

Language Translation and Code-Breaking

Kevin Knight
Information Sciences Institute
University of Southern California

joint work with:
Sujith Ravi, Qing Dou (USC)
Beáta Megyesi, Christiane Schaefer (Uppsala)
Regina Barzilay, Ben Snyder (MIT)
Sravana Reddy (Chicago)

LXMLS Lisbon

July 2013

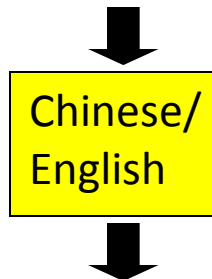
Machine Translation

美国关岛国际机场及其办公室均接获一名自称沙地阿拉伯富商拉登等发出的电子邮件，威胁将会向机场等公众地方发动生化袭击後，关岛经保持高度戒备。

Kowane mutum na da hakkin ya sami yancin yin tunani da na sanin yakamata da na bin addini; saboda haka yana da yancin sake addini ko ra'ayin da ya bada gaskiya gare shi, da kuma yancin nuna addininsa ko ra'ayinsa, shi daya ko a cikin taro kuma a fili ko a boye ta hanyar koyarwa ko yin ibada, ko bauta wa abin da ya bada gaskiya gare shi da yin abubuwan da abin da yake bauta wa din ya nuna masa.

Machine Translation

美国关岛国际机场及其办公室均接获一名自称沙地阿拉伯富商拉登等发出的电子邮件，威胁将会向机场等公众地方发动生化袭击後，关岛经保持高度戒备。



The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport.

Kowane mutum na da hakkin ya sami yancin yin tunani da na sanin yakamata da na bin addini; saboda haka yana da yancin sake addini ko ra'ayin da ya bada gaskiya gare shi, da kuma yancin nuna addininsa ko ra'ayinsa, shi daya ko a cikin taro kuma a fili ko a boye ta hanyar koyarwa ko yin ibada, ko bauta wa abin da ya bada gaskiya gare shi da yin abubuwan da abin da yake bauta wa din ya nuna masa.



Everyone has the right to freedom of thought, conscience and religion; this right includes freedom to change his religion or belief, and freedom, either alone or in community with others and in public or private, to manifest his religion or belief in teaching, practice, worship and observance.

Statistical Machine Translation

“When I look at an article in Russian, I say: this is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.” -- Warren Weaver (1947)

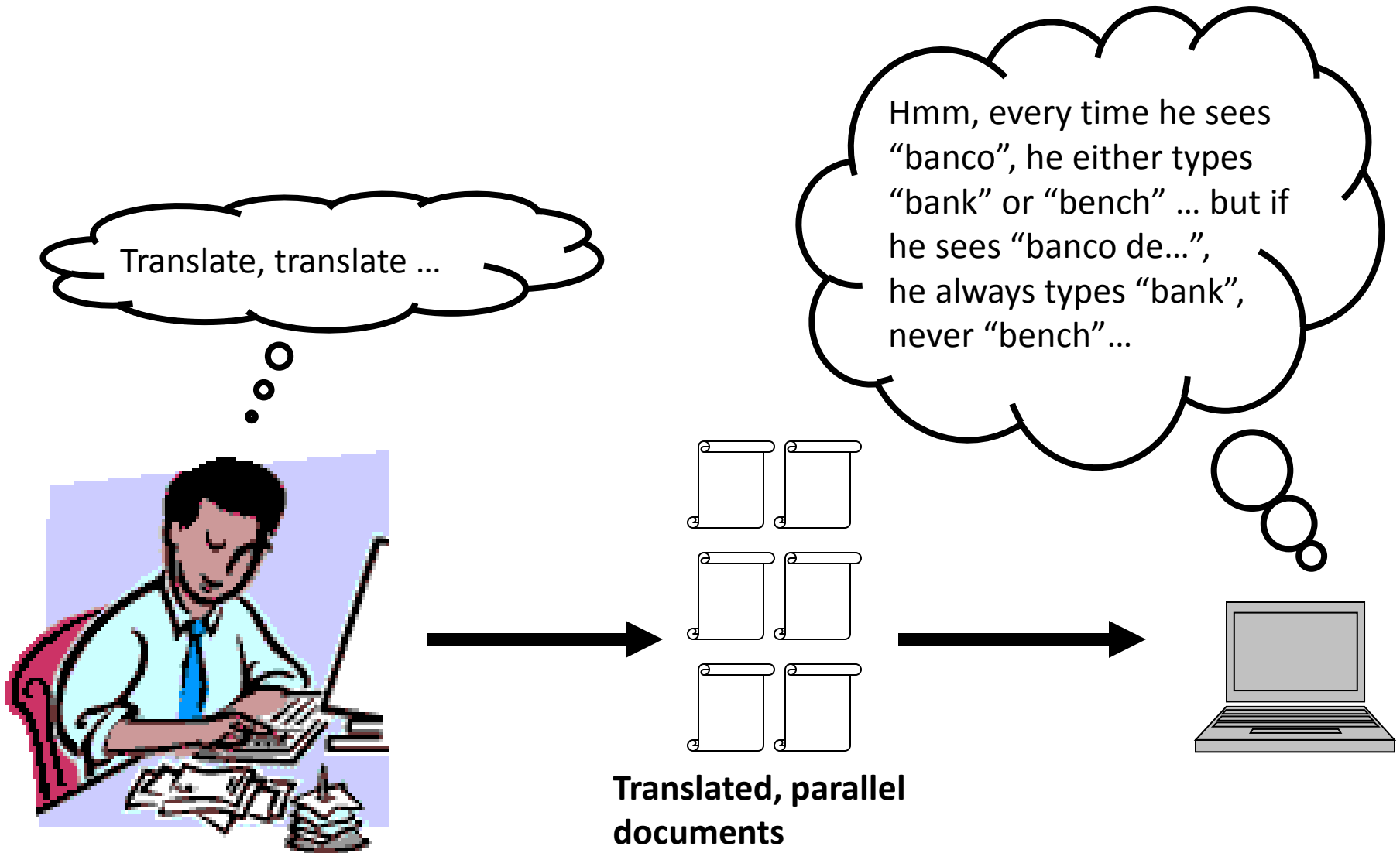


OUR HERO

Weaver saw a colleague decoding intercepts into Turkish, without “knowing” Turkish.

... maybe a computer could translate into English without “knowing” English?

Statistical Machine Translation



Parallel Corpus

12 English sentences in English and Spanish.

1a. Garcia and associates .
1b. Garcia y asociados .

7a. the clients and the associates are enemies .
7b. los clients y los asociados son enemigos .

2a. Carlos Garcia has three associates .
2b. Carlos Garcia tiene tres asociados .

8a. the company has three groups .
8b. la empresa tiene tres grupos .

3a. his associates are not strong .
3b. sus asociados no son fuertes .

9a. its groups are in Europe .
9b. sus grupos estan en Europa .

4a. Garcia has a company also .
4b. Garcia tambien tiene una empresa .

10a. the modern groups sell strong pharmaceuticals .
10b. los grupos modernos venden medicinas fuertes .

5a. its clients are angry .
5b. sus clientes estan enfadados .

11a. the groups do not sell zenzanine .
11b. los grupos no venden zanzanina .

6a. the associates are also angry .
6b. los asociados tambien estan enfadados .

12a. the small groups are not modern .
12b. los grupos pequenos no son modernos .

Centauri/Arcturan

Your assignment, translate this to Arcturan: farok crrrok hihok yorok klok kantok ok-yurp

1a. ok-voon ororok sprok .	7a. lalok farok ororok lalok sprok izok enemok .
1b. at-voon bichat dat .	7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anak plok sprok .	8a. lalok brok anak plok nok .
2b. at-drubel at-voon pippat rrat dat .	8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghrok .	9a. wiwok nok izok kantok ok-yurp .
3b. totat dat arrat vat hilat .	9b. totat nnat quat oloat at-yurp .
4a. ok-voon anak drok brok jok .	10a. lalok mok nok yorok ghrok klok .
4b. at-voon krat pippat sat lat .	10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok .	11a. lalok nok crrrok hihok yorok zanzanak .
5b. totat jjat quat cat .	11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok .	12a. lalok rarok nok izok hihok mok .
6b. wat dat krat quat cat .	12b. wat nnat forat arrat vat gat .

Centauri/Arcturan

Your assignment, translate this to Arcturan: **farok** crrrok hihok yorok klok kantok ok-yurp

1a. ok-voon ororok sprok .	7a. lalok farok ororok lalok sprok izok enemok .
1b. at-voon bichat dat .	7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anak plok sprok .	8a. lalok brok anak plok nok .
2b. at-drubel at-voon pippat rrat dat .	8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghrok .	9a. wiwok nok izok kantok ok-yurp .
3b. totat dat arrat vat hilat .	9b. totat nnat quat oloat at-yurp .
4a. ok-voon anak drok brok jok .	10a. lalok mok nok yorok ghrok klok .
4b. at-voon krat pippat sat lat .	10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok .	11a. lalok nok crrrok hihok yorok zanzanak .
5b. totat jjat quat cat .	11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok .	12a. lalok rarok nok izok hihok mok .
6b. wat dat krat quat cat .	12b. wat nnat forat arrat vat gat .

Centauri/Arcturan

Your assignment, translate this to Arcturan: farok **crrok** hihok yorok klok kantok ok-yurp

1a. ok-voon ororok sprok .	7a. lalok farok ororok lalok sprok izok enemok .
1b. at-voon bichat dat .	7b. wat jjat bichat wat dat vat eneak .
2a. ok-drubel ok-voon anak plok sprok .	8a. lalok brok anak plok nok .
2b. at-drubel at-voon pippat rrat dat .	8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghrok .	9a. wiwok nok izok kantok ok-yurp .
3b. totat dat arrat vat hilat .	9b. totat nnat quat oloat at-yurp .
4a. ok-voon anak drok brok jok .	10a. lalok mok nok yorok ghrok klok .
4b. at-voon krat pippat sat lat .	10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok .	11a. lalok nok crrok hihok yorok zanzanak .
5b. totat jjat quat cat .	11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok .	12a. lalok rarok nok izok hihok mok .
6b. wat dat krat quat cat .	12b. wat nnat forat arrat vat gat .

Centauri/Arcturan

Your assignment, translate this to Arcturan: **farok** crrok **hihok** **yorok** **clok** kantok ok-yurp

1a. ok-voon ororok sprok .	7a. lalok farok ororok lalok sprok izok enemok . /
1b. at-voon bichat dat .	7b. wat jjat bichat wat dat vat eneak .
2a. ok-drubel ok-voon anak plok sprok .	8a. lalok brok anak plok nok . /
2b. at-drubel at-voon pippat rrat dat .	8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghrok . / /	9a. wiwok nok izok kantok ok-yurp .
3b. totat dat arrat vat hilat .	9b. totat nnat quat oloat at-yurp .
4a. ok-voon anak drok brok jok .	10a. lalok mok nok yorok ghrok clok . / / /
4b. at-voon krat pippat sat lat .	10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . /	11a. lalok nok crrok hihok yorok zanzanok . / / /
5b. totat jjat quat cat .	11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 	12a. lalok rarok nok izok hihok mok . / / /
6b. wat dat krat quat cat .	12b. wat nnat forat arrat vat gat .

Centauri/Arcturan

Your assignment, translate this to Arcturan: **farok** crrok **hihok** **yorok** **clok** kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . / 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anak plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anak plok nok . / 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghrok . / / 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anak drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghrok clok . / / 10b. wat nnat gat mat bat hilat . process of elimination
5a. wiwok farok izok stok . / 5b. totat jjat quat cat .	11a. lalok nok crrok hihok yorok zanzanok . / / 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . / / 12b. wat nnat forat arrat vat gat .

Statistical Machine Translation

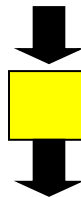
1999

- one page per day, low quality
- limited domains, languages
- no commercial offerings

2013

- fast, 50+ languages
- quality much improved
- products for business, intelligence, end users

أعلن الرئيس الصومالي شريف شيخ أحمد أثناء زيارة لمواقع للقوات الحكومية والأفريقية في حي هودن في مقديشو أن الحملة العسكرية الحالية "لن تتوقف حتى تتحرر الصومال من الشباب والقاعدة".

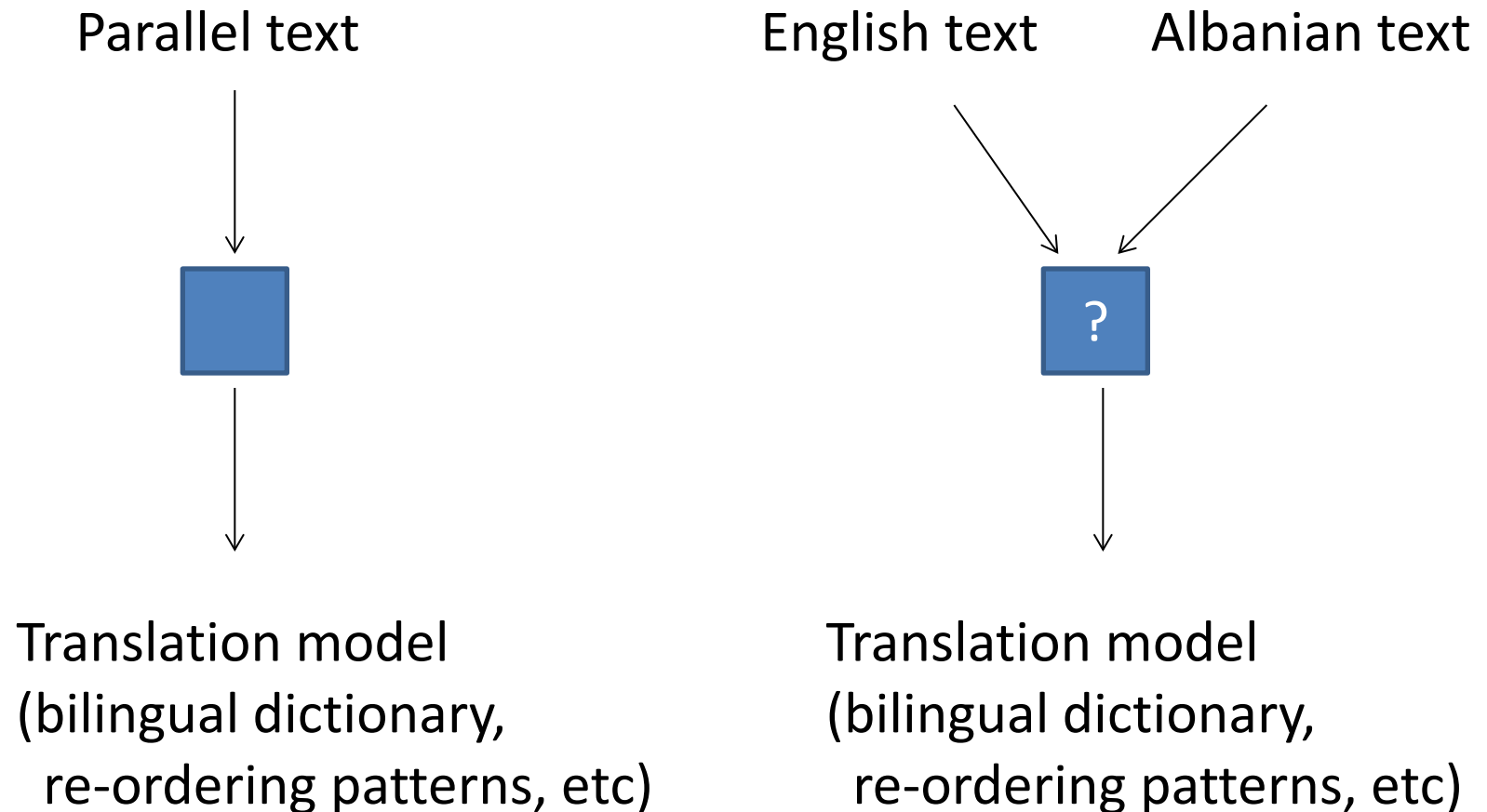


Somali President Sharif Sheikh Ahmed during a visit to sites of government forces and African in the district of Howden in Mogadishu that the current military campaign "will not stop until Somalia is liberated from the youth and al-Qaeda. "

Statistical Machine Translation

- Phrase-based MT
- Syntax-based MT
- Semantics-based MT
- Many ways to mix and match

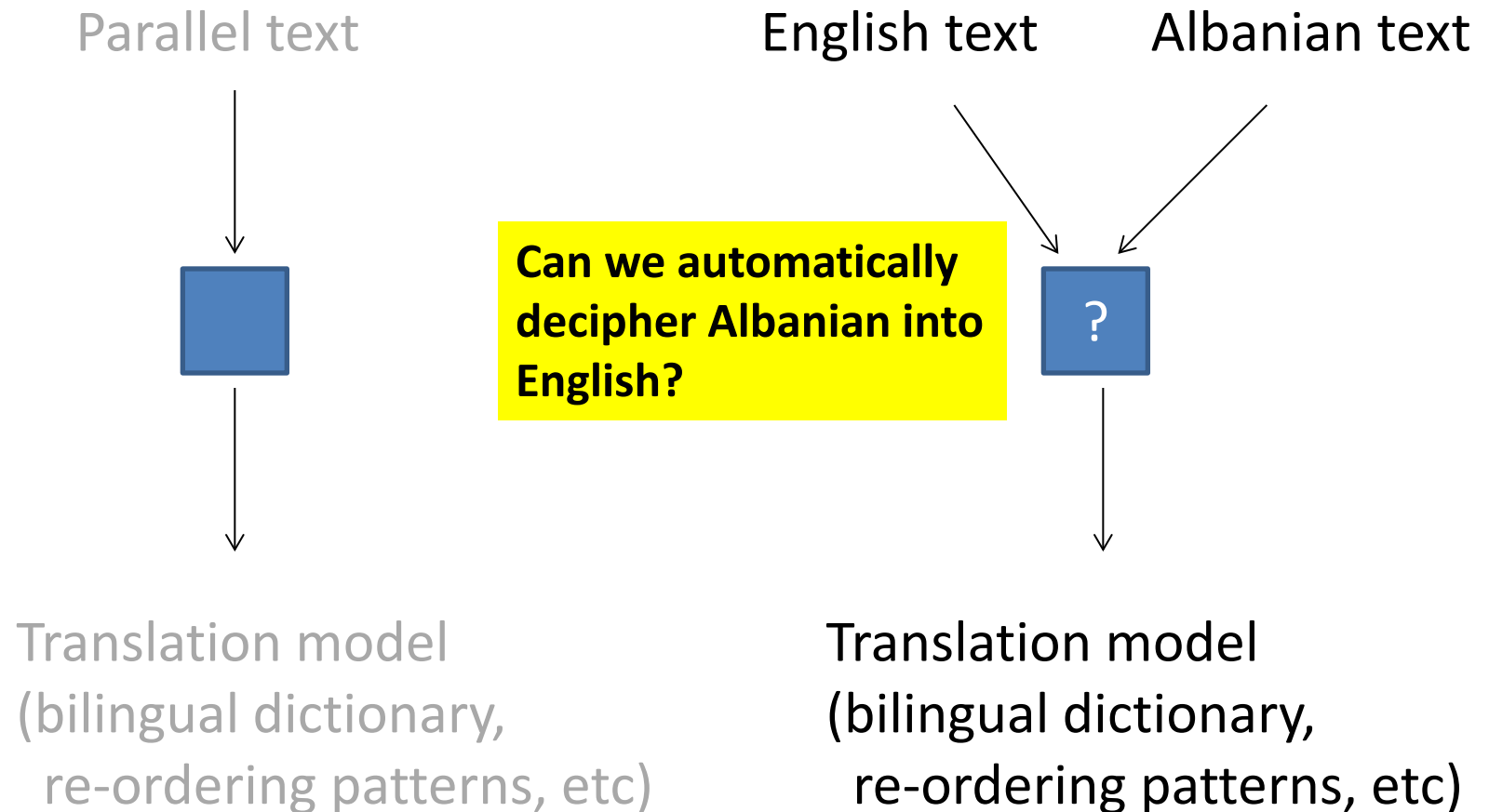
Learn Translation Knowledge from Non-Parallel Text?



