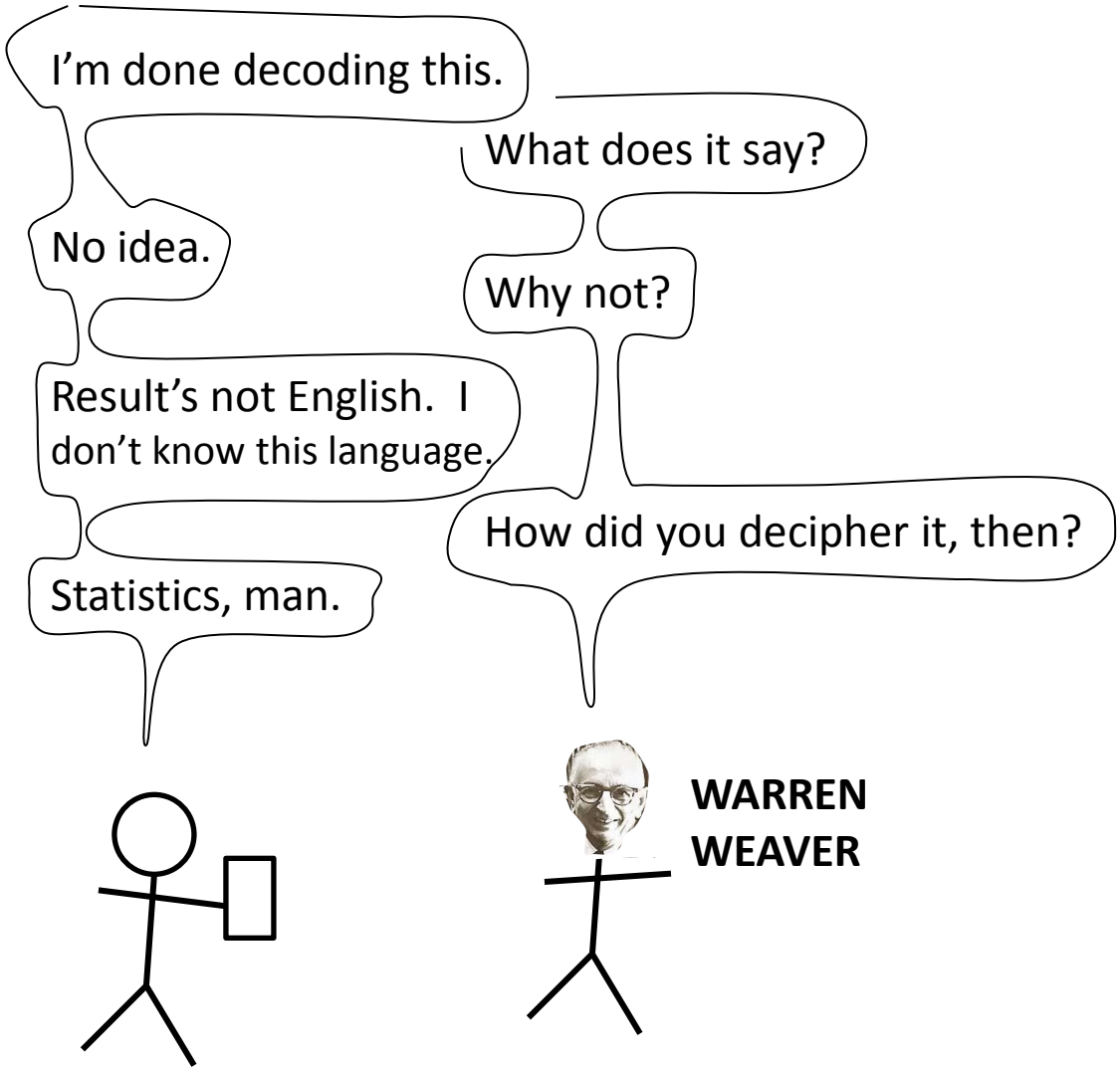
EM clustering (10 classes, a-j)

```
==============
  a  c.     0.17
  a bas     0.16
  a  ru     0.15
  a  mu     0.12
  a   g     0.10
  a  nu     0.06
  a sqi     0.04
  a bar     0.03
  a lam     0.03
  a   j     0.02
  a  oh     0.02
  a plus    0.02
  a   r     0.02
  a  ah     0.01
  a   d     0.01
  a zzz     0.01
==============
  b iot     0.13
  b  ni     0.13
  b  ns     0.13
  b y..     0.11
  b  h.     0.09
  b  n.     0.09
  b  ki     0.08
  b  p.     0.06
  b tri     0.06
  b  o.     0.04
  b  ih     0.02
  b fem     0.01
  b  no     0.01
  b  uh     0.01
==============
  c sqi     0.11
  c lam     0.10
  c sqp     0.10
  c nee     0.08
  c o..     0.06
  c   :     0.05
  c cross   0.05
  c   e     0.05
  c  s.     0.05
  c   i     0.04
  c lip     0.04
  c tri     0.04
  c  x.     0.04
  c ...     0.03
  c tri..   0.03
  c   n     0.02
  c   l     0.01
  c   s     0.01
==============
  d   :     0.20
  d   z     0.17
  d  hd     0.12
  d  pi     0.11
  d bar     0.08
  d lam     0.07
  d del     0.05
  d  x.     0.05
  d gam     0.03
  d  s.     0.02
  d sqi     0.01
  d sqp     0.01
==============
  e  uu     0.50
  e lam     0.14
  e  oh     0.06
  e  eh     0.05
  e zzz     0.05
  e bar     0.04
  e  ah     0.03
  e grr     0.03
  e  uh     0.03
  e  hk     0.02
  e  ih     0.02
==============
  f arr     0.17
  f lam     0.08
  f  ru     0.08
  f   g     0.07
  f del     0.06
  f bar     0.05
  f  mu     0.05
  f  nu     0.05
  f plus    0.05
  f grc     0.04
  f sqi     0.04
  f sqp     0.04
  f  x.     0.04
  f  oh     0.03
  f three   0.03
  f  ah     0.02
  f  r.     0.02
  f zzz     0.02
  f  c.     0.01
  f  hk     0.01
==============
  g zzz     0.16
  g grr     0.14
  g  uh     0.13
  g  ih     0.11
  g  oh     0.10
  g  eh     0.09
  g  ah     0.08
  g tri     0.04
  g grl     0.03
  g  h.     0.03
  g   d     0.02
  g  ns     0.02
==============
  h   j     0.22
  h three   0.18
  h  r.     0.15
  h  hd     0.14
  h bar     0.07
  h lam     0.07
  h inf     0.04
  h plus    0.03
  h y..     0.02
  h  c.     0.01
  h cross   0.01
  h  ns     0.01
==============
  i   b     0.09
  i   c     0.08
  i  gs     0.08
  i   p     0.08
  i   n     0.07
  i  ds     0.05
  i   k     0.05
  i   l     0.05
  i   f     0.04
  i   a     0.03
  i   h     0.03
  i   i     0.03
  i   m     0.03
  i   v     0.03
  i  zs     0.03
  i   e     0.02
  i   o     0.02
  i   r     0.02
  i   s     0.02
  i   t     0.02
  i   u     0.02
  i iot     0.01
  i lam     0.01
  i   q     0.01
  i   w     0.01
==============
  j   z     0.11
  j bar     0.10
  j  m.     0.09
  j  c.     0.07
  j plus    0.07
  j  pi     0.06
  j  hd     0.05
  j mal     0.04
  j  s.     0.04
  j sqp     0.04
  j  x.     0.04
  j   g     0.03
  j del     0.02
  j gam     0.02
  j  h.     0.02
  j lam     0.02
  j sqi     0.02
  j three   0.02
  j bas     0.01
  j   j     0.01
  j  n.     0.01
  j  nu     0.01
  j  p.     0.01
  j  ru     0.01
```

circumflexes

unaccented letters

Plaintext Language is Unknown



Knight et al [2006]:

<u>Method 1:</u>
"Universal" laws of human language

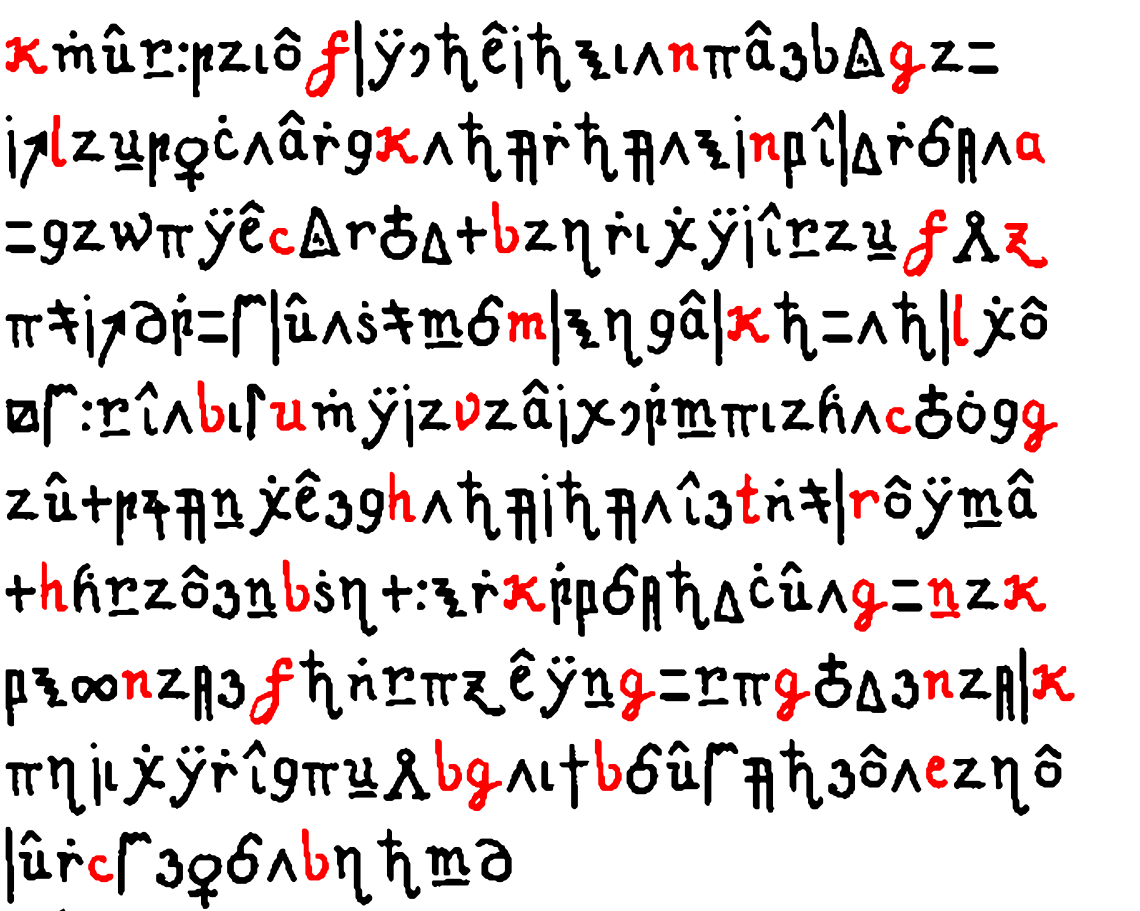
<u>Method 2:</u>
Brute-force decipher, versus all known languages

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 ...

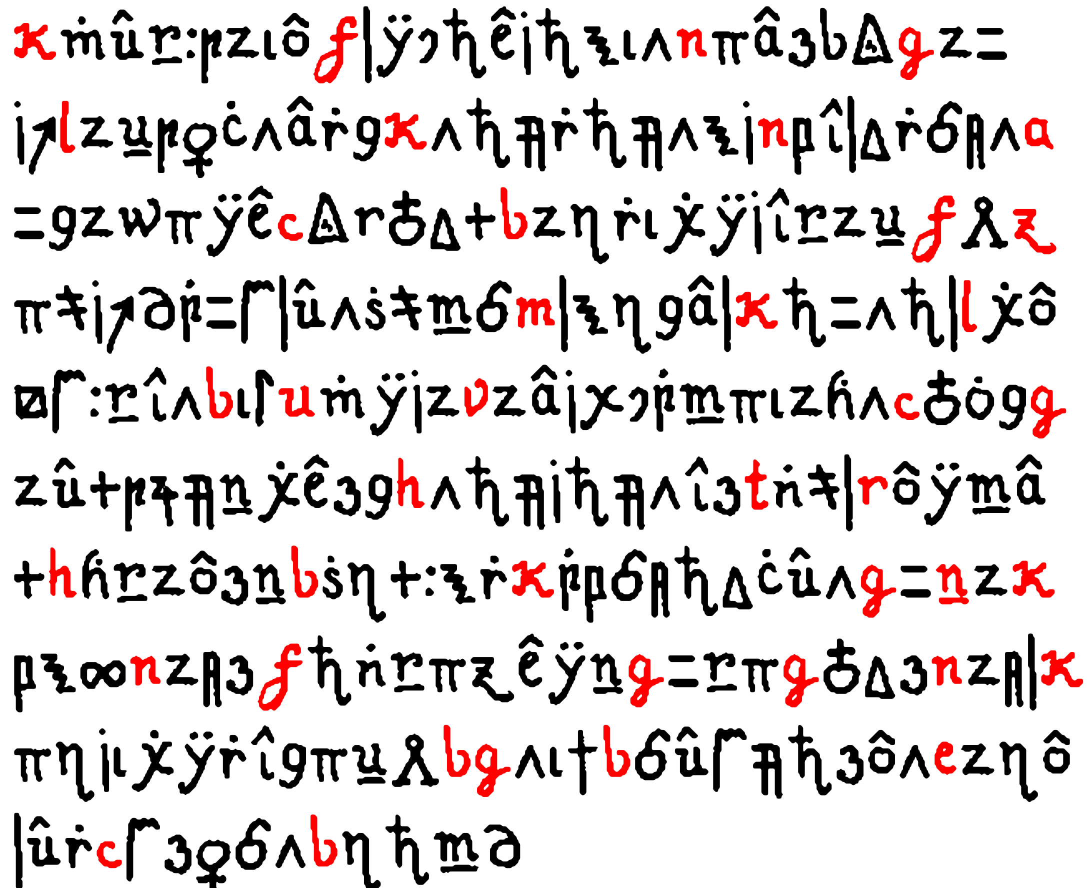
Decipher against 80 plaintext languages.

First Decipherment Approach

unaccented Roman letters that cluster:

a b c ∂ 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



kfnglknacbfzmk
lbuvcghtrhbkgnkn
fggnkbgbecb ...

D... ...nst
80 ... ... languages.