African Languages

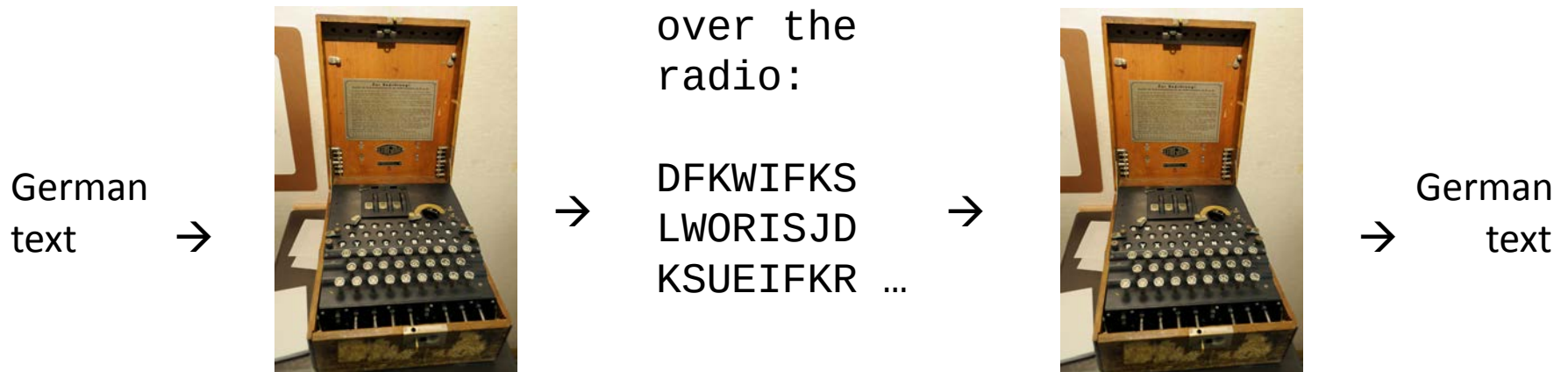


Learn Translation Knowledge from Non-Parallel Text?



“First NLP Task Ever” (1930s-40s)

Breaking the German Enigma Cipher



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

$N \rightarrow J$

Substitution table **changes**
with every keystroke:

$NNN \rightarrow JTE$

Flattens out ciphertext
letter distributions.

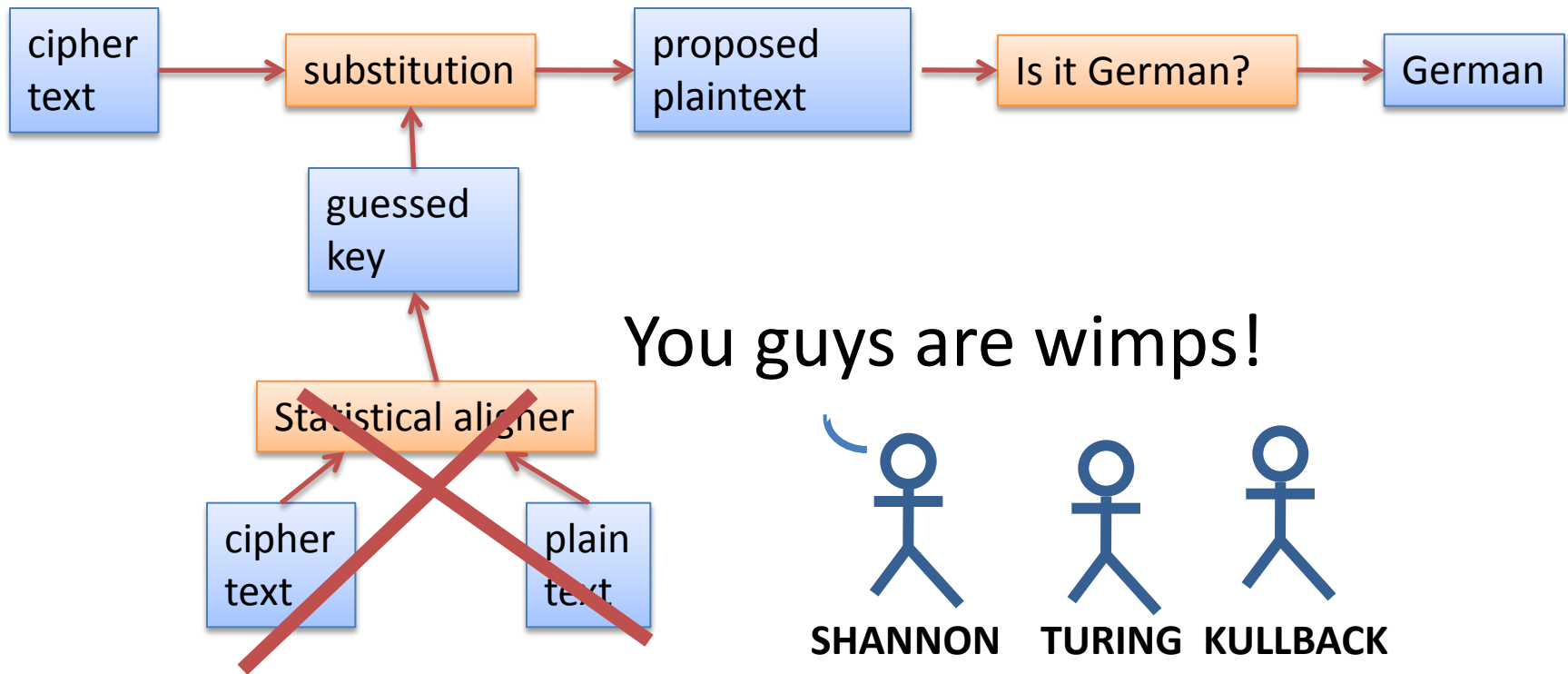


Secret key =
initial rotor
ordering and
settings

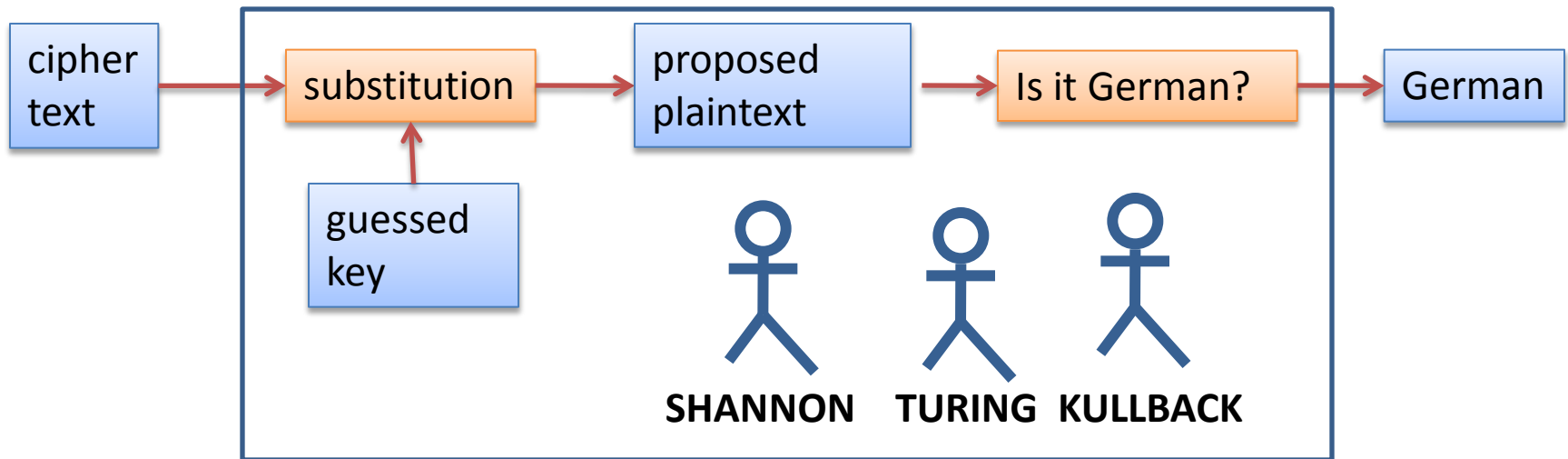
Reversible behavior

$NNN \rightarrow JTE \rightarrow NNN$

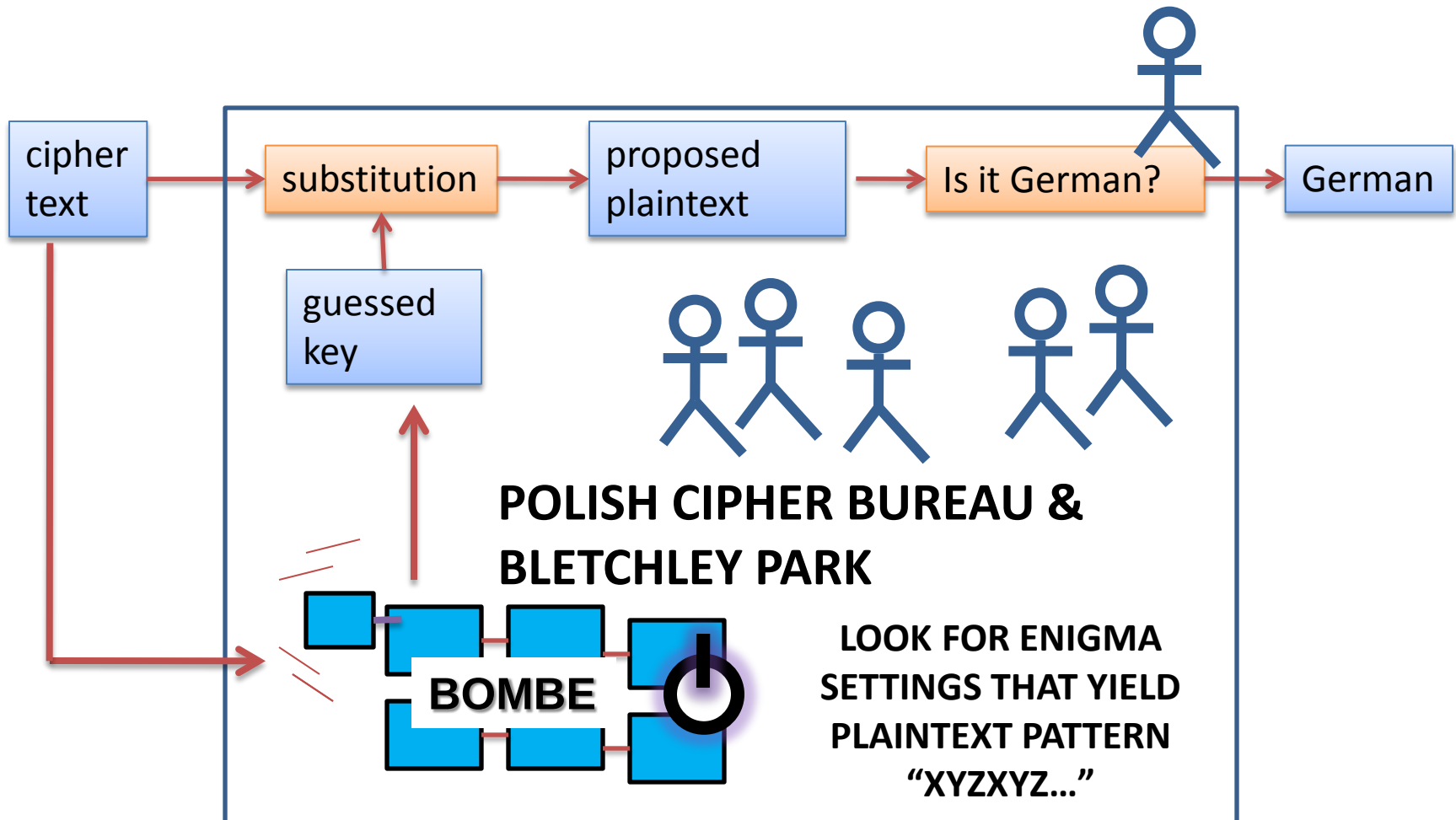
Breaking Enigma



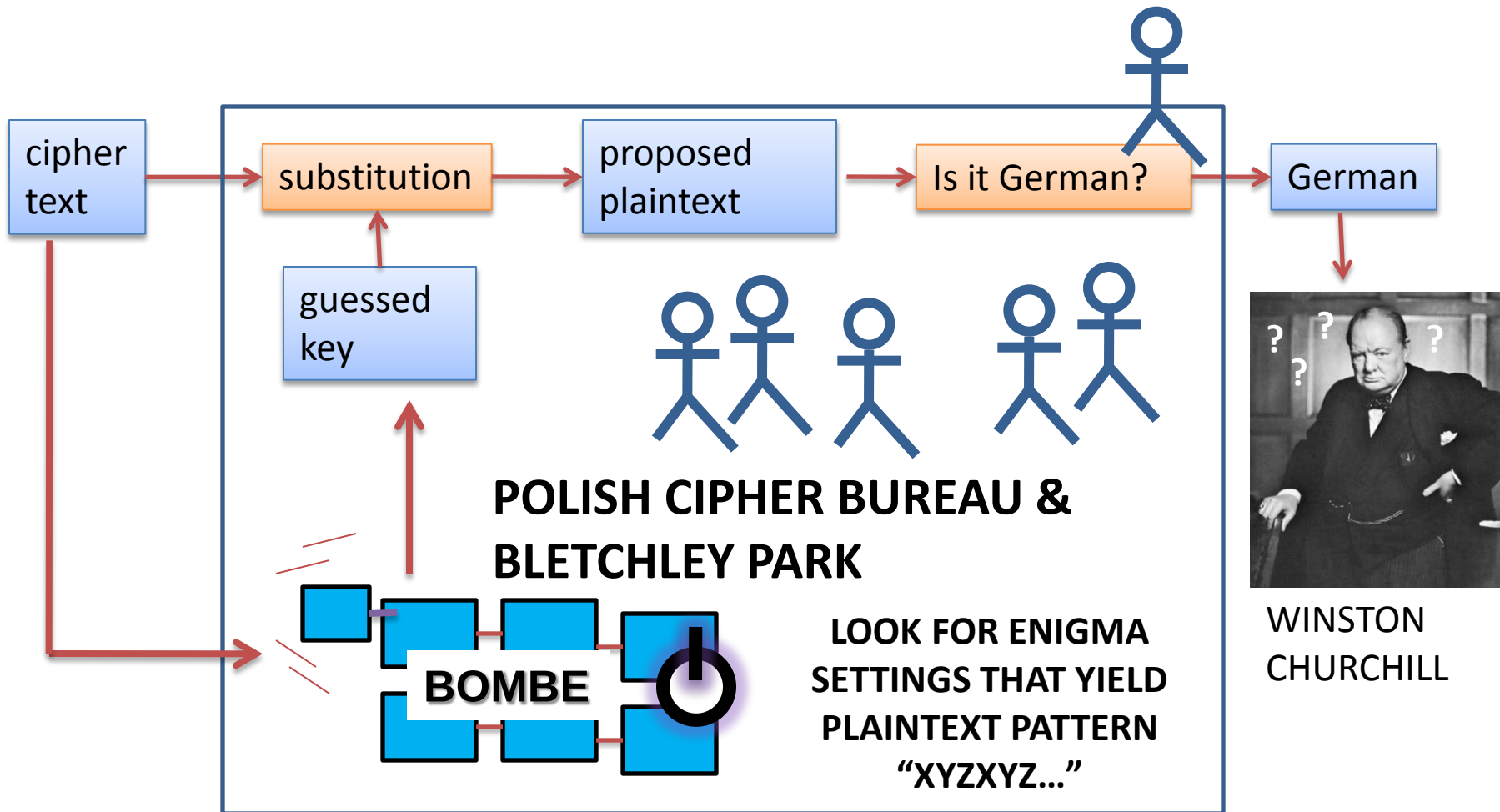
Breaking Enigma



Breaking Enigma



Breaking Enigma

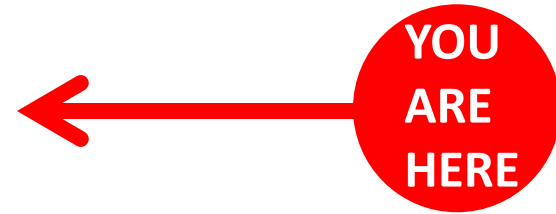


What's the Plan for This Talk?

- Break a series of codes
 - Simple letter substitution
 - Phonetic substitution
 - archaeology
 - transliteration
 - Word substitution
 - Foreign language as cipher
- Bonus
 - Two historical ciphers
 - Final thought on translation and cryptography

What's the Plan for This Talk?

- Break a series of codes
 - Simple letter substitution
 - Phonetic substitution
 - archaeology
 - transliteration
 - Word substitution
 - Foreign language as cipher
- Bonus
 - Two historical ciphers
 - Final thought on translation and cryptography



Letter Substitution Cipher

- Encipherment key:

PLAIN: ABCDEFGHIJKLMNOPQRSTUVWXYZ

CIPHER: P L O K M I J N U H B Y G V T F C R D X E S Z A Q W

- 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	
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 . 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."

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	.

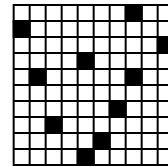
Letter Substitution Cipher

plaintext samples,
unrelated to ciphertext



$P(p)$

plaintext p

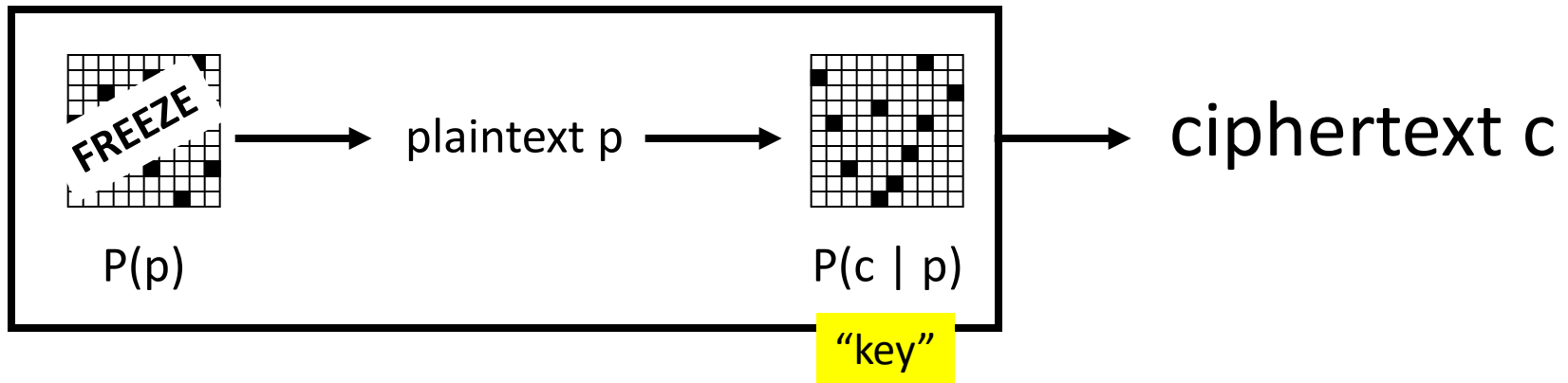


$P(c \mid p)$

"key"

ciphertext c

Letter Substitution Cipher

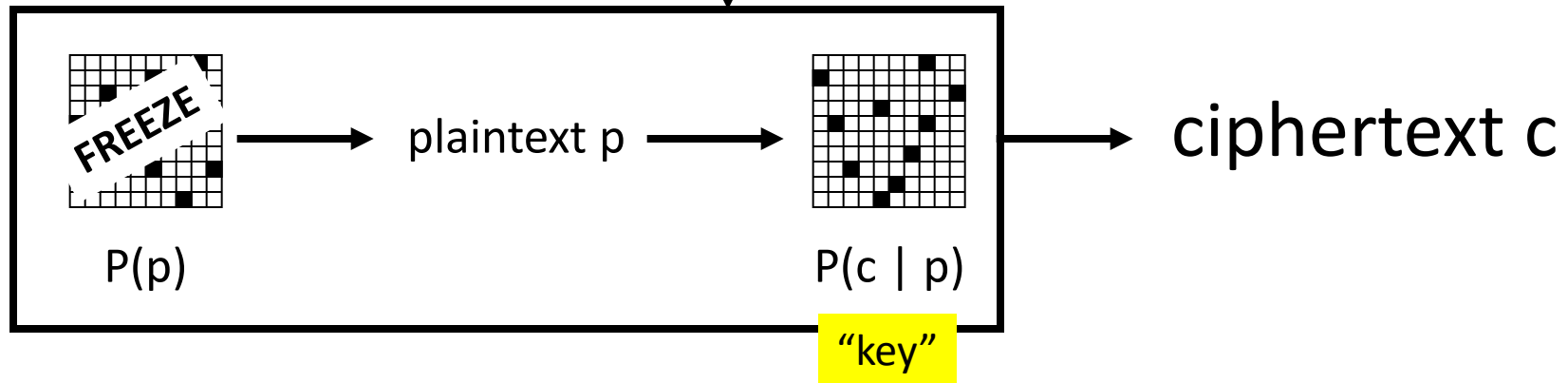


Letter Substitution Cipher

Find substitution-table values that maximize

$$\begin{aligned} P(c) &= \sum_p P(p, c) \\ &= \sum_p P(p) P(c \mid p) \end{aligned}$$

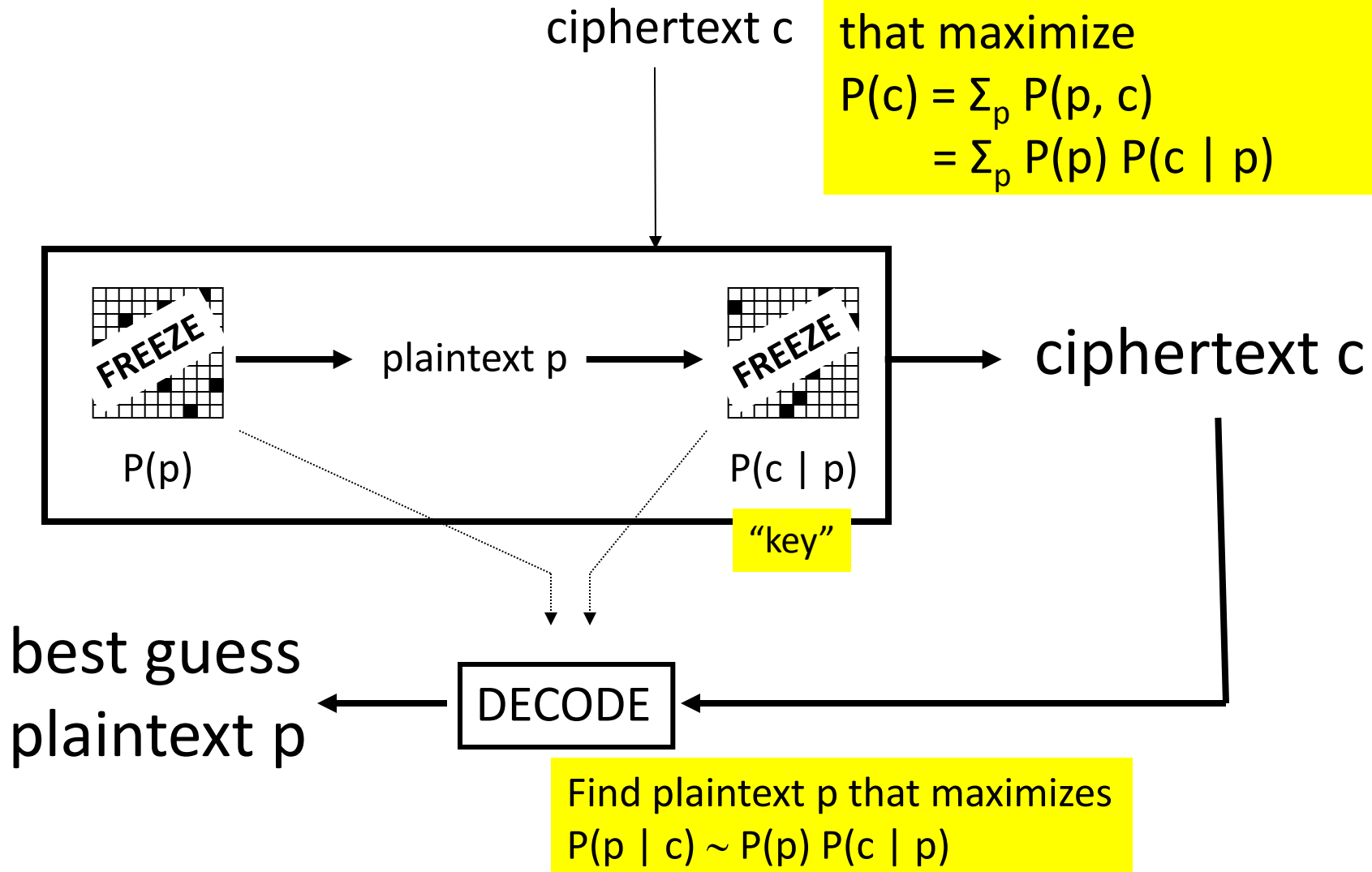
ciphertext c



Letter Substitution Cipher

Find substitution-table values that maximize

$$P(c) = \sum_p P(p, c) \\ = \sum_p P(p) P(c | p)$$



Letter Substitution Cipher

plaintext samples,
unrelated to ciphertext

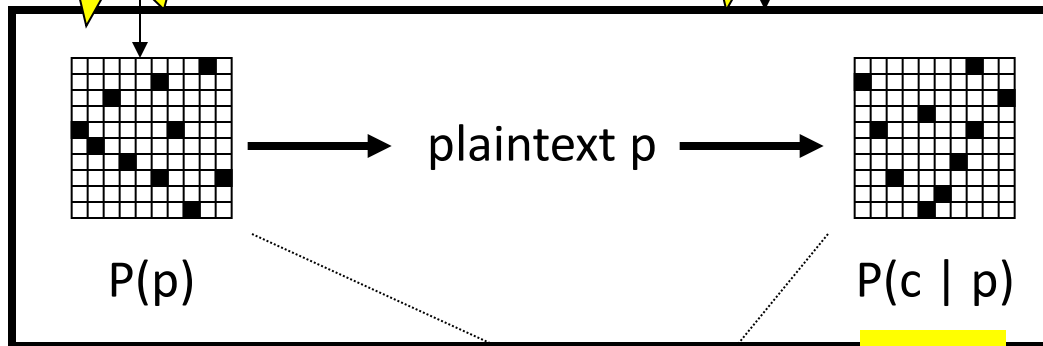
LM

ciphertext c

EM

Find substitution-table values
that maximize

$$P(c) = \sum_p P(p, c) \\ = \sum_p P(p) P(c | p)$$

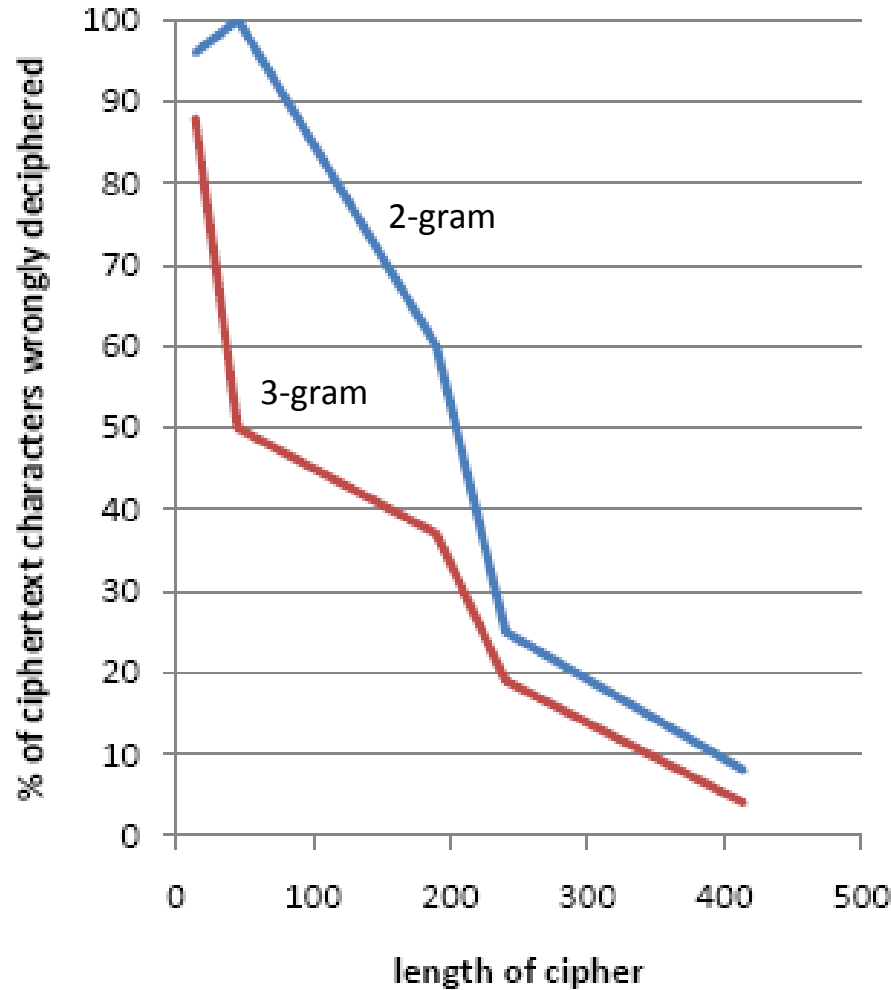


best guess
plaintext p

Viterbi

Find plaintext p that maximizes
 $P(p | c) \sim P(p) P(c | p)$

Decipherment Accuracy vs. Cipher Length



Letter Substitution Cipher

plaintext samples,
unrelated to ciphertext

LM

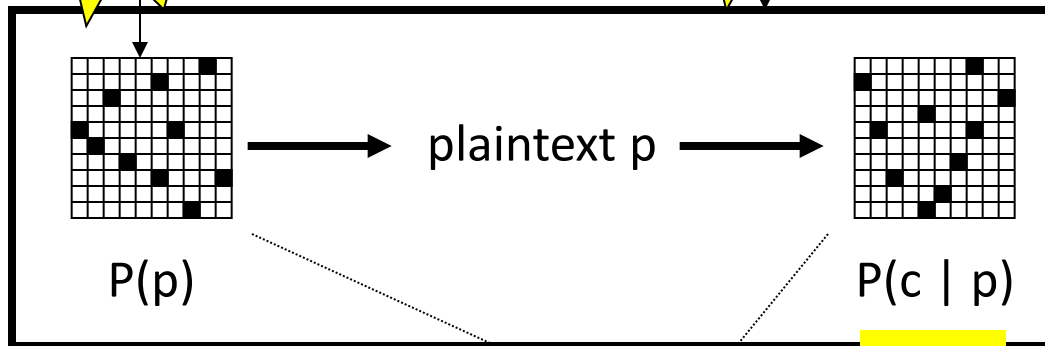
ciphertext c

EM

Find substitution-table values
that maximize

$$P(c) = \sum_p P(p, c) \\ = \sum_p P(p)^{0.5} P(c | p)$$

[Ravi & Knight 09b]



best guess
plaintext p

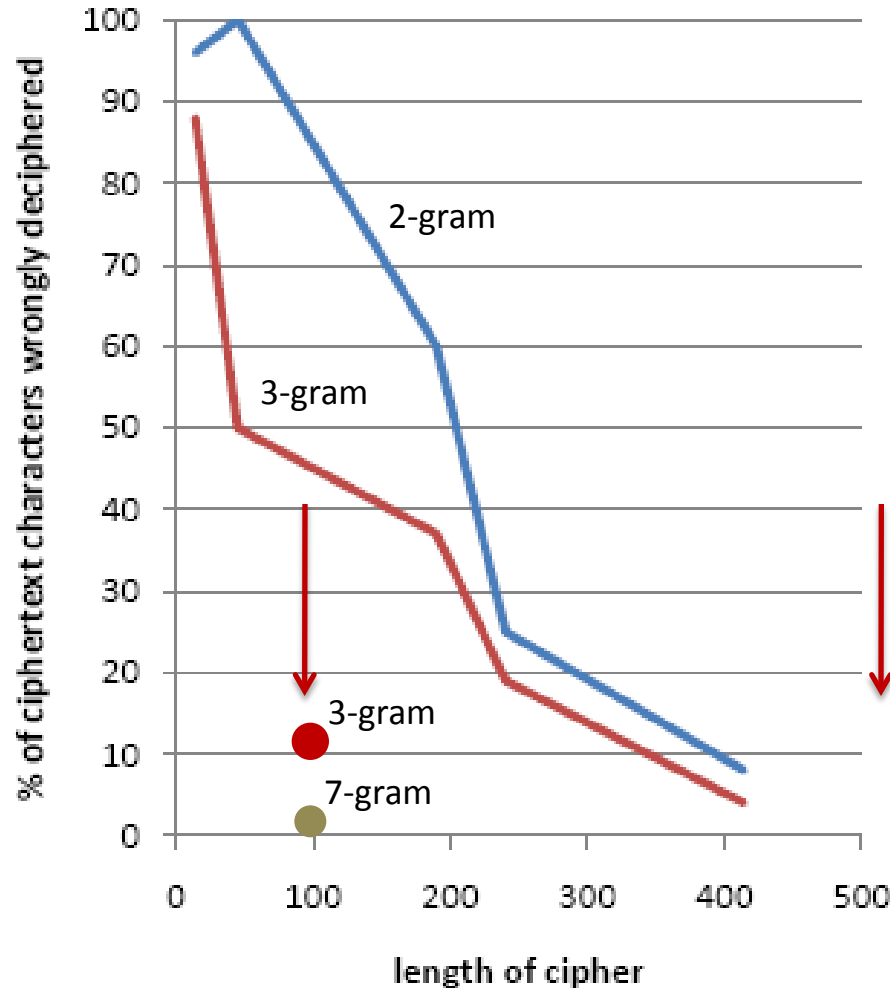
Viterbi

Find plaintext p that maximizes

$$P(p | c) \sim P(p) P(c | p)^3$$

[Knight/Yamada 99]

Reducing LM Weight During EM



Set EM to maximize

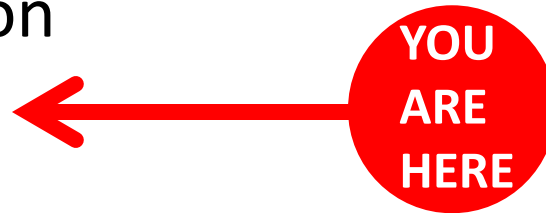
$$P(c) \approx \sum_p P(p)^{0.5} P(c | p)$$

instead of

$$P(c) \approx \sum_p P(p) P(c | p)$$

Plan for This Talk

- Break a series of codes
 - Simple letter substitution
 - Phonetic substitution
 - archaeology
 - transliteration
 - Word substitution
 - Foreign language as cipher
- Bonus
 - Two historical ciphers
 - Final thought on translation and cryptography



Phonetic Decipherment

ciphertext

**primera parte
del ingenioso
hidalgo don ...**

Phonetic Decipherment

ciphertext

**primera parte
del ingenioso
hidalgo don ...**

Phonetic Decipherment

“When I look at these squiggles, I say to myself, this is **really a sequence of Spanish phonemes**, but it has been encoded in some strange symbols...”