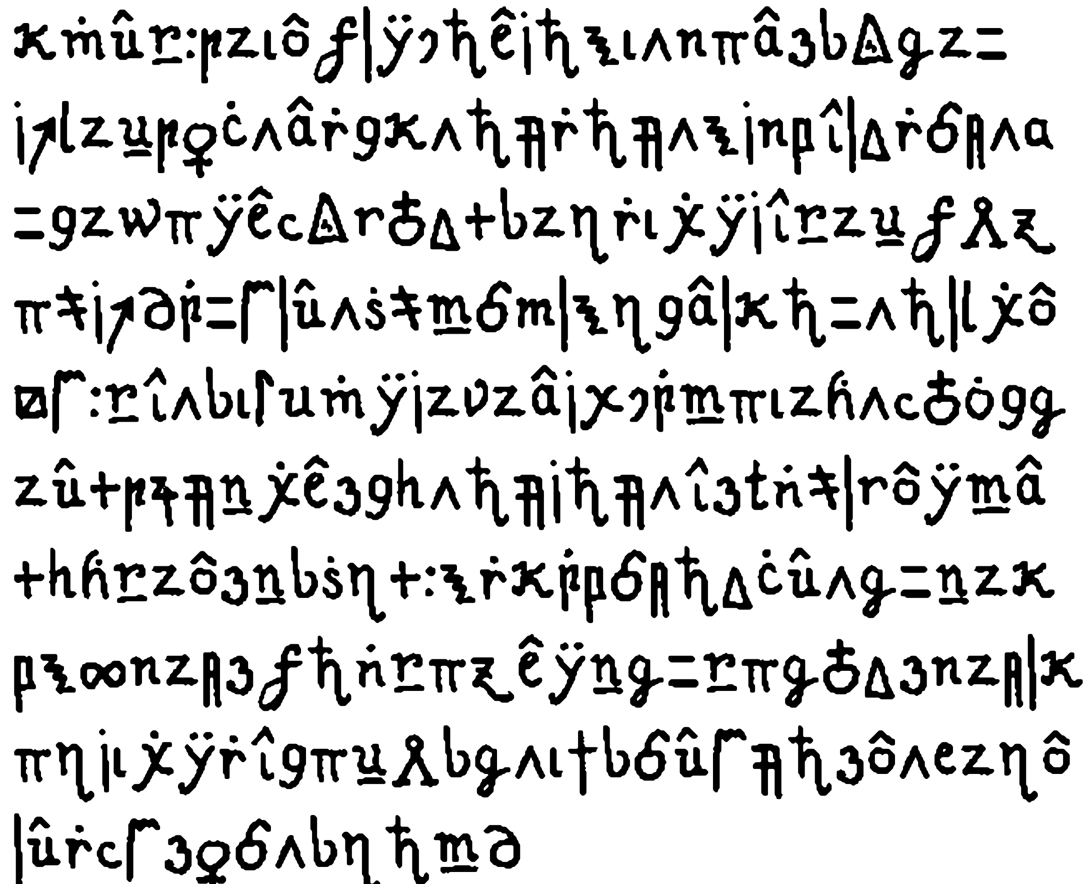
FAIL

Second Decipherment Approach

Homophonic cipher, e.g.:

A =
B =
C =
D =
E =
F =
G =

etc.



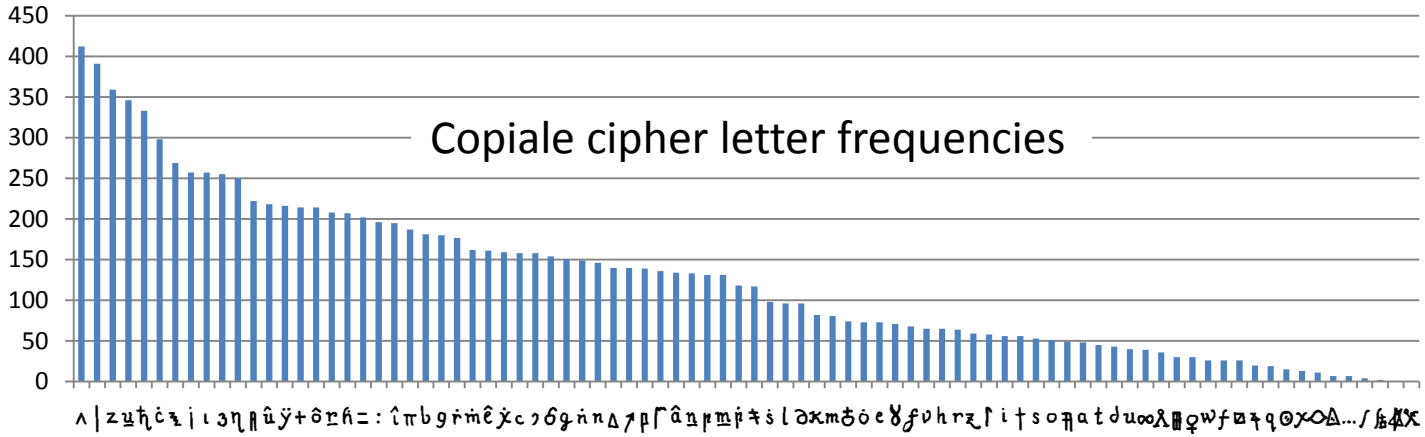
Homophonic Cipher

Assign cipher-letter types based on plaintext-letter frequencies

→ uniform distribution of cipher types!

→ harder to crack

but recall:



on the other hand:

Homophonic Cipher

- Let's try computer attack, assuming plaintext English, then Latin, then French ...
  - [Knight et al 06] method, using letter n-gram LM
- But suppose we fail:
  - was it because the cipher is not homophonic?
  - or because homophonic ciphers at this length cannot be reliably solved by our method?
- Reality check:
  - first set up synthetic, actual homophonic cipher
  - yes, we can solve it! (and ID the plaintext language)

Homophonic Cipher

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

FAIL

Extremely slight numerical preference for German ... random noise?

What Next?

- Concentrate on German
  - book located in Germany
  - slight computer preference for German
  - inscription “Philipp 1866”

- Roll up sleeves

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

Digraph Tables

Key Observation #1

In Copiale, ⲥ almost always followed by ⲏ

In German, C almost always followed by H

(German CH is like English QU)

So guess: ⲥ = C, ⲏ = H

One Thing Leads to Another

ɔħ = CH → ɔħʌ = CHT → ʌ = T ?

Each step is guesswork.

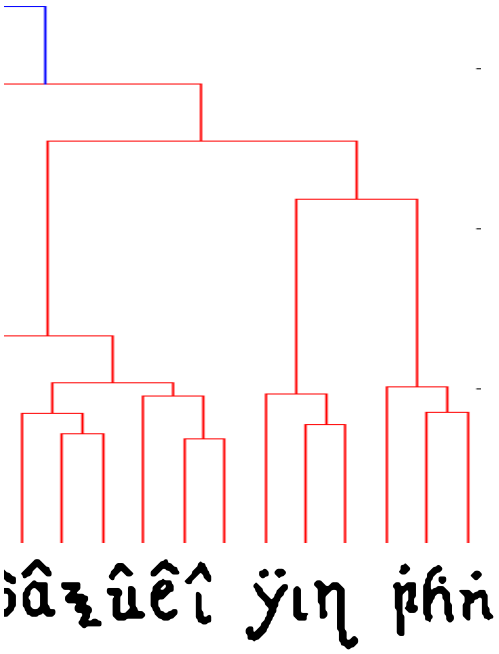
Must be willing to retract.

Weird task, not knowing German.

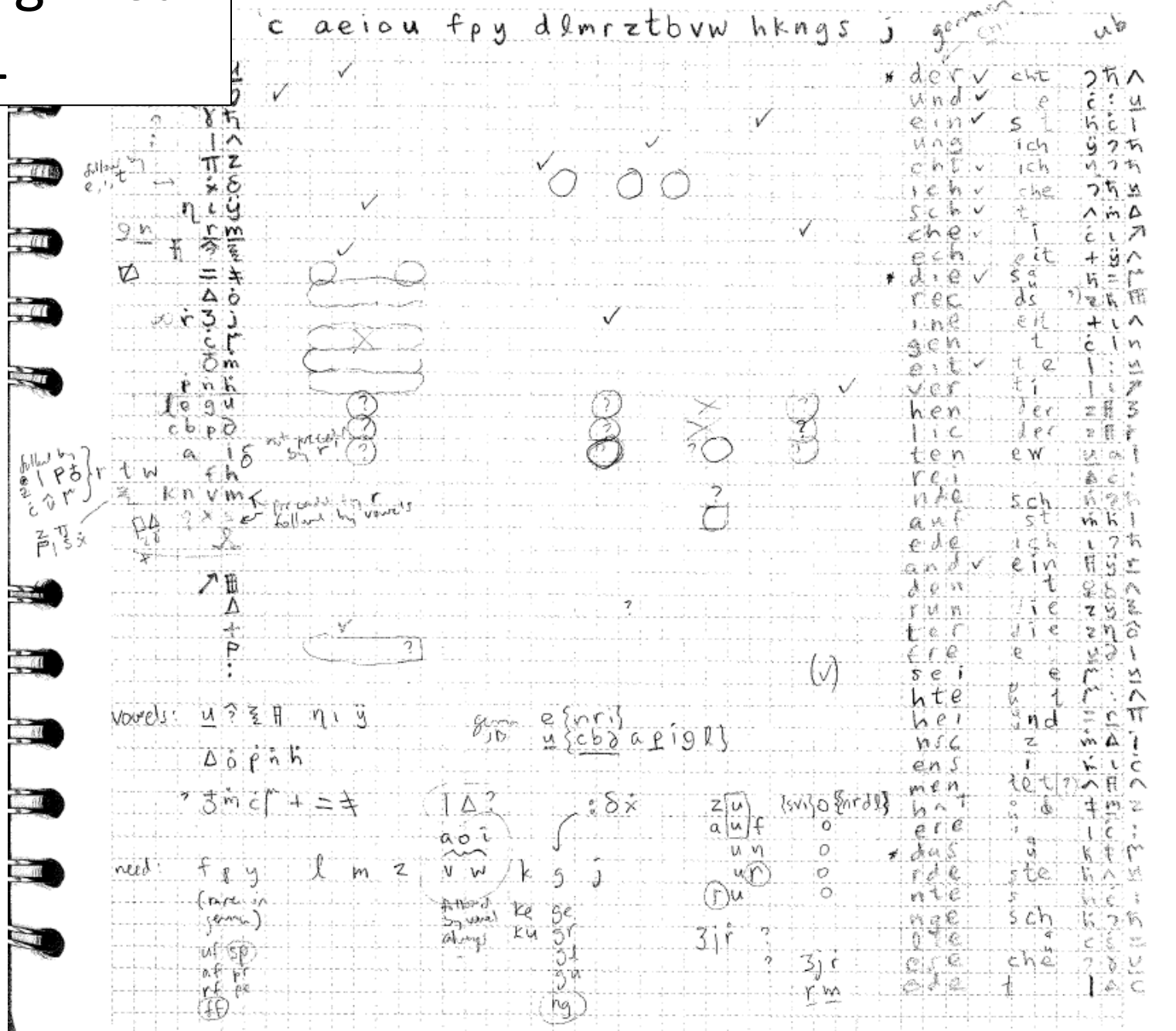
No longer care what the book says.