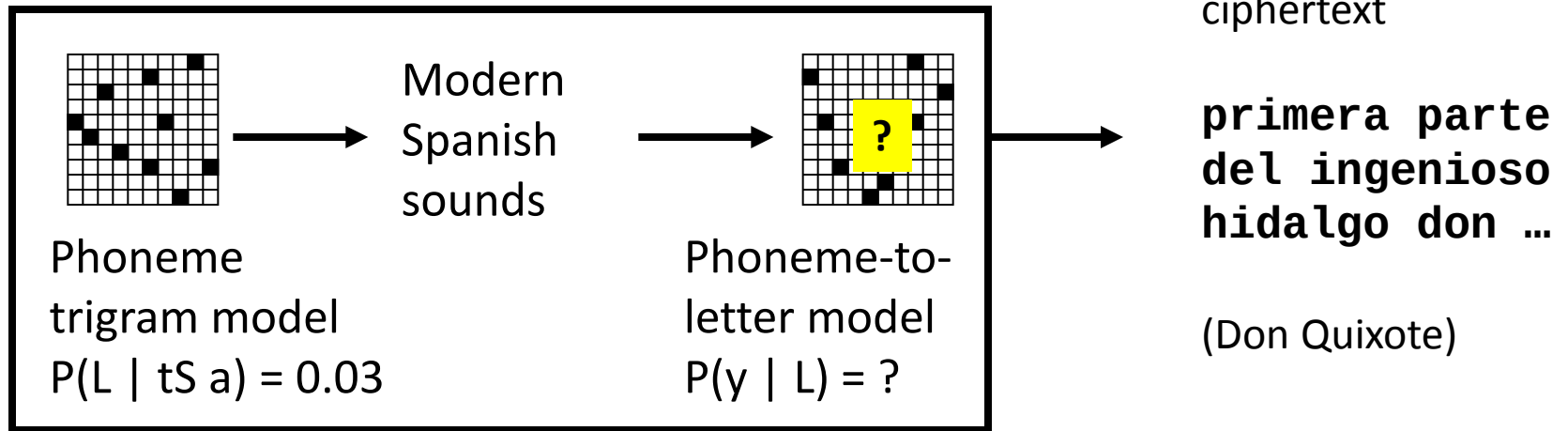
OUR HERO

ciphertext

**primera parte
del ingenioso
hidalgo don ...**

(Don Quixote)

Phonetic Decipherment



26 sounds:

B, D, G, J (canyon),
L (yarn), T (thin), a,
b, d, e, f, g, i, k, l,
m, n, o, p, r,
rr (trilled), s,
t, tS, u, x (hat)



32 letters:

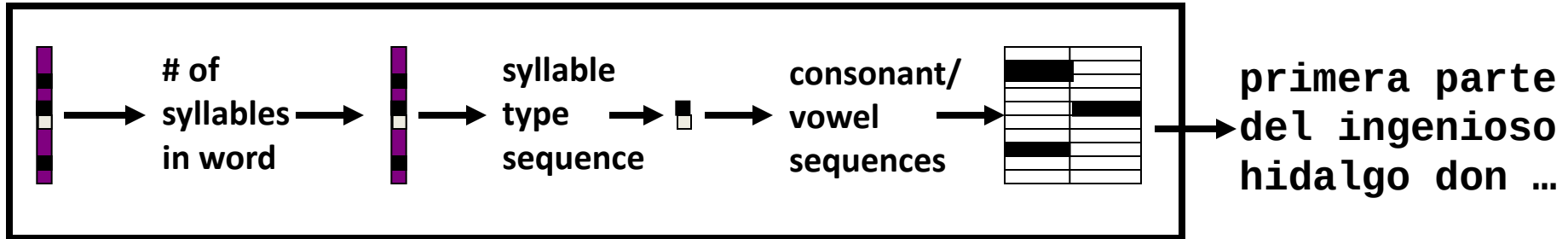
ñ, á, é, í, ó, ú,
a, b, c, d, e, f, g,
h, i, j, k, l, m, n,
o, p, q, r, s, t, u
v, w, x, y, z

EM approach = 93% accurate phonetic decipherment

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 $P(c \mid p)$ 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

Unknown Source Language



$P(1) = ?$
 $P(2) = ?$
 etc.

$P(CV) = ?$
 $P(V) = ?$
 $P(CVC) = ?$
 + 7 others

$P(V | V) = ?$
 $P(VV | V) = ?$

$P(a | V) = ?$
 $P(a | C) = ?$
 etc.

Input: **primera parte del ingenioso ...**
 Output: **NSV.NV.NV NVS.NV NVS VS.NV.SV.V.NV ...**

S = sonorous consonant phoneme

N = non-sonorous consonant phoneme

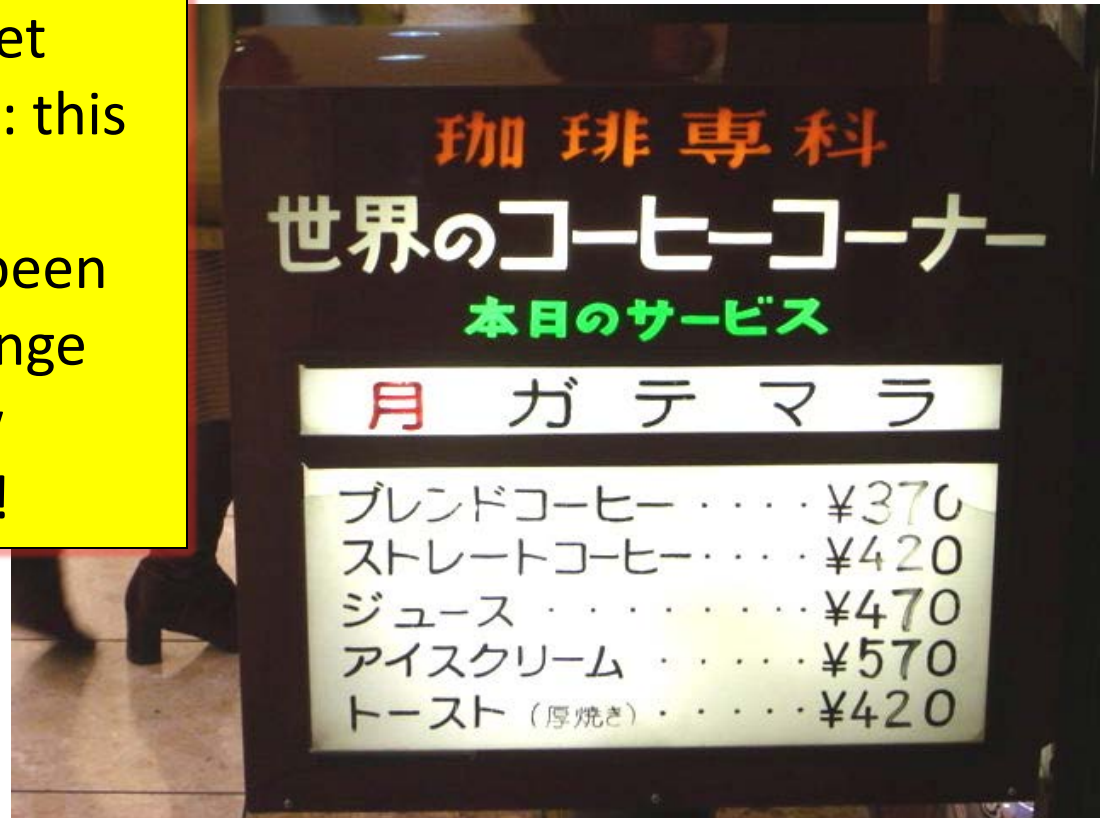
V = vowel phoneme

Phoneme Substitution Ciphers

When I look at street signs in Tokyo, I say: this is **really written in English**, but it has been coded in some strange symbols. I will now proceed to decode!



OUR HERO



Parallel data:

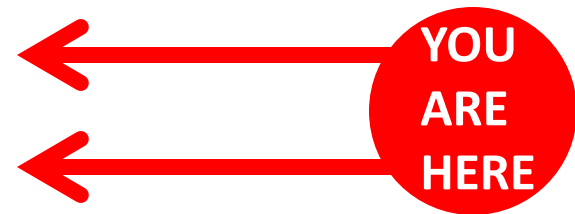
[Knight & Graehl 97]

Non-parallel data:

[Ravi & Knight 09a]

Plan for This Talk

- Break a series of codes
 - Simple letter substitution
 - Phonetic substitution
 - archaeology
 - transliteration
 - Word substitution
 - Foreign language as cipher
- Bonus
 - Two historical ciphers
 - Final thought on translation and cryptography

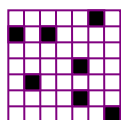


Word Substitution Cipher

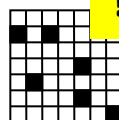
.....France.....Britain.....Canada...
Mexico.....Indonesia.....Malaysia...
Britain.....Canada.....Australia...
Britain.....France.....Indonesia.....
Mexico.....Australia.....France...
 ...Britain.....

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

TRAIN



plaintext e



?

$P(\text{sentence has } w1 \mid$
 $\text{sentence has } w2)$

$P(f \mid e) =$
 $7 \times 7 \text{ subst table}$

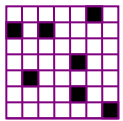
.....knd!.....bryT!ny!
knd!.....
 !lmksyk.....
 ...!ndwnysy!.....!lmksyk.....
bryT!ny!.....m!lyzy!...
bryT!ny!.....frns!.....
!str!ly!.....!ndwnysy!...
frns!.....frns!
frns!.....bryT!ny!.....
!str!ly!.....

Word Substitution Cipher

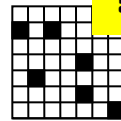
.....France.....Britain.....Canada...
Mexico.....Indonesia.....Malaysia...
Britain.....Canada.....Australia...
Britain.....France.....Indonesia.....
Mexico.....Australia.....France...
 ...Britain.....

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

TRAIN



plaintext e



?

$P(\text{sentence has } w1 \mid \text{sentence has } w2)$

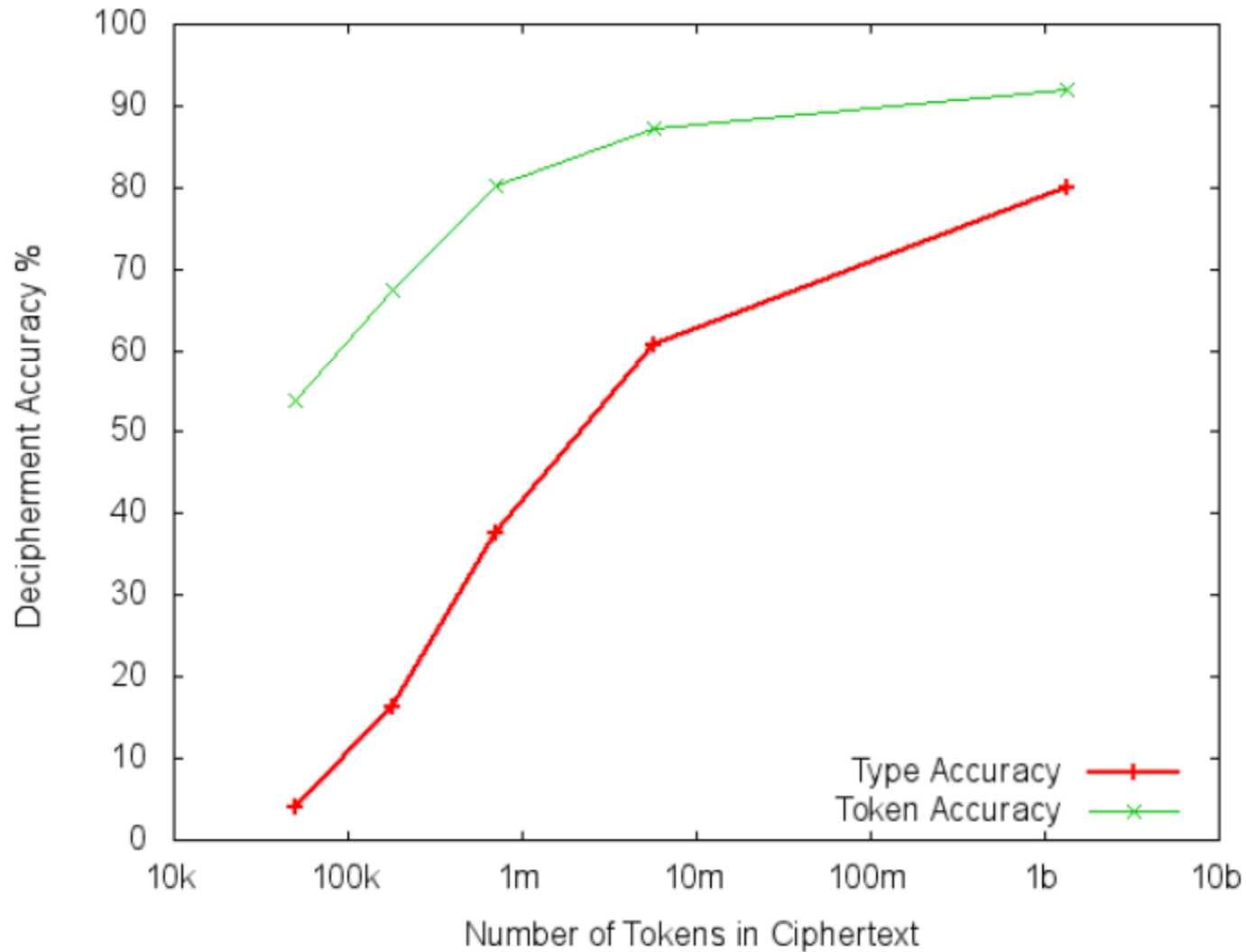
$P(f \mid e) =$
7 x 7 subst table

.....knd!.....bryT!ny!
knd!.....
 !lmksyk.....
 ...!ndwnysy!.....!lmksyk.....
bryT!ny!.....m!lyzy!...
bryT!ny!.....frns!.....
!str!ly!.....!ndwnysy!...
frns!.....frns!
frns!.....bryT!ny!.....
!str!ly!.....

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]

Word Substitution (Full Vocabulary, Giga-scale)



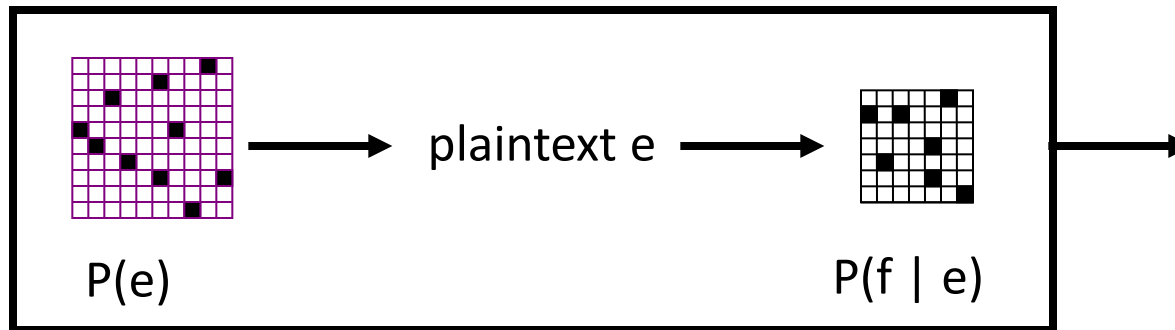
(Dou & Knight 2011)

Foreign Language as a Cipher

“When I look at **this giant corpus of Arabic**, I say to myself, this is really English, but it has been encoded in some strange symbols!!! Let's decode!!!”



OUR
HERO



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 نظّمته 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 الجانب

Time Expressions

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

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

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

Time Expressions

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

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

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

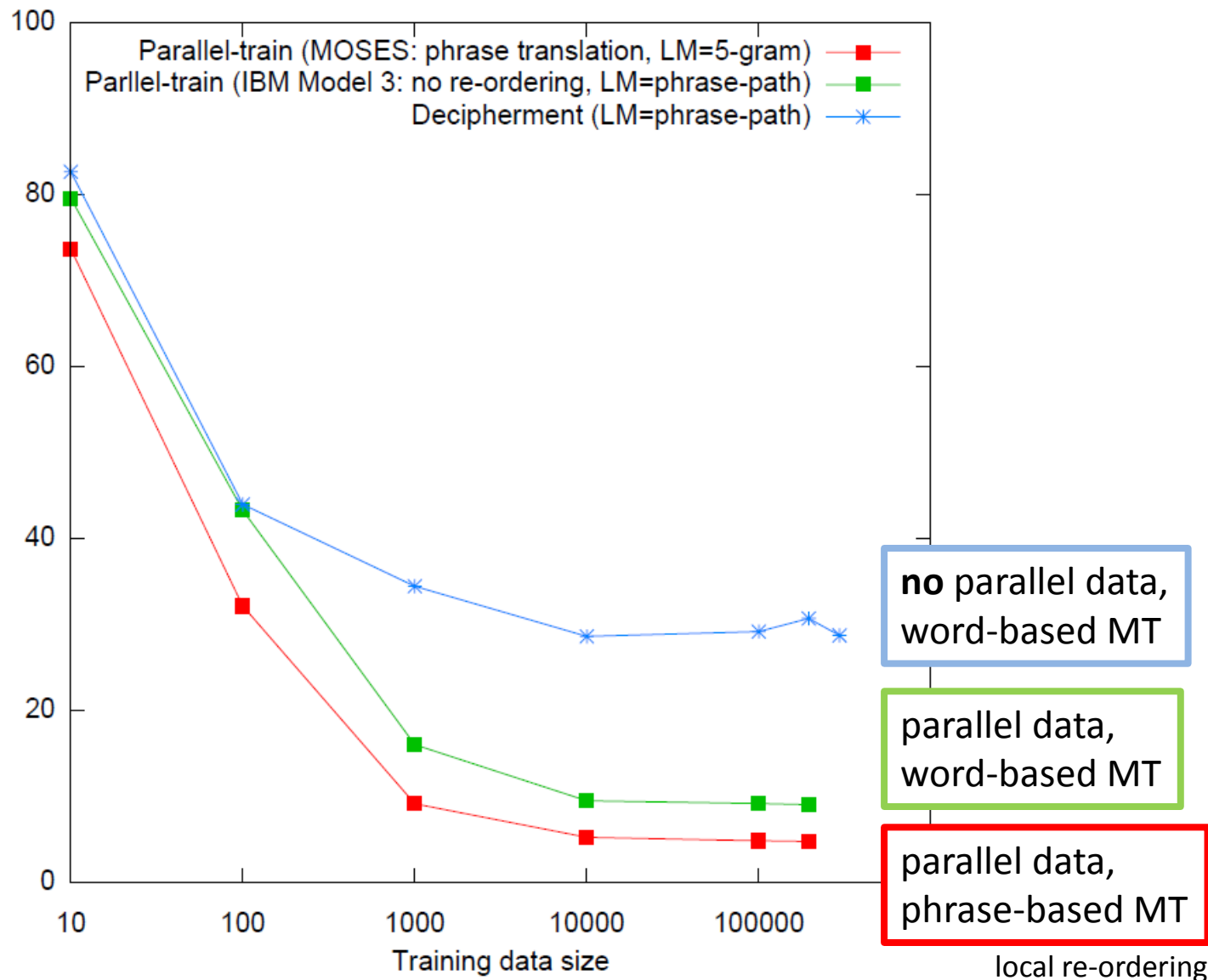
229 fy Hzyr!n !lm!Dy
207 fy Hzyr!n
75 fy Hzyr!n !lmqbl
61 fy Hzyr!n 1993
31 fy Hzyr!n 1992
27 !lr!b@ mn Hzyr!n
27 fy Hzyr!n 1967
19 fy 30 Hzyr!n !lm!Dy
18 fy n=h!y& Hzyr!n !lm!Dy
18 fy Hzyr!n 1991
17 mn Hzyr!n
17 mndh Hzyr!n !lm!Dy
17 4 Hzyr!n

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

Deciphering Spanish Time Expressions

MT quality
on test set

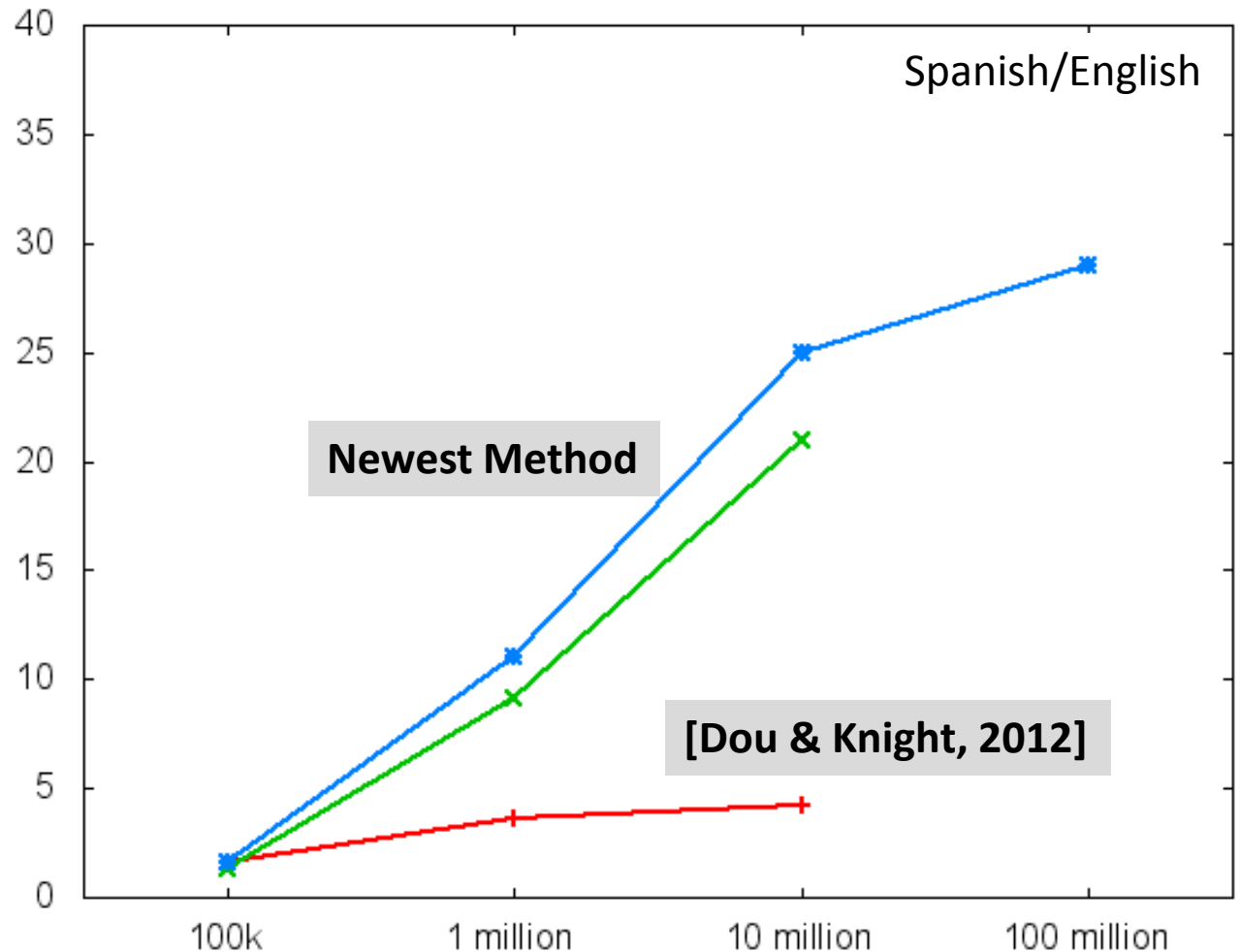
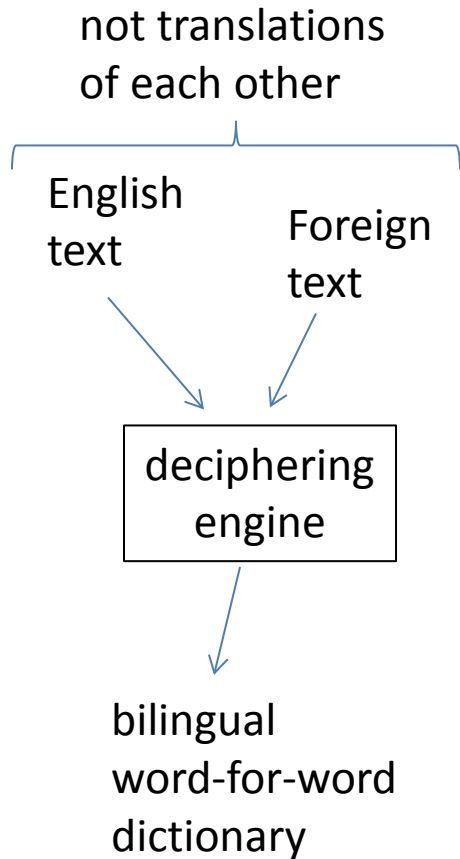
(Edit distance,
lower is better)



Exploiting Giga-Scale Non-Parallel Text

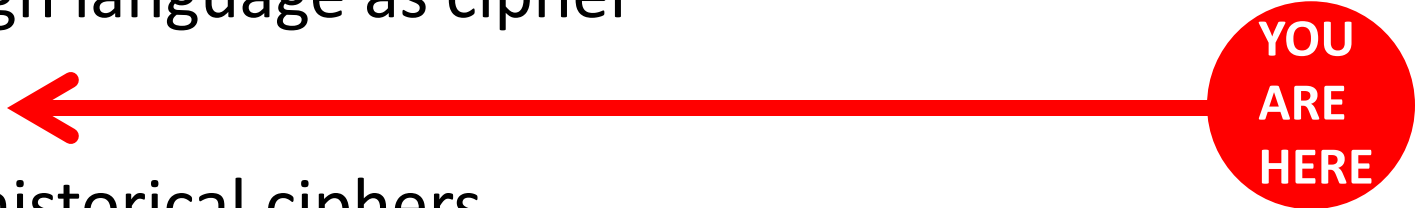
(Dou & Knight subm.)

Accuracy of learned
bilingual dictionary



How much foreign text (running words)

Plan for This Talk

- Break a series of codes
 - Simple letter substitution
 - Phonetic substitution
 - archaeology
 - transliteration
 - Word substitution
 - Foreign language as cipher
- Bonus 
 - Two historical ciphers
 - Final thought on translation and cryptography

Copiale Cipher

=mzfzhi|ny|cz|nem|c|iz|rbg|ra|c|zōy|ab+qanpāz
 h|c|j|y|an+qan|p|rg|p|z|y|dāy|a|g|a|f|+|am|z|c|ū|t|p|p|ri|d|y|a
 e. h=anū|d|c|p|z|p|g|ō|j|v|t|+|b|z|ō|s|ē|t|p|z|=rōm|z|p|b|l|ū|d|:q|p|o
 c|p|l|a|t|t|p|d|g|p|u|c|p|a|k|z|ē|l|u|o|d|p|:ab|z|n|p|u|b|z|p|:ō|l|b|y|t|p|l|g|p|p
 m|p|ō|r|u|d|ō|p|n|:s|ē|t|p|u|+|ū|j|d|+|n|t|c|u|g|=r|z|g|m|a|z|u|l|ā|y|g|p|w|+|q|z
 j|z|m|=m|p|a|c|t|t|c|t|p|z|j|=g|z|n|c|p|l|b|t|t|+|v|z|y|z|ē|b|g|t|p|m|z|j|g|z|+|t|ā|l
 h|=f|w|z|ē|w|l|b|c|t|p|u|d|t|t|p|b|t|t|z|=+|p|m|p|+|t|p|u|z|ō|l|z|n|p|j
 g|n|+|v|u|z|p|h|z|z|m|z|+|r|ē|+|a|r|ē|c|p|c|:p|t|p|q|p|p|p|a|o|l|p|u|+|v|z|ō|f|j|u
 p|u|l|h|s|=l|ō|t|ā|j|n|ē|p|u|b|f|t|p|v|p|c|ē|p|a|e|z|n|n|f|f|c|y|ē|l|w
 z|ū|j|ē|l|b|z|y|ā|ū|m|h|j|p|ā|l|o|=g|z|g|+|h|g|t|ū|t|l|p|l|p|b|t|t|z|l|ō|z|s|+|:p|z
 a|=+|g|z|p|t|p|d|p|z|m|p|t|t|c|y|l|v|z|ā|b|c|p|u|p|ā|l|p|z|f|u|o|p|t|l|w|ū|p|d|c|y|u
 d|z|ō|p|l|a|z|ū|v|p|ō|d|=l|p|z|l|u|v|z|j|d|p|l|v|h|g|z|p|+|b|d|t|p|l|p|ā|z
 m|ō|r|z|ō|ō|p|ō|m|a|m|z|l|ā|h|l|Δ|n|l|u|p|z|v|l|d|d|+|b|f|=t|c|u|p|t|p|g|z|p|ē
 h|=r|p|t|t|ā|b|t|p|r|:p|l|ō|l|u|z|p|p|ē|z|b|d|j|f|ū|t|z|ō|z|h|r|:l|p|ō|g|g|z|=u|w
 p|j|z|ā|g|a|=g|p|v|ū|z|c|p|z|h|c|:ū|r|g|h|p|m|ō|l|u|z|u|w|p|z|p|ū|t|z|=g|z
 g|p|ō|j|u|v|l|h|z|p|n|z|p|t|t|p|z|p|z|p|d|u|v|p|p|+|z|l|d|d|+|h|m|c|t|p|p|f|z|p
 d|p|ū|m|:ē|l|o|s|=l|j|y|r|d|u|e|m|p|r|:d|t|ā|z|g|p|p|m|p|c|:p|p|d|z|l|ō|z|d|c|:q

zāl

zāl|o|m|t|p|p|p|f|j|+|c|p|p|k|z|z|c|d|o|c|:j|z|m|z|p|ō|n|Δ|d|d|ē|f|c|d|:u|d
 t|y|ē|c|:ā|p|a|n|h|=p|r|ō|y|m|p|+|p|t|t|c|s|ā|p|a|n|z|t|t|h|c|l|u|b|.

Hpāo+dx|z|c|u|b|.

Δ|t|t|z|p|c|ā|p|a|m|p|y|ē|g|ō|+|c|p|p|ū|h|c|p|p|y|p|l|=m|b|z|z|t|z
 u|p|z|:z|z|z|h|p|b|d|ē|y|r|ē|g|c|ā|t|t|p|c|p|z|ō|b|s|=z|p|d|u|h|l|p|p|a|z|m|p|v|p|p
 w|p|ē|z|ū|c|p|p|ō|m|l|ū|d|:p|p|e|p|p|z|f|h|g|z|p|p|l|b|m|y|z|n|ū|p|u|d|d|u|p|d
 p|p|v|y|g|d|z|p|m|ā|z|p|ō|c|ā|p|y|ō|u|w|d|c|p|y|z|=p|z|u|p|u|v|t|=z|l|s|b|z|z
 r|p|y|ē|b|m|ō|l|:u|d|t|t|p|m|z|t|z|t|p|m|h|g|t|p|p|u|z|a|z|g|p|w|+|p|a|k|z|l|:z|u|l|p|z
 x|u|c|ū|p|z|l|p|t|t|p|ā|l|o|h|=p|b|=+|v|p|u|l|ū|d|:y|p|m|s|m|p|o|+|h|t|c|g|p|r
 z|v|b|z|ē|z|u|d|p|l|s|ū|m|t|t|ē|z|+|e|z|y|z|h|z|=m|h|t|t|ō|r|ō|ū|t|p|t|t|l|c|ū
 f|ō|b|m|ō|r|p|ē|o|+|d|c|z|t|z|c|y|p|z|ā|g|z|w|z|p|p|n|ū|z|h|z|ō|l|p|ō|r|ā|m
 p|ē|z|d|l|c|p|p|b|g|p|p|g|m|y|p|p|ā|z|t|t|c|ē|l|u|l|m|z|j|m|p|p|z|j|c|d|z|f|z
 c|z|p|ō|y|ō|r|u|z|u|l|Δ|t|p|=l|m|p|y|ā|b|d|z|j|p|m|d|z|r|m|ō|l|u|b|m|r|p|o|z
 z|ō|c|p|ā|p|m|p|c|:z|+|h|t|c|v|ē|y|r|z|ū|p|t|p|d|p|s|z|n|l|ā|o|m|d|d|g|p|u|z|z|+
 w|p|t|p|p|+|g|p|ē|=l|c|y|p|u|d|z|h|z|d|p|z|d|t|y|p|m|.

Apēo+b|Δ|n|.

Capvōc|ē|p|v|ō|p|r|ā|r|Δ|Δ|n|ā|y|x|u|l|c|p|p|b|g|=z|g|h|z|p|

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

Copiale Cipher

Section headers

Lines ≈
equal length

Paragraphs and section titles
always begin with
capitalized Roman letters.