Cluster diagram crucial:

ÿ = I → ɩ = I , ɳ = I



Spring Break 2011

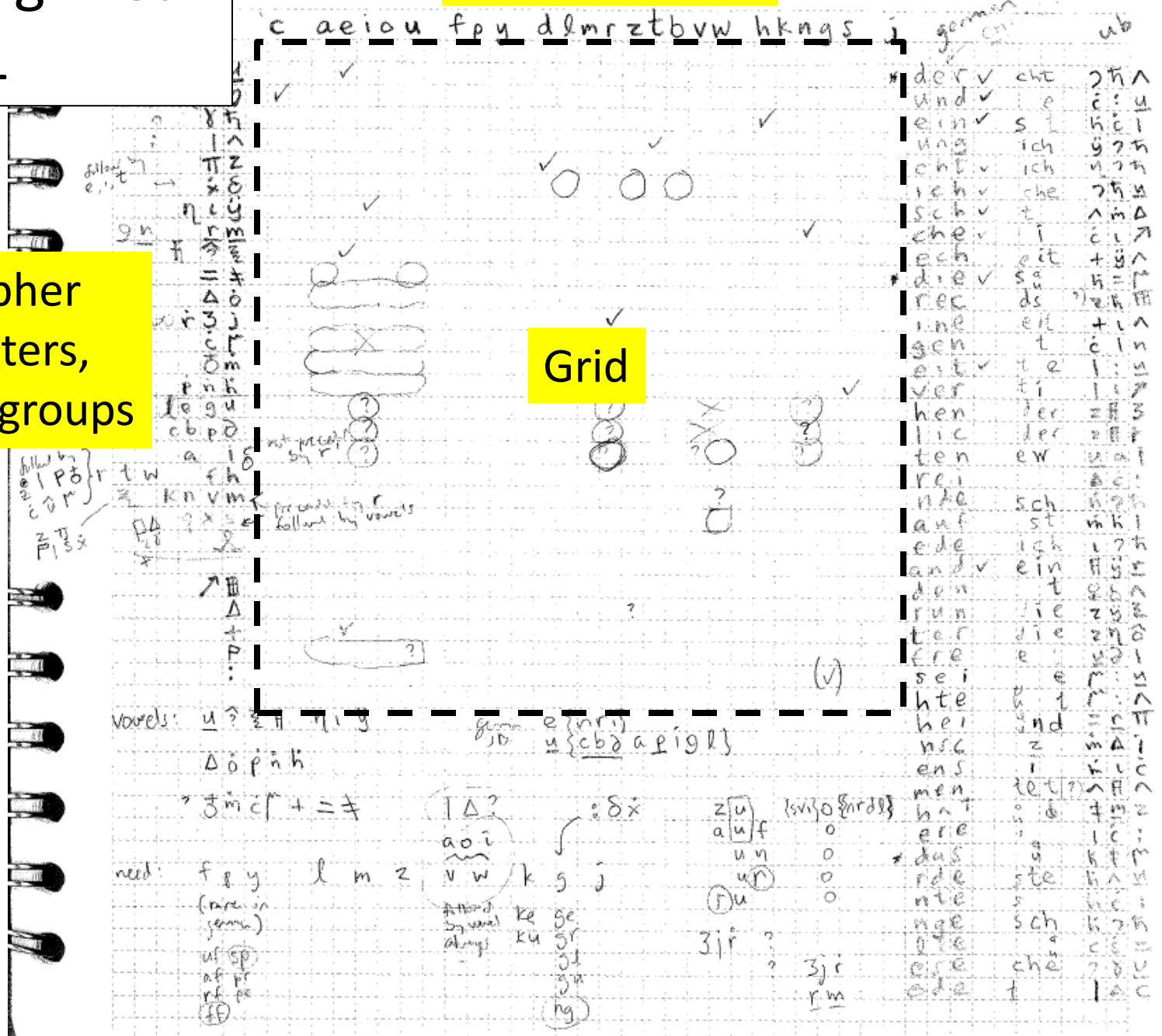


Spring Break 2011

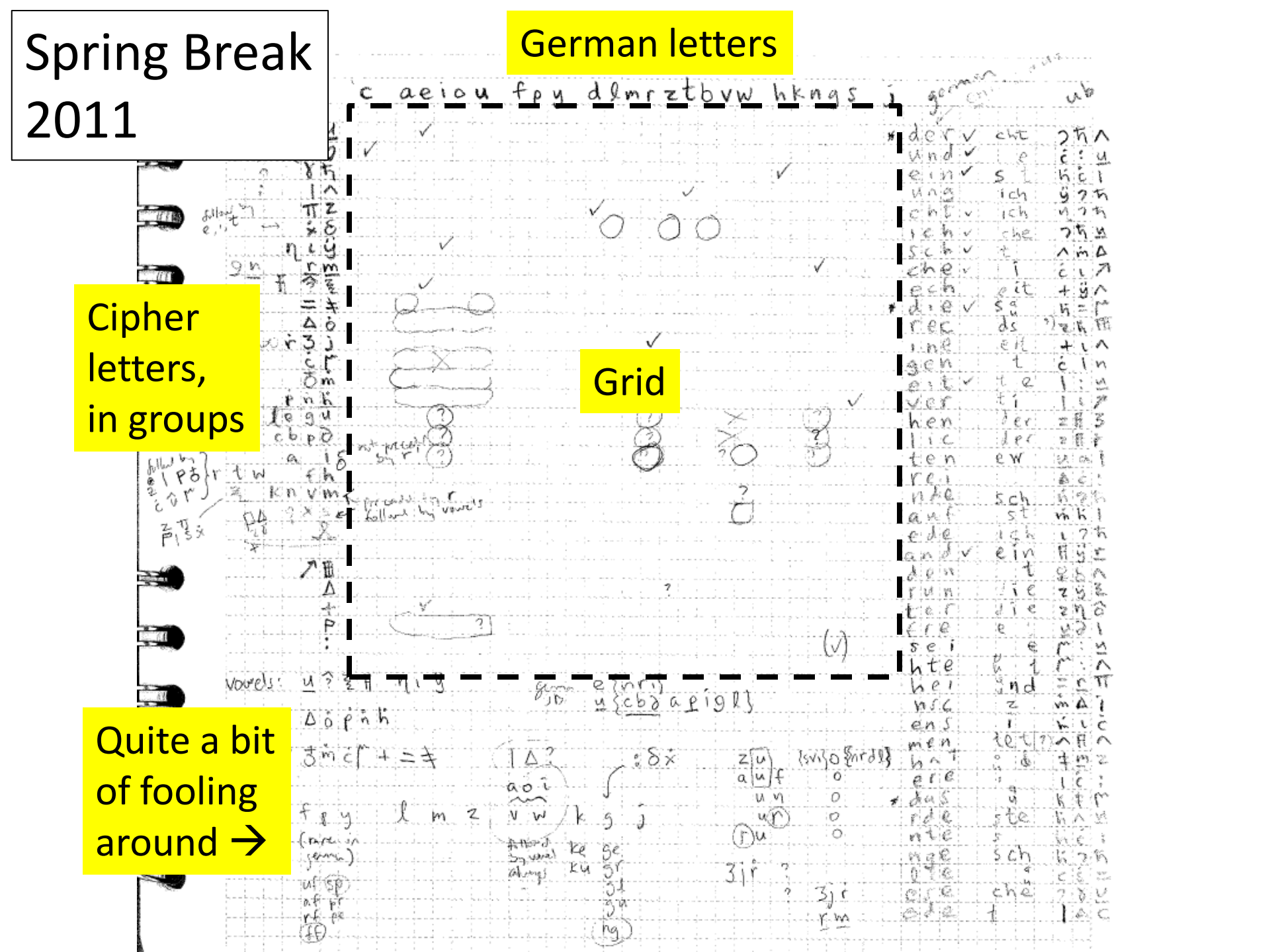
German letters

Cipher letters, in groups

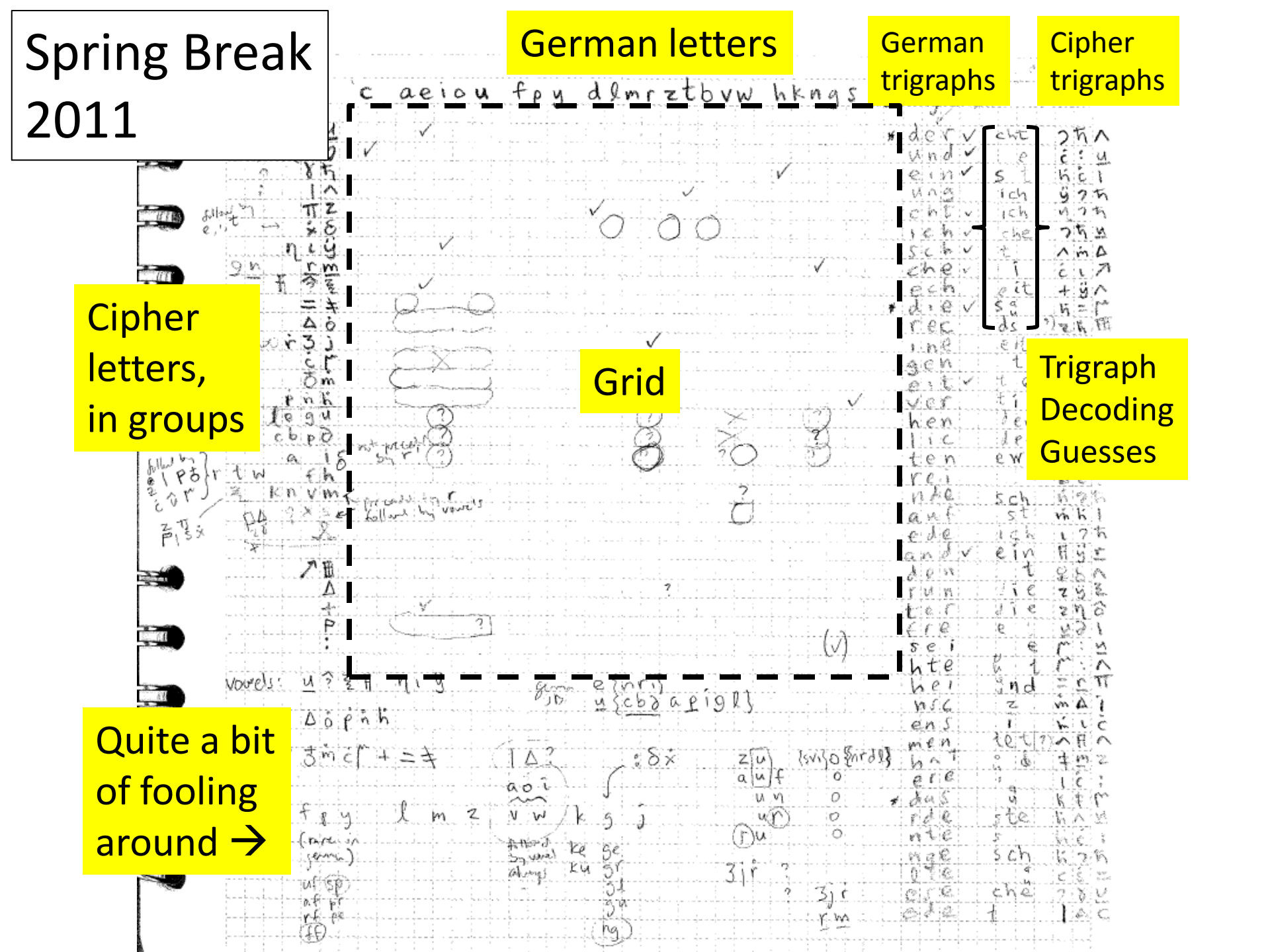
Grid



Spring Break 2011



Spring Break 2011



Problems Remain

Good progress:

- But very few SCH coming out
- No mappings yet for unaccented Roman letters

Decipherment with **?** for cipher letters still without mappings:

?GEHEIMER**?**UNTERLIST**?**VOR**?**DIE**?**GESELLE

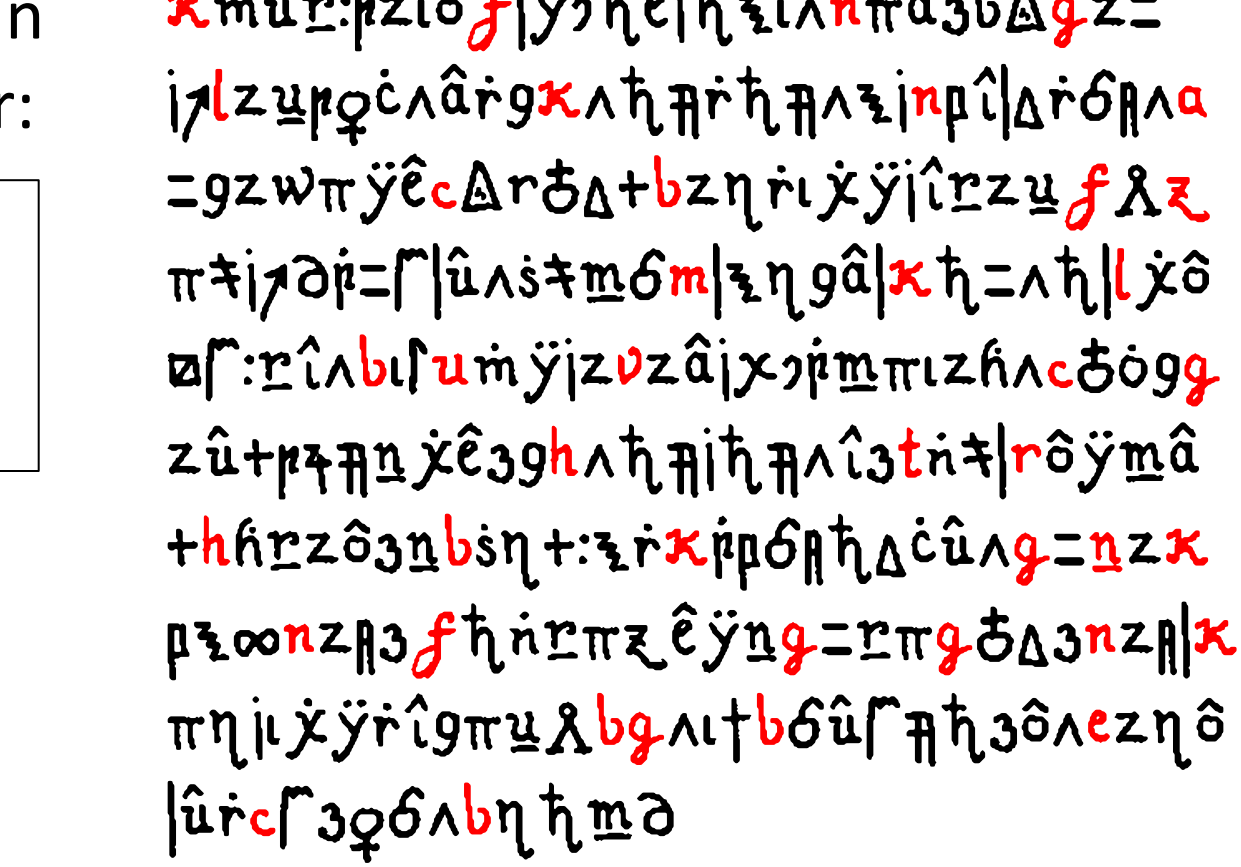
?ERDER**?**TITUL

?CEREMONIE**?**DER**?**AUFNAHME

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

Use Online German-English Translator www.freetranslation.com

 → Erder Abschnitl → Erder Abschnitl

Use Online German-English Translator www.freetranslation.com

→ Erder Abschnitl → Erder Abschnitl

Erder Abschnita → Erder Abschnita

Erder Abschnitb → Erder Abschnitb

Erder Abschnitc → Erder Abschnitc

…

Erder Abschnitt → Erder Section

Cipher ":" maps to German L

Cipher ":" maps to German T

Another Strange Letter: †

†AFLNER, †NUPFTUCHS, and GESELL†AFLT

Tried substituting A-Z for †

Found GESELLSCHAFLT in dictionary, **†=SCH**

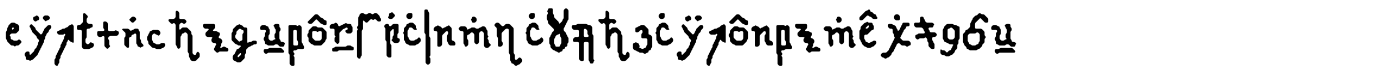
Led to others multi-letter substitutions:

ûıſêɜ ṗp†ɡɳʌ: → Erder Abschnitt → First Section

st

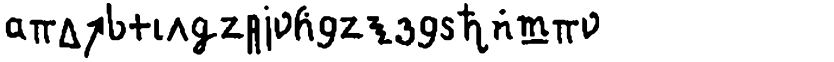
ſ=ST

German Native Edits Decipherment



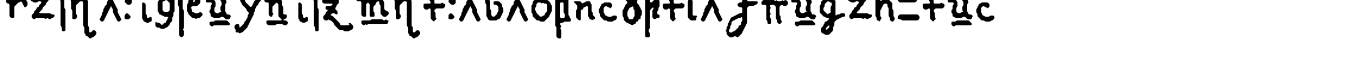
hypothesis: is mache ebenfals wilhuhrlise bewegunge

correction: i**ch** mache ebenfals wil**kü**hrli**ch**e bewegunge**n**



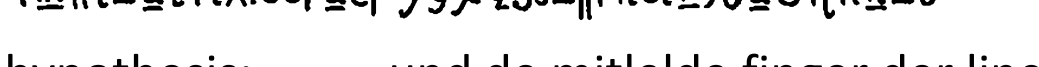
hypothesis: dos mit der andern hand

correction: do**ch** mit der andern hand



hypothesis: dritlens einer n mlt tobach mit de daume

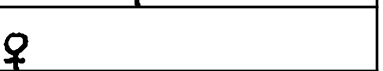
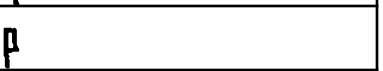
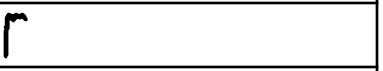
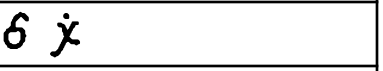
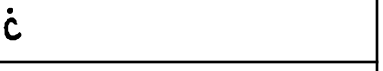
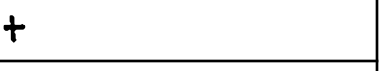
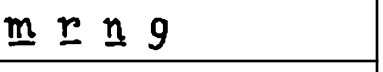
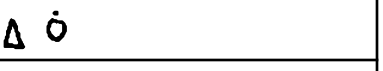
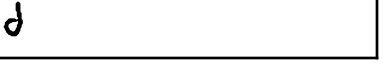
correction: drit**t**ens einer **? ???** tobach mit de**m** daum**en**

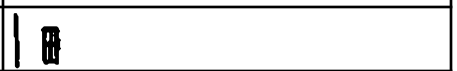
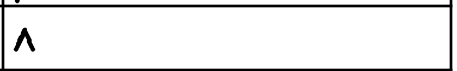
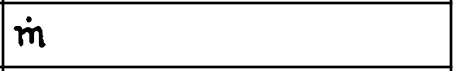
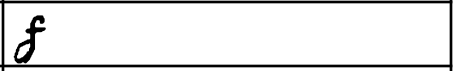
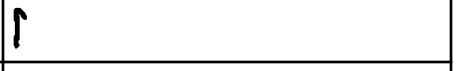
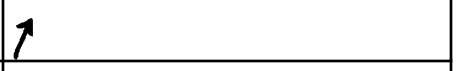
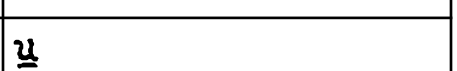
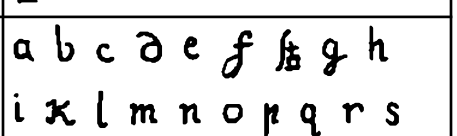


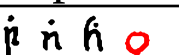
hypothesis: und de mitlelde finger der linche hand

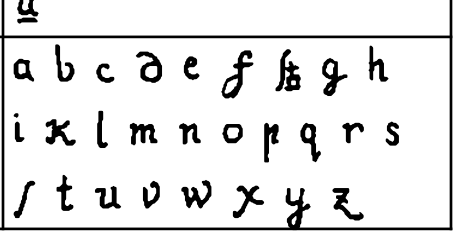
correction: und de**m** mit**t**el**st**en finger der linche**n** hand

Substitution Table

German	Ciphertext
A	
Ä	
B	
C	
D	
E	
F	
G	
H	
I	
J	
K	
L	
M	
N	
O	
Ö	
P	

German	Ciphertext
R	
S	
T	
U	
Ü	
V	
W	
X	
Y	
Z	
SCH	
SS	
ST	
CH	
repeat	
EN / EM	
space	

German	Ciphertext
A	
Ä	
B	
C	
D	
E	
F	
G	
H	
I	
J	
K	
L	
M	
N	
O	
Ö	
P	

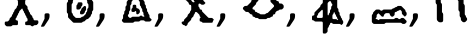
German	Ciphertext
R	
S	
T	
U	
Ü	
V	
W	
X	
Y	
Z	
SCH	
SS	
ST	
CH	
repeat	
EN / EM	
space	

Substitution Table

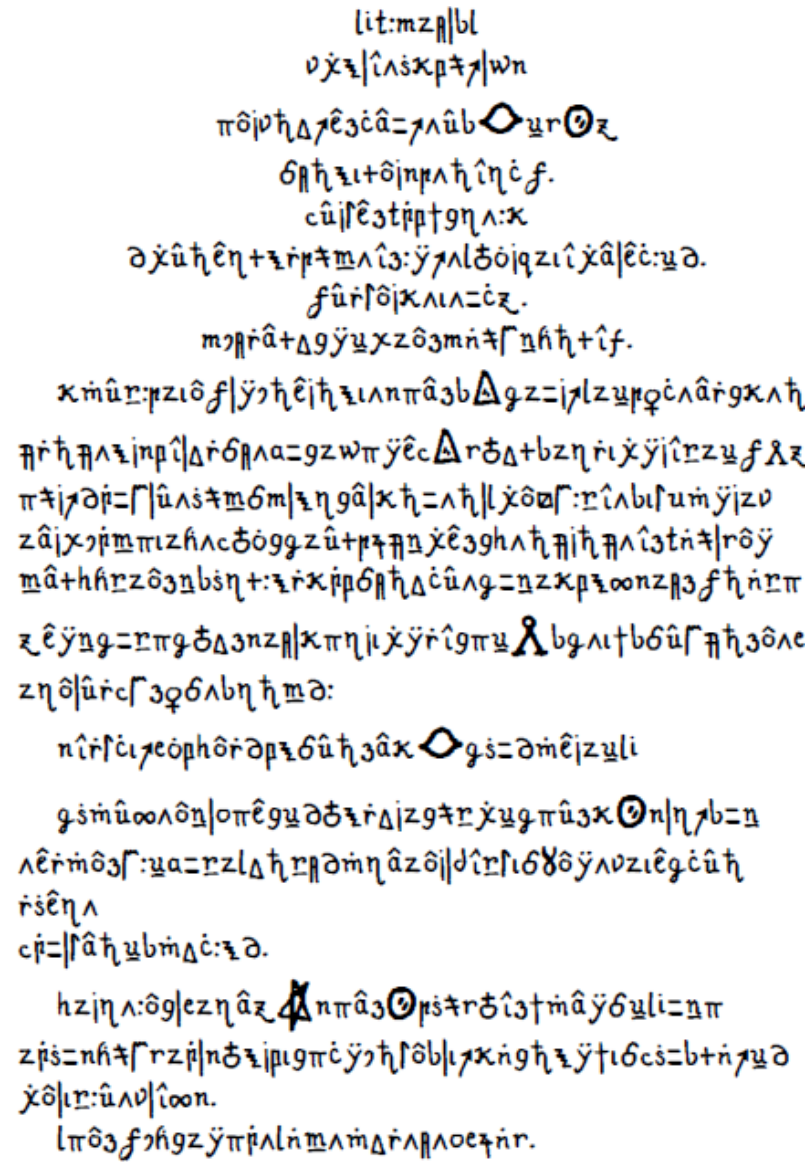
Cipher letters in **RED** decode ambiguously ...