Non-enciphered inscriptions:
Copiales 3 and Philipp 1866

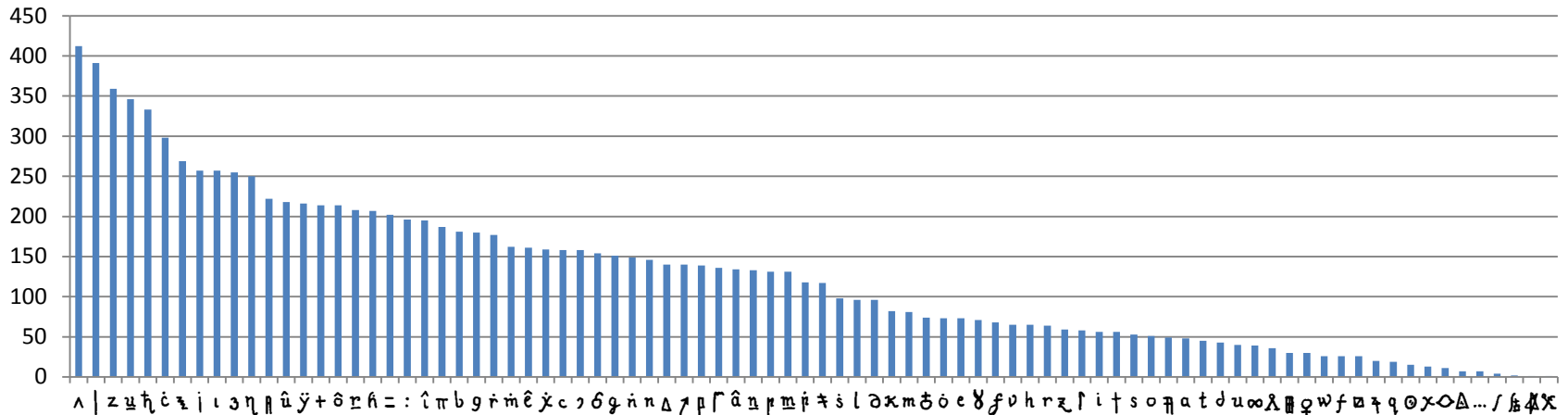
Some scratch-outs, rare

Preview text fragments
("catchwords")

hijciyānu+ηλπiγδρζυχδδγλαγαρf+iamz3cūh̄r̄p̄riδ̄yā
e. li=anūv̄δ̄c̄p̄z̄agōiv̄h̄t̄+bz̄iōs̄ēh̄r̄m̄=rōm̄z̄||b̄lūd̄:ηr̄|σ
c̄p̄||λr̄h̄n̄dḡp̄uc̄|ηr̄κλz̄ēλuōd̄||:λb̄x̄n̄πiōb̄z̄||:ōl̄b̄ȳh̄m̄lḡr̄p̄
m̄p̄ōr̄uōd̄h̄n̄:s̄ēh̄r̄ūi+ūj̄v̄d̄+ñt̄c̄ūḡ=r̄z̄gm̄Δz̄λūl̄āȳḡ||w̄+q̄x̄
||m̄=mp̄ac̄īh̄c̄īr̄x̄||īgz̄nc̄p̄||λb̄h̄t̄+v̄zȳ||ēb̄gn̄r̄m̄z̄īḡz̄ūh̄āl
h̄i=īw̄z̄ē|w̄.λbc̄h̄m̄ūd̄h̄t̄h̄n̄πb̄h̄r̄=+p̄m̄r̄p̄+h̄r̄ūi|ōl̄zn̄||n̄q̄j̄
gn̄+v̄uz̄||h̄z̄īm̄z̄r̄ē+Δr̄īēc̄p̄ic̄:|r̄r̄n̄q̄r̄s̄m̄p̄āōl̄p̄ū+v̄z̄ōb̄īu
πū.λh̄s̄=|l̄ōīr̄āj̄n̄r̄ēr̄m̄ub̄|f̄h̄ūv̄r̄c̄ēīp̄l̄ez̄ñn̄:f̄f̄c̄ȳē||λw̄
z̄īj̄h̄.λbz̄ȳā|ūm̄h̄īp̄āīo=gz̄ḡ+līgt̄īūh̄l̄h̄r̄||b̄ōh̄z̄īl̄ōx̄s̄ī+||z̄
a=+gz̄z̄h̄m̄ōr̄||x̄m̄p̄h̄ēȳl̄v̄z̄ā|bc̄īm̄p̄āl̄||q̄z̄uōp̄ī|w̄ūr̄ōb̄ȳ|u
d̄x̄z̄ō||λaz̄ū|v̄p̄ōd̄=λh̄z̄λūv̄x̄j̄Δ.λv̄h̄gr̄z̄||b̄Δr̄λn̄Δp̄āz̄
m̄ōr̄īo|ōp̄m̄Δm̄z̄h̄āh̄Δn̄īūp̄z̄|v̄.λd̄Δ+b̄f̄=|h̄c̄ūp̄īr̄gz̄h̄ē
h̄i=|r̄r̄īt̄+āb̄ōr̄r̄:|r̄λōλū|ηh̄ēz̄b̄ōΔj̄|ūh̄īz̄oz̄h̄r̄:|p̄ō|gḡz̄=ūw̄
p̄īz̄z̄ōr̄ga=gp̄v̄ñz̄c̄h̄r̄z̄h̄c̄:ūr̄gh̄r̄m̄ō|ūz̄ūw̄p̄z̄p̄p̄ūr̄l̄=gz̄
gp̄o|ūv̄n̄~~h̄īt̄~~||ñz̄x̄h̄h̄z̄h̄ōv̄n̄p̄r̄ī+||a|Δd̄+h̄im̄c̄f̄m̄p̄īz̄n̄
d̄h̄ūm̄:ēl̄ōs̄=λīj̄r̄v̄ūem̄r̄r̄:dh̄īāz̄ḡq̄r̄m̄r̄ic̄:||v̄d̄Δl̄ōΔc̄:η
z̄āλ

z̄āλom̄r̄m̄l̄f̄j̄+ēh̄r̄κx̄r̄z̄x̄d̄ōc̄:ix̄m̄z̄h̄ōn̄Δd̄ōēf̄c̄a|:ūa
d̄ȳēc̄:āīr̄an̄h̄ī=r̄ōd̄ȳm̄||b̄+ñt̄c̄s̄āh̄an̄x̄h̄h̄c̄l̄ūb̄.
H̄p̄āōo+dx̄ī|ic̄:ūb̄.
Δh̄īx̄p̄c̄āh̄r̄am̄p̄ȳērḡō+ēh̄r̄ūh̄ēh̄r̄īȳr̄λ=mb̄bz̄īst̄z̄
ūr̄x̄:z̄z̄x̄z̄h̄||b̄d̄ēȳr̄ēgc̄āh̄r̄ic̄r̄x̄ōb̄s̄=χ̄|ōūh̄|r̄p̄ax̄=mp̄v̄p̄r̄
w̄p̄ēz̄|ūc̄p̄h̄ōēm̄l̄īd̄:ηr̄ep̄r̄z̄f̄h̄iḡz̄īp̄r̄l̄bm̄ȳz̄n̄ūp̄ūd̄|Δūp̄ī
h̄v̄ȳḡd̄|ηm̄āx̄p̄ōc̄īāp̄ȳōūw̄d̄c̄p̄ȳz̄=ηx̄ūp̄ūv̄|=z̄λs̄b̄x̄z̄
r̄p̄ȳēb̄m̄ōī:ūōh̄n̄m̄z̄f̄x̄h̄m̄h̄ḡh̄īp̄ūz̄ax̄gp̄w̄+ηλx̄Δf̄:z̄r̄ūl̄p̄ī
x̄ūc̄ūm̄x̄ī||h̄r̄āλoh̄=rb̄=+v̄p̄ūl̄īd̄:ȳr̄m̄s̄m̄p̄o+h̄h̄c̄gn̄r̄
z̄r̄b̄z̄ēx̄ū|d̄h̄λs̄ūm̄h̄ēx̄+ēz̄ȳz̄h̄z̄m̄h̄t̄+ōr̄ōūh̄r̄h̄īl̄c̄ū
f̄l̄ōbm̄ōr̄p̄ēō+dc̄īh̄z̄c̄ȳr̄x̄īaḡx̄w̄z̄r̄||n̄ūz̄b̄z̄īl̄m̄āōn̄
p̄ēz̄x̄d̄c̄īr̄λb̄ḡh̄r̄m̄ḡm̄ȳp̄āz̄h̄h̄c̄ēλū|Δm̄z̄
cz̄p̄d̄ȳōr̄ūz̄ū.Δt̄r̄ī=|m̄p̄ȳāb̄ōΔz̄ȳj̄m̄ōz̄ī
x̄l̄ōc̄r̄āj̄m̄r̄ic̄:z̄+h̄h̄c̄v̄ēȳr̄x̄īp̄īn̄d̄p̄s̄x̄n̄|
w̄īh̄r̄||ḡp̄ē=λc̄ȳr̄ūd̄|z̄ōr̄z̄|d̄r̄ȳr̄m̄.
H̄p̄ēōo+b̄.Δn̄.
C̄ap̄v̄d̄c̄ēīr̄v̄ōh̄r̄ār̄.ΔΔn̄āȳx̄ūλc̄h̄r̄p̄bḡ=z̄gh̄x̄|

Letter Frequencies



digraphs:

ɔ h 99

ç : 66

h ^ 49

: u 48

z ʀ 44

trigraphs:

ɔ h ^ 47

ç : u 23

η ɔ h 22

ÿ ɔ h 18

h ç | 17

tendencies:

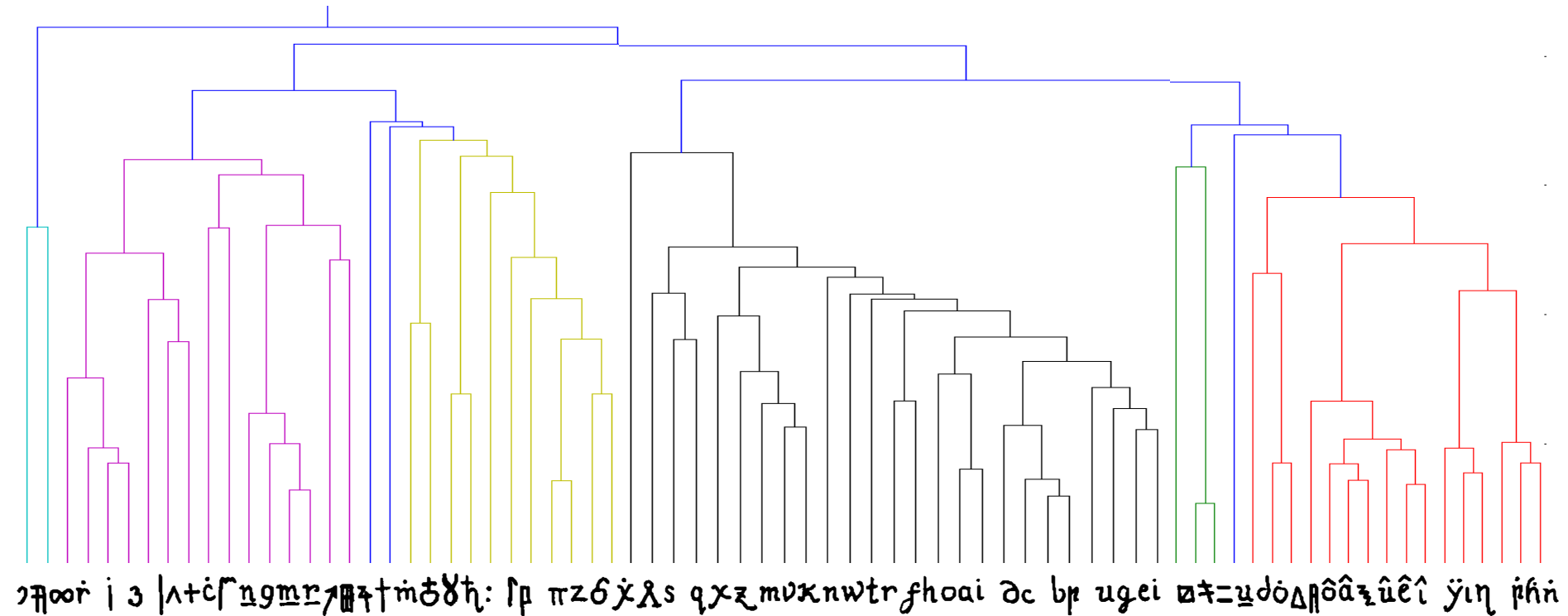
â, ê, î, ô, û followed by ɔ and j

â, ê, î, ô, û preceded by z and π

Clustering of Cipher Letters

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

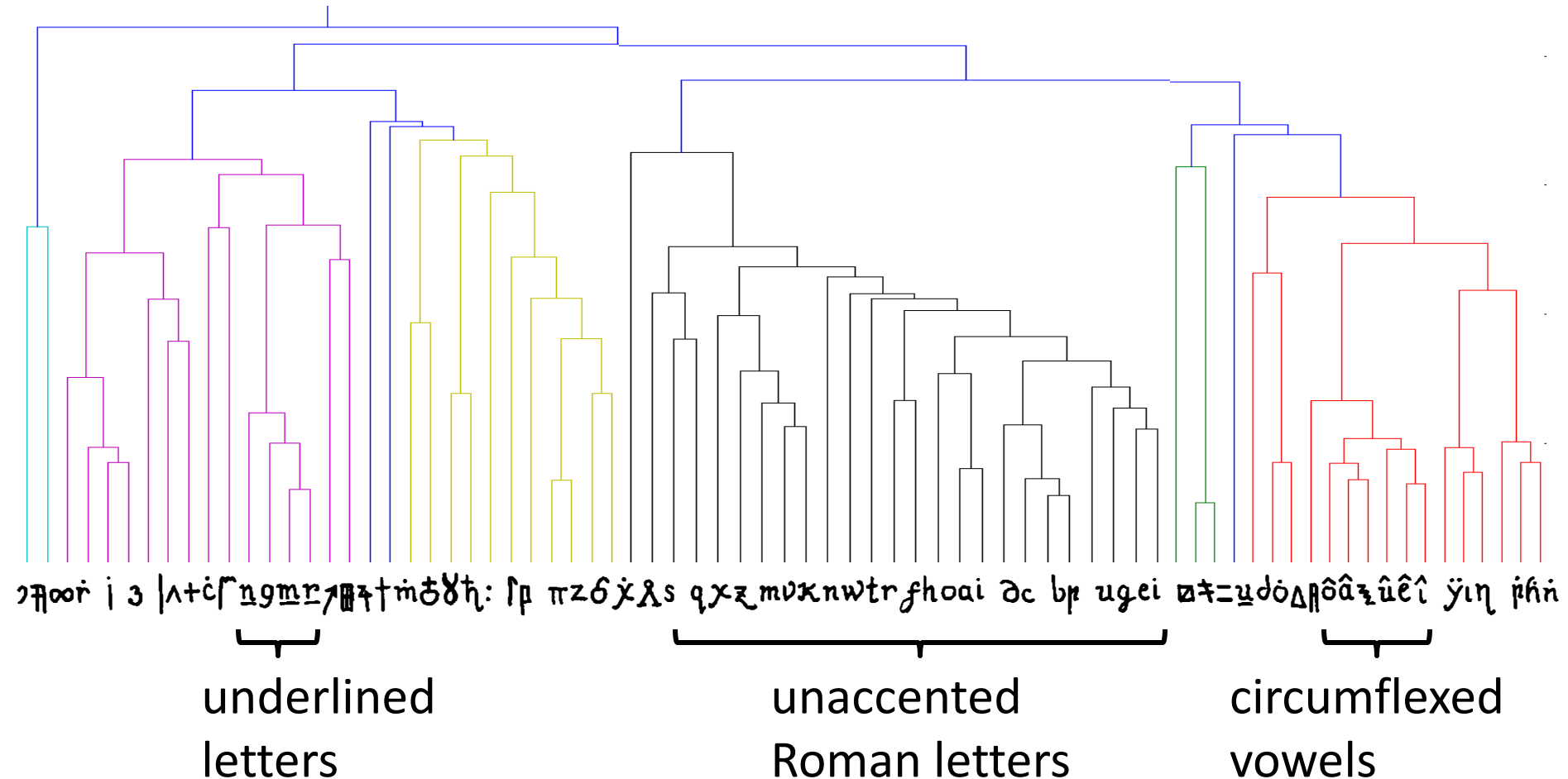
Scipy software



Clustering of Cipher Letters

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

Scipy software



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

κ m û r : p z i ô f | y ʔ h ê j t z i λ n π â z b Δ g z =
i j l z u p q ç λ â r g κ λ t p r t p λ z i n p î | Δ r δ p λ a
= g z w π y ê c Δ r δ Δ + b z η r i x y j î r z u f λ z
π x j ʔ p r = f | û λ s x m δ m | z η g â | κ t = λ t | l x ô
ø f : r î λ b i f u m y j z v z â j x ʔ r m π i z h λ c δ ô g g
z û + p x p n x ê z g h λ t p i t p λ î z t n x | r ô y m â
+ h h r z ô z n b s η + : z r κ r p δ p t Δ c û λ g = n z κ
p x ∞ n z p z f t n r π z ê y n g = r π g δ Δ z n z p | κ
π η j i x y r î g π u λ b g λ i t b δ û f p t z ô λ e z η ô
| û r c f z q δ λ b η t m θ

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

Decipher against
80 plaintext 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

κ m û r : p z i ô f | y o h ê i h z i λ n π â z b Δ g z =
i j l z u p q c l â r g κ λ h p r h p λ z i n p î | Δ r δ p λ a
= g z w π y ê c Δ r δ Δ + b z η r i x y i î r z u f λ z
π x i j d r = f | û λ s x m δ m | z η g â | κ h = λ h | l x ô
ø f : r î λ b i f u m y j z v z â i x o r m π i z h λ c δ o g g
z û + p x p n x ê z g h λ h p i h p λ î z t n x | r ô y m â
+ h h r z ô z n b s η + : z r κ r p δ p h Δ c û λ g = n z κ
p x ∞ n z p z f h n r π z ê y n g = r π g δ Δ z n z p | κ
π η i x y r î g π u λ b g λ i t b δ û f p h z ô λ e z η ô
| û r c f z q δ λ b η h m d

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

D, 80, **FAIL** nst
... languages.

Second Decipherment Approach

Homophonic cipher,
e.g.:

A = ʒ j l y r

B = û

C = ô ñ

D = ʈ

E = ʁ ʃ Δ * f î ɜ ɔ

F = ʀ

G = ʏ

etc.



κmûr:ρziô ʃ|ÿʔhêjth ɜιλnπâɜbΔgz=
jʀlzurɔçlâŕgκλh ʈŕh ʈλɜjnrî|Δŕδʀλa
=gzwπÿêcΔrδΔ+ɜzŋŕi ʁÿjîrzu ʃλɜ
π*|ʀðŕ=ʃ|ûλs*ɜδm|ɜŋ gâ|κh=λh|l ʁô
øʃ:ŕîλbιʃumÿjzυzâjxʔŕɜπizhλcδôgg
zû+ŕʔʈŋ ʁêɜghλh ʈjth ʈλîɜtn*|rôÿɜâ
+hŋŕzôɜɜbŝŋ+:ɜŕκŕδʀhΔcûλg=ɜzκ
ρɜooŋzʀɜʃhŋŕπɜêÿŋg=ŕπgδΔɜŋzʀ|κ
πŋjι ʁÿŕîgπuλbɜλι†bδûʃ ʈhɜôλɜzŋô
|ûŕcʃɜɔδλbŋh ɜɜð

Homophonic Cipher

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

FAIL

But, slight numerical preference for
German

Cipher Characteristics

digraphs:

o t 99

c : 66

t ^ 49

: u 48

z R 44

trigraphs:

o t ^ 47

c : u 23

n o t 22

y o t 18

k c | 17

tendencies:

â, ê, î, ô, û followed by z and j

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 ʰ

In German, C almost always followed by H
(German CH is like English QU)

So guess: ʔ = C, ʰ = H

One Thing Leads to Another

$\text{ʁ}^{\text{h}} = \text{CH} \rightarrow \text{ʁ}^{\text{h}}\text{ʌ} = \text{CHT} \rightarrow \text{ʌ} = \text{T} ?$

Each step is guesswork.

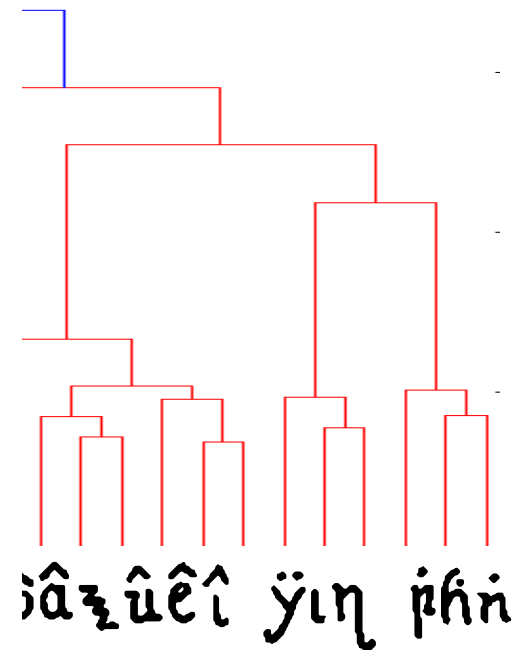
Must be willing to retract.

Weird task, not knowing German.

No longer care what the book says.

Cluster diagram crucial:

$\text{ȳ} = \text{ɪ} \rightarrow \text{ɪ} = \text{ɪ} , \eta = \text{ɪ}$



Spring Break 2011

c aeiou fpy dlmrztbv w hknsgs

* der	✓	cht	7h	Λ
und	✓	e	ē	: u
ein	✓	s	hē	l
und		ich	g	7h
cht	✓	ich	u	7h
ich	✓	che	7h	u
sch	✓	t	Λ	m Δ
che	✓	i	ē	l
ech		it	+ g	Λ
* die	✓	sa	+ h	Λ
rec		ds	7h	u
ne		elt	+ l	Λ
gen		t	ē	l n
eit	✓	t	l	: u
ver		ti	l	7h
hen		ter	= h	3
lic		der	u	h
ten		ew	u	h
rei			u	ē
nfe		sch	h	7h
auf		st	m	h
ede		ich	l	7h
and	✓	ein	h	7h
den		t	g	h
run		die	z	h
ter		die	z	h
sei		e	u	h
hte		e	m	: l n
hei		h	m	: l n
ns		und	=	h
ens		z	m	Δ
men		i	h	l
h		let	Λ	h
ere		h	7h	z
* das		g	l	h
nte		ste	h	Λ
nge		s	h	ē
te		sch	h	7h
ere		che	ē	h
ede		t	l	Δ

follow
 e, 't →

9n
 7
 10
 11
 12
 13
 14
 15
 16
 17
 18
 19
 20
 21
 22
 23
 24
 25
 26
 27
 28
 29
 30
 31
 32
 33
 34
 35
 36
 37
 38
 39
 40
 41
 42
 43
 44
 45
 46
 47
 48
 49
 50
 51
 52
 53
 54
 55
 56
 57
 58
 59
 60
 61
 62
 63
 64
 65
 66
 67
 68
 69
 70
 71
 72
 73
 74
 75
 76
 77
 78
 79
 80
 81
 82
 83
 84
 85
 86
 87
 88
 89
 90
 91
 92
 93
 94
 95
 96
 97
 98
 99
 100

vowels: $\underline{u} \hat{?} \underline{\hat{e}} \underline{a} \eta \underline{i} \underline{y}$
 $\Delta \underline{o} \underline{p} \underline{n} \underline{h}$

gives $\{nr\}$
u $\{cb\}$ a pig

need: f g y l m z
(more in
german)
ut sp
af pr
rf ee
(ff)

1 Δ? e
 ao i j
 v w k g
 Afford ke ge
 37 wual ku or
 always
 st
 ga
 ng

$$\begin{array}{c} \boxed{u} \\ a \boxed{u} f \\ u \quad u \\ u \quad r \\ \textcircled{r} u \end{array} \quad \begin{array}{c} \{u\} \circ \{u\} \\ \circ \\ \circ \\ \circ \\ \circ \end{array}$$

$$3j \quad ? \quad ? \quad 3j \quad r$$

Spring Break 2011

German letters

Cipher letters, in groups

Grid

vowels: u ʔ ɛ ɪ ɨ ɜ ɹ ɻ

$$\Delta \hat{O} \hat{P} \hat{n} \hat{h}$$
$$\rightarrow \oint \vec{m} \cdot d\vec{r} = \neq$$

need: f g y l m z

(rare in German)
uf sp
af pr
rf pe
ff

aoi
vw k g

1. $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$
 2. $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$
 3. $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$
 4. $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$
 5. $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$
 6. $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$
 7. $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$
 8. $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$
 9. $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$
 10. $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$

z	u	0	0
a	u	0	0
	u	0	0
	u	0	0
	u	0	0

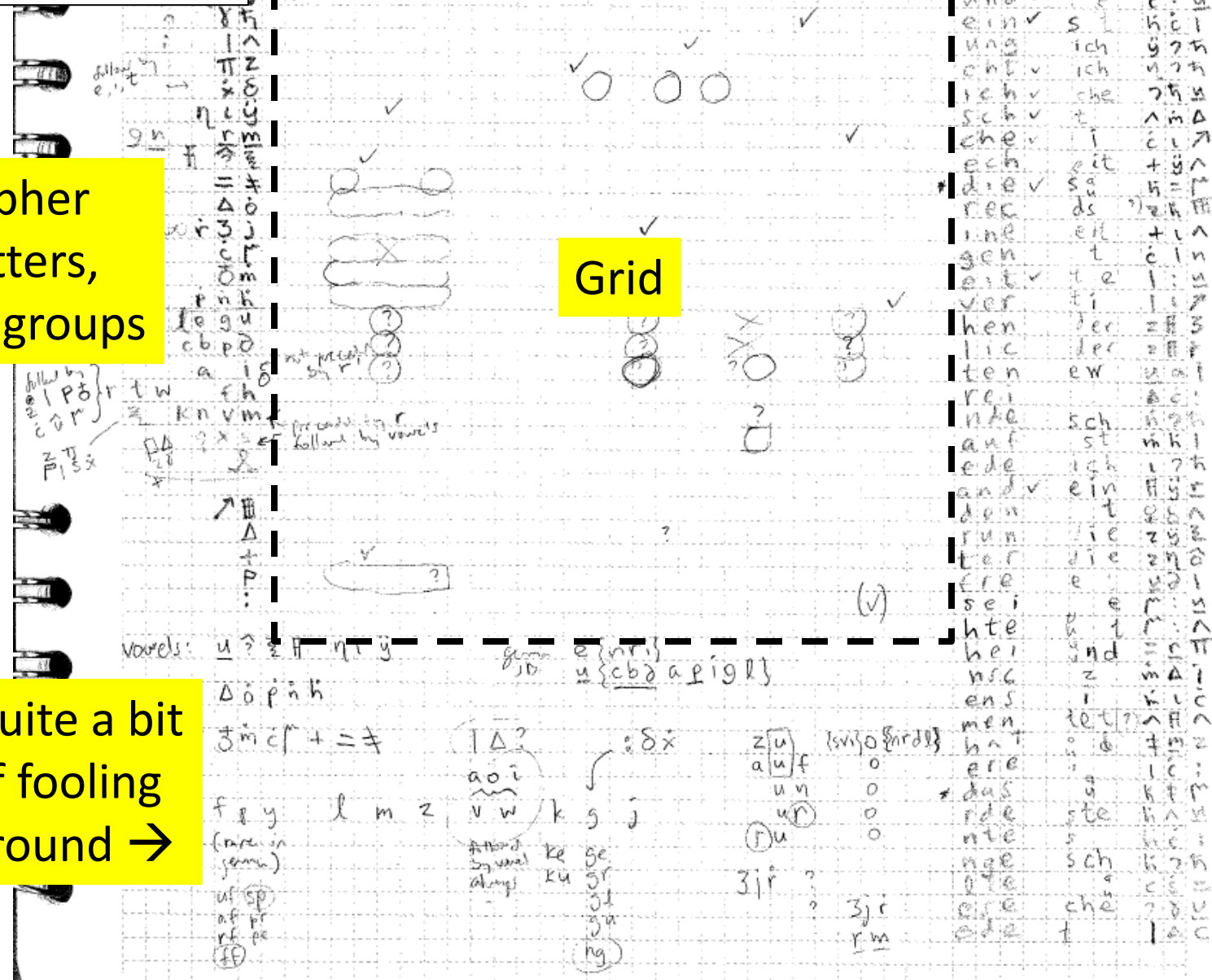
$$\begin{array}{ccc} 31r & ? & \\ & ? & 3j r \\ & & r m \end{array}$$

Spring Break 2011

German letters

Cipher letters, in groups

Quite a bit
of fooling
around →



Spring Break 2011

German letters

German trigraphs

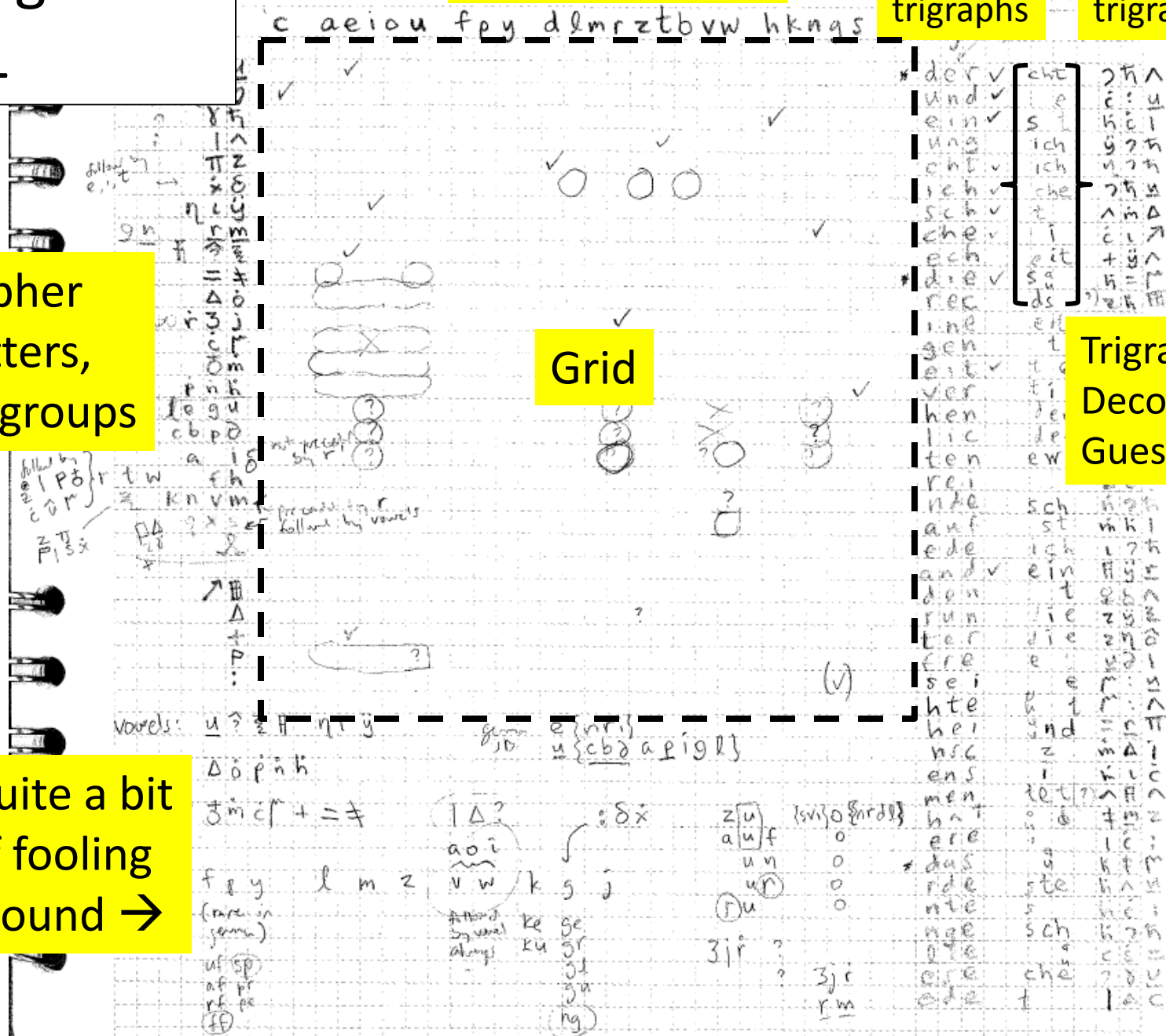
Cipher trigraphs

Cipher letters, in groups

Grid

Trigraph Decoding Guesses

Quite a bit
of fooling
around →



Key Observation #2

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

κ m û r: p z i ô f | y ʔ h ê j h z i λ n π â z b Δ g z =
i ʔ l z u p q ç λ â r g κ λ h ʔ r h ʔ λ z i n p î | Δ r δ ʔ λ a
= g z w π y ê c Δ r ð Δ + b z η r i x y j î r z u f λ z
π x i ʔ ð r = ʔ | û λ s x m δ m | z η g â | κ h = λ h | l x ô
ø ʔ : r î λ b i f u m y j z v z â i x ʔ r m π i z h λ c ð ô g g
z û + r x ʔ n x ê z g h λ h ʔ i h ʔ λ î z t n x | r ô y m â
+ h h r z ô z n b s η + : z r κ r p δ ʔ h Δ c û λ g = n z κ
p z ∞ n z ʔ z f h n r π z ê y n g = r π g ð Δ z n z ʔ | κ
π η j i x y r î g π u λ b g λ i t b δ û ʔ ʔ h z ô λ e z η ô
| û r c ʔ z q δ λ b η h m ð

Actually, those are space bars

Copiale Decipherment

lit: m z a l l b l
 v x z i l s x p x j w n
 π ο ι ρ η Δ ρ ε ζ ε α = ρ λ ũ b ◊ u r ⊗ z

δ ρ η ζ ι + ο ι η ρ λ η ι η ε ρ .

c ũ j f e z t i p t g η λ : x

ο χ ũ η ε η + ε ρ ρ x m λ ι z : γ ρ λ ε ο ι ρ z ι i x a l e c : ũ ο .

f ũ r f o j x λ ι λ = ε z .

m o η r ā + Δ g γ ũ x z o z m n x f n h η + i f .

x m ũ r : p z i o f j γ o η e i η ζ ι λ n π ā z b Δ g z = j ρ l z ũ p o c λ ā r g x λ η

η ρ η η λ ε i η ρ i l Δ r δ ρ λ a = g z w π γ ē c Δ r o δ a + b z η ρ i x γ i η z ũ f λ z

π x j o d i c f ũ λ s x m o m i η g ā i x η = a η i l x o m f : r i l b i f u m γ i z o

z ā i x o i m π i z h λ c o o g g z ũ + p x η n x ē z g h λ η η i η η λ i z t n x i r o γ

m ā + h h r z o z n b s i t + : ε ρ x ρ ρ o η Δ c ũ λ g = n z x p x o o n z ρ z f η n r π

z ē γ n g = r π g o δ a n z a l x π η i x γ r i g π ũ Λ b g λ i t b o ũ f η η z o λ e

z η o i ũ r c f z o f λ b η η m o :

n i r i c i ρ e o p h o r o p x o ũ η z ā x ◊ g s = o m e j z ũ i

g s m ũ o o l o n | o π e g ũ o o x r Δ j z g x r x ũ g π ũ z x ⊗ n i η ρ b = n

λ ē r m o z f : ũ a = r z l Δ η r ρ o m η ā z o i j o i n f i o δ o γ λ o z i e g c ũ η

r s ē η λ

c i r = f i ā η ũ b m Δ c : x o .

h z i η λ : o g i e z η ā z Δ n π ā z ⊗ p s x r o i z t m ā γ o ũ i = n π

z r s = n h x f r z p i n o x i p u g π c γ o η i o b i l x n g η x γ t i o c s = b + n γ ũ o

x o i r : ũ λ o i o o n .

l π o z f o h g z γ p i r l n m λ m Δ r λ o e x n r .

gesetz buchs

der hocheleuchte ◊ e ⊗

geheimer theil.

erster abschnitt

geheimer unterricht vor die gesellen.

erster titul.

ceremonien der aufnahme.

wenn die sicherheit der Δ durch den ältern

thürheter besorget und die Δ vom dirigirenden Λ durch aufsetzung seines huths geöffnet ist wird der candidat von dem jüngern thürhüter aus einem andern zimmer abgeholt und bey der hand ein und vor des dirigirenden Λ tisch geführt dieser fragt ihn:

erstlich ob er begehre ◊ zu werden

zweytens denen verordnungen der ⊗ sich

unterwerffen und ohne widerspenstigkeit die lehrzeit ausstehen wolle.