But only with respect to modern German, not old German.

No entries for large characters:



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 ⅄ 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 ⅄ tisch geführet dieser frägt ihn:

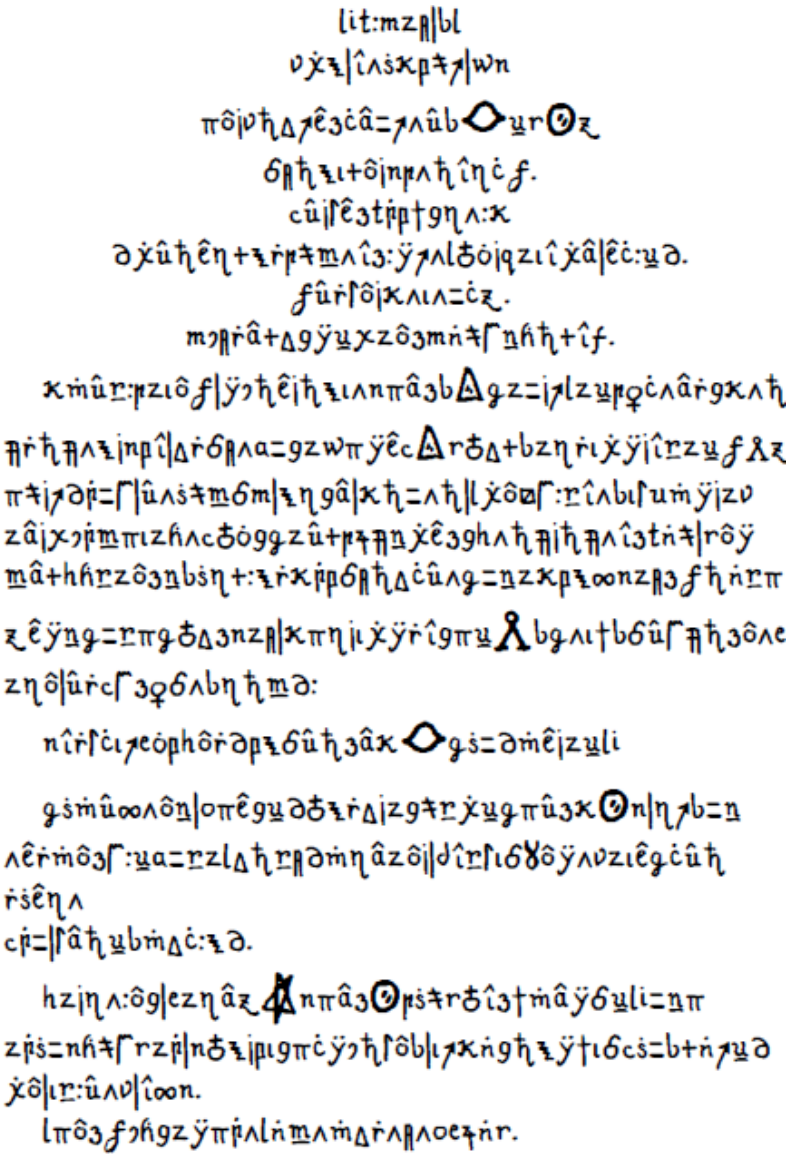
erstlich ob er begehre ◇ zu werden

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

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

der candidat antwortet ja.

Copiale Decipherment



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.

Lessons

- Important to “dig in”
- Computer methods are inflexible
  - they only find what you tell them to find
- People are very flexible
  - excellent pattern matchers

- That’s why noisy MT corpora are cleaned up by grad students, not by computers …

Back to Machine Translation!

- “When I look at an article in Russian, I say to myself, this is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.” -- Warren Weaver (1947)
- What kind of a code is Russian?
  - At the very least, words get translated, inserted, deleted, and re-ordered ...
- Can we learn a translation model without parallel text?

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& @shr&
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

!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 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

Time Expressions

Hzyr!n

229	fy Hzyr!n !lm!Dy
207	fy Hzyr!n
75	fy Hzyr!n !lmqbl
61	fy Hzyr!n 1993
31	fy Hzyr!n 1992
27	!lr!b@ mn Hzyr!n
27	fy Hzyr!n 1967
19	fy 30 Hzyr!n !lm!Dy
18	fy n=h!y& Hzyr!n !lm!Dy
18	fy Hzyr!n 1991
17	mn Hzyr!n
17	mndh Hzyr!n !lm!Dy
17	4 Hzyr!n
16	n=h!y& Hzyr!n !lm!Dy
16	fy Hzyr!n 1990
15	sh=hr Hzyr!n
15	fy sh=hr Hzyr!n !lm!Dy
15	fy Hzyr!n 1994
14	mn 17 Hzyr!n
14	fy Hzyr!n 1996
14	fy 30 Hzyr!n
13	fy sh=hr Hzyr!n
13	fy 20 Hzyr!n !lm!Dy
13	4 Hzyr!n 1967
12	n=h!y& Hzyr!n
12	!lr!b@ mn Hzyr!n 1967

Time Expressions

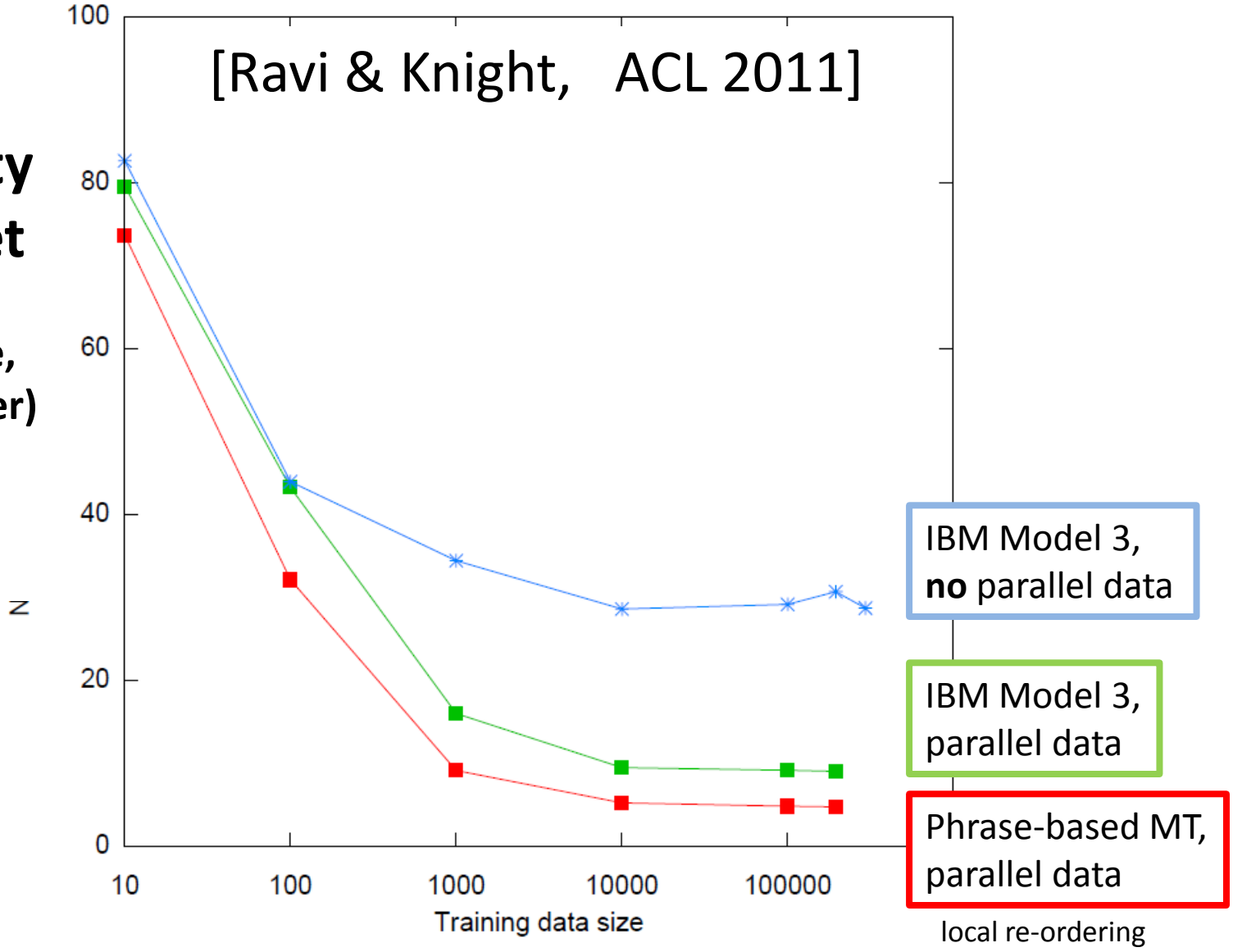
Hzyr!n

229	fy Hzyr!n !lm!Dy
207	fy Hzyr!n
75	fy Hzyr!n !lmqbl
61	fy Hzyr!n 1993
31	fy Hzyr!n 1992
27	!lr!b@ mn Hzyr!n
27	fy Hzyr!n 1967
19	fy 30 Hzyr!n !lm!Dy
18	fy n=h!y& Hzyr!n !lm!Dy
18	fy Hzyr!n 1991
17	mn Hzyr!n
17	mndh Hzyr!n !lm!Dy
17	4 Hzyr!n
16	n=h!y& Hzyr!n !lm!Dy
16	fy Hzyr!n 1990
15	sh=hr Hzyr!n
15	fy sh=hr Hzyr!n !lm!Dy
15	fy Hzyr!n 1994
14	mn 17 Hzyr!n
14	fy Hzyr!n 1996
14	fy 30 Hzyr!n
13	fy sh=hr Hzyr!n
13	fy 20 Hzyr!n !lm!Dy
13	4 Hzyr!n 1967
12	n=h!y& Hzyr!n
12	!lr!b@ mn Hzyr!n 1967

Deciphering Spanish Time Expressions

MT quality on test set

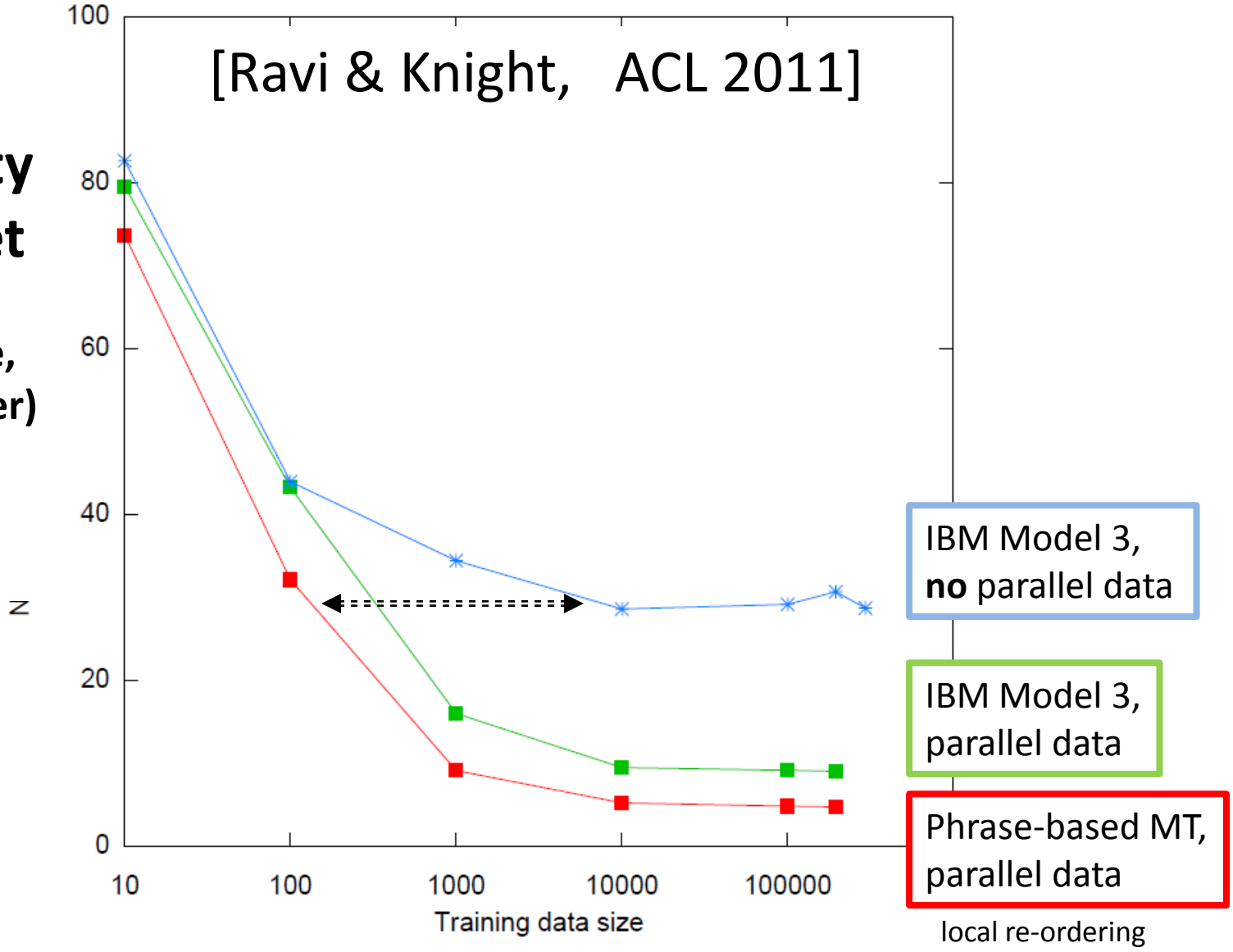
(Edit distance, lower is better)



Deciphering Spanish Time Expressions

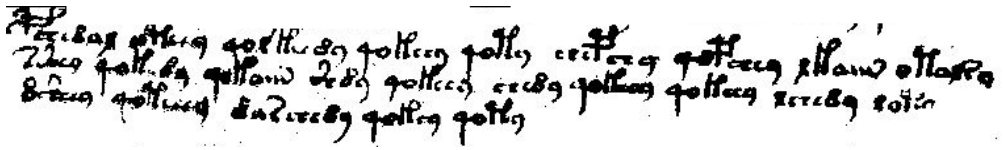
MT quality on test set

(Edit distance, lower is better)



More to Do

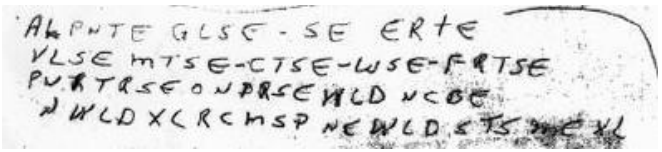
Voynich Manuscript (1400)



Machine Translation (1947-)

Zodiac 408 Serial Killer (1967)

FBI cipher (1999)



Kryptos (1990)

```
                                                OBKR
UOXOGHULBSOLIFBBWFLRVQQPRNGKSSO
TWTQSJQSSEKZZWATJKLUDIAWINFBNYP
VTTMZFPKWGDKZXTJCDIGKUHUAUEKCAR
```

WARREN WEAVER

ለኪዮሠጸጨ !

Homophonic P(cipher) for Copiale Cipher

EM, multiple restarts, uniform start first, best start **bold**. About 5 hours per language.

english	… **26303** …
french	29846 … 27137 … **26759** …
german	29514 26561 **25672** 26590 …
italian	29410 … 26979 26922 26828 26911 … 27370 … **26741**
latin	27487 … **26349**
portuguese	… **26479**
spanish	… **26533** …
welsh	26723 26936 26817 26904 26682 26749 26573 **26501** 26520 26687 26811

Clustering English Letters