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

der candidat antwortet ja.

Copiale Decipherment

lit:mz||bl
 0x3|1s5xpx|wn
 π0j0tΔpê3câ=7λûb⊙ur⊙z

$$\delta R_{\mu\nu} + \delta j_{\mu\nu} \hat{h}^{\mu\nu} \dot{f}.$$

cûjſêztrp†9η∧:κ

$\partial \hat{x} \hat{u} \hat{\eta} + \varepsilon \hat{r} \hat{k} \hat{m} \hat{l} \hat{i} z : \ddot{y} \hat{\gamma} \hat{\lambda} \hat{\sigma} \hat{o} \hat{q} z i \hat{x} \hat{a} | \hat{e} c : u \partial.$

$$f \hat{u} \hat{r} \hat{o} | \kappa \wedge \iota = \dot{c} z.$$
$$m\omega\hat{r}\hat{a}+\Delta g\dot{y}\underline{x}z\hat{o}3mn\ddagger\Gamma\underline{n}\hat{h}\hat{h}+\hat{i}f.$$

κτûπ:πziôf|ÿ)hê|hΞιλππâzbΔgz=|γlzπρçλâτgκλh

[illegible]

nîrîcî, eôphôrôp, 6ûh, 3âk  g, s = 3mêjzuli

g s m u o o l o n | o p e g u d o t r d | z g k r x u g p u z k ⑤ n | r l b = n

$\Lambda \hat{e} r m o z f : u a = r z l \Delta h r p m n \hat{a} z o i | d \hat{i} r f i s \delta y \lambda o z i e g c u t$
 $r s \hat{e} n \Lambda$

$$c\dot{\rho} = |\hat{f}\hat{a}| \hbar \underline{u} b m_{\Delta} \dot{c} : \frac{1}{2} \partial.$$

hziql:ôg|ezqâz 4 nπâ3 9 ms†r†îz†mâÿδulî= nπ

[illegible]

lπô3 fʒhgz ŷπr l n m Δ r l l o e z n r.

First lawbook

of the $\mathbb{C}^*$ e $\mathbb{C}^*$

Secret part.

First section

Secret teachings for apprentices.

First title.

Initiation rite.



If the safety of the Δ is guaranteed, and the Δ is opened by the chief λ , 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 λ , who asks him:

First, if he desires to become $\odot$.

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

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

The candidate answers yes.

Copiale Decipherment

Documentation at
<http://stp.lingfil.uu.se/~bea/copiale>
Google: "copiale"

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.

lit:mz||bl
v̇ẋż|l̇ṡk̇ṗk̇j̇|wn
πδ|ρ̇ḣΔ̇ẋēżċā=γ̇λ̇ū̇ḃ☉̇u̇ṙ☉̇ż

δ̇ḣḣżi+δ̇i̇ṅṗλ̇ḣi̇ṅċḟ.

ċū̇j̇ḟēżṫṗṫġη̇λ̇:κ̇

δ̇ẋū̇ḣē̇η̇+ε̇ṙṗk̇ṁλ̇i̇ż:γ̇j̇λ̇δ̇ō̇j̇q̇żi̇ẋā̇|ē̇ċ:u̇δ̇.

ḟū̇ṙḟō̇j̇κ̇λ̇i̇λ̇=̇ċż.

ṁj̇ḣṙā̇+Δ̇ġj̇ẏẋżδ̇ṁṅk̇ḟṅḣḣ+i̇ḟ.

κ̇ṁū̇ṙ:π̇żi̇δ̇ḟ|j̇ẏḣē̇j̇ḣżi̇λ̇ṅπ̇ā̇żḃΔ̇ġż=̇j̇l̇żu̇ṗq̇ċλ̇ā̇ṙġκ̇λ̇ḣ

ḣṙḣḣλ̇żi̇ṅṗi̇|Δ̇ṙδ̇ḣλ̇a=̇ġżẇπ̇j̇ē̇ċΔ̇ṙδ̇Δ̇+̇ḃżη̇ṙi̇ẋj̇i̇ṅżu̇ḟλ̇ż

π̇k̇j̇żδ̇ṙ=̇ḟ|ū̇λ̇ṡk̇ṁδ̇ṁ|żη̇ġā̇|κ̇ḣ=̇λ̇ḣ|l̇ẋδ̇ẇḟ:π̇λ̇ḃi̇ḟu̇ṁj̇żv̇

żā̇j̇ẋj̇ṙṁπ̇żḣλ̇ċδ̇ō̇ġżżū̇+̇ṗk̇ḣṅẋē̇żġḣλ̇ḣḣi̇ḣḣλ̇i̇żṫṅk̇|ṙō̇ẏ

ṁā̇+̇ḣḣṙżδ̇żṅḃṡη̇+̇żṙk̇ṙṗδ̇ḣΔ̇ċū̇λ̇ġ=̇ṅżκ̇ṗẋȯoṅżḣj̇ḟḣi̇ṅπ̇

żē̇j̇ṅġ=̇ṙπ̇ġδ̇Δ̇żṅżḣ|κ̇π̇η̇j̇i̇ẋj̇ṙi̇ġπ̇u̇λ̇ḃġλ̇i̇ḟḃδ̇ū̇ḟḣḣżδ̇λ̇ċ

żη̇δ̇|ū̇ṙċḟżq̇δ̇λ̇ḃη̇ḣṁδ̇:

ṅi̇ṙḟċi̇ẋē̇ō̇ṗḣδ̇ṙδ̇ṗżδ̇ū̇ḣżā̇κ̇☉̇ġṡ=̇δ̇ṁē̇j̇żu̇l̇i̇

ġṡṁū̇ȯλ̇δ̇η̇|ȯπ̇ē̇ġu̇δ̇δ̇żṙΔ̇j̇żġk̇ṙẋu̇ġπ̇ū̇żκ̇☉̇ṅ|η̇j̇ḃ=̇ṅ

λ̇ē̇ṙṁδ̇żḟ:u̇a=̇ṙżl̇Δ̇ḣṙḣδ̇ṁη̇ā̇żδ̇i̇|δ̇i̇ṙḟi̇δ̇δ̇ẏλ̇ȯżi̇ē̇ġċū̇ḣ

ṙṡē̇η̇λ̇

ċṙ=̇ḟḟā̇ḣżḃṁΔ̇ċ:żδ̇.

ḣżj̇η̇λ̇:δ̇ġ|ē̇żη̇ā̇ż☐̇ṅπ̇ā̇ż☉̇ṗṡk̇ṙδ̇i̇żṫṁā̇j̇δ̇u̇l̇i̇=̇ṅπ̇

żṙṡ=̇ṅḣk̇ḟṙżṙ|ṅδ̇żi̇ṗġπ̇ċj̇ẏḣḟδ̇ḃi̇j̇κ̇ṅġḣżj̇ḟi̇δ̇ċṡ=̇ḃ+̇ṅj̇u̇δ̇

ẋδ̇i̇ṙ:ū̇λ̇v̇|ō̇oṅ.

l̇π̇δ̇żḟj̇ḣġżj̇ṗṙλ̇ṅṁλ̇ṁΔ̇ṙλ̇ḣl̇ȯē̇q̇ṅṙ.

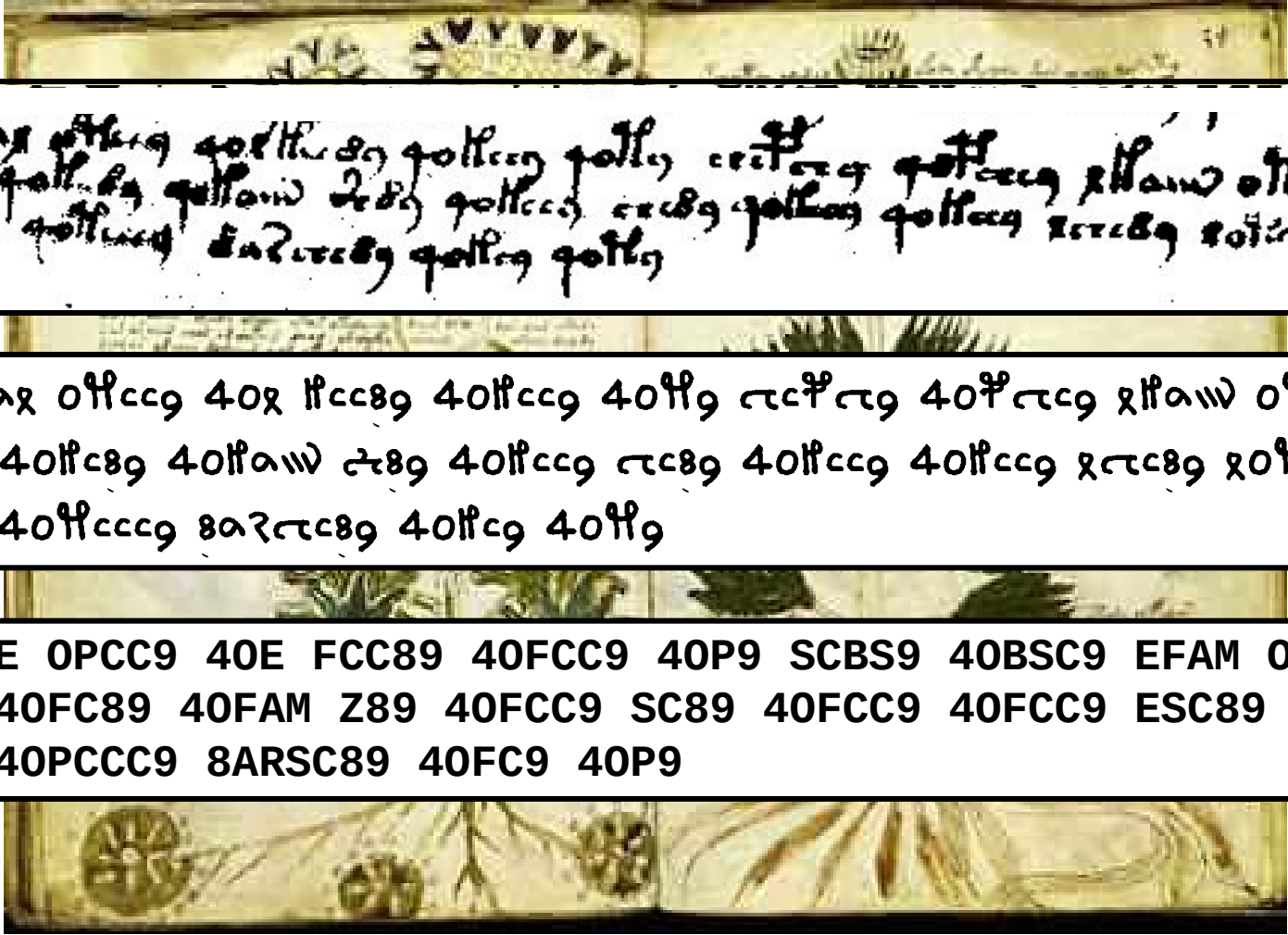
Voynich Manuscript



- Medieval illustrated manuscript
- 235 pages, 6 sections, 38k word tokens, 35 letter types
- Undeciphered



Voynich Manuscript



qolccg qolccg qolccg qolccg qolccg qolccg qolccg qolccg
qolccg qolccg qolccg qolccg qolccg qolccg qolccg qolccg
qolccg qolccg qolccg qolccg qolccg qolccg qolccg qolccg

qolccg qolccg qolccg qolccg qolccg qolccg qolccg qolccg
qolccg qolccg qolccg qolccg qolccg qolccg qolccg qolccg
qolccg qolccg qolccg qolccg qolccg qolccg qolccg qolccg

BSC8AE 0PCC9 40E FCC89 40FCC9 40P9 SCBS9 40BSC9 EFAM 0PAE29
2ZC9 40FC89 40FAM Z89 40FCC9 SC89 40FCC9 40FCC9 ESC89 EOP9
8ZC9 40PCCC9 8ARSC89 40FC9 40P9

Voynich Letter Substitution

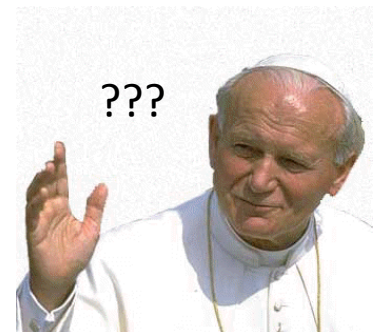
Latin letter trigram model

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

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

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

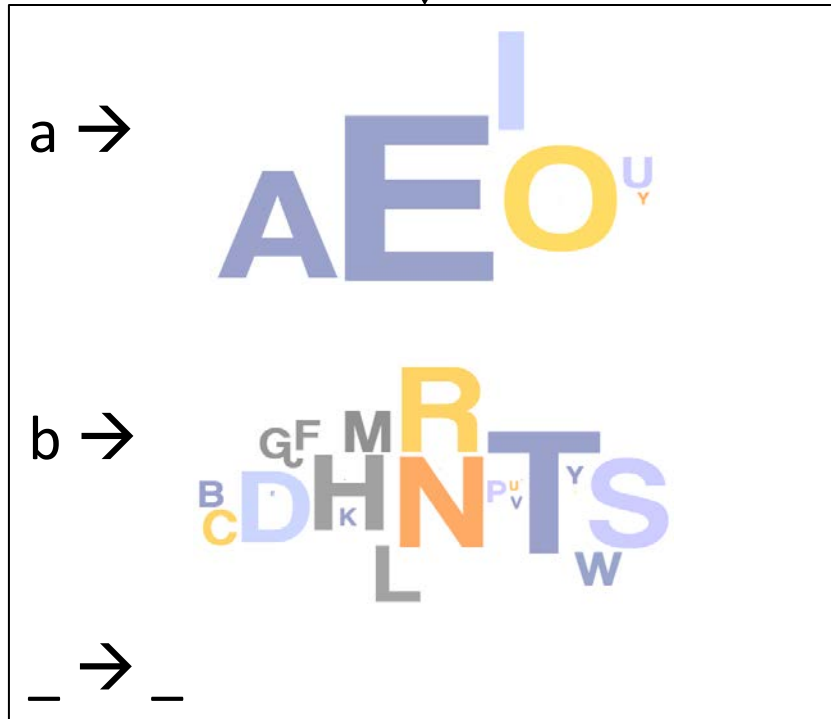
Input	Decipherment
VAS92 9FAE AR APAM ZOE ZOR9 QOR92 9 FOR ZOE89 ...	quiss squm is onum pom quiss hates s qum hatis ...



Letter Clustering

Trigram model over {a, b, _}

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



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 _ ...

Letter Clustering

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 _ ...

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	_	...

Letter Clustering

Trigram model over {a, b, _}

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



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

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 _ ...

Letter Clustering

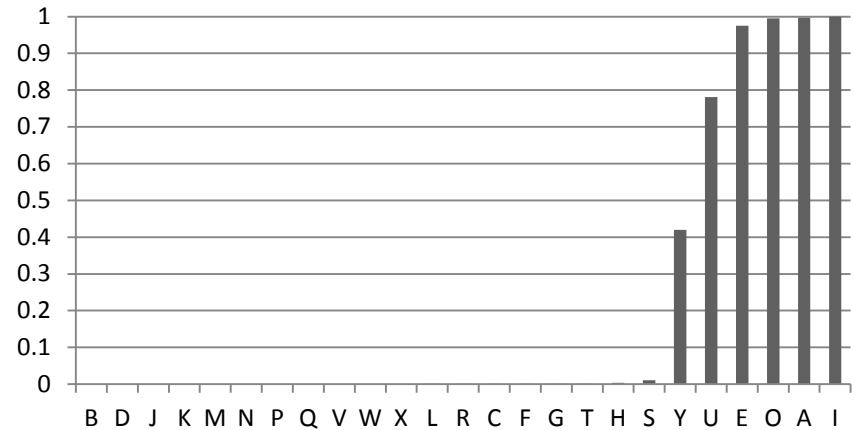
$P(\text{letter} \mid \text{tag})$

English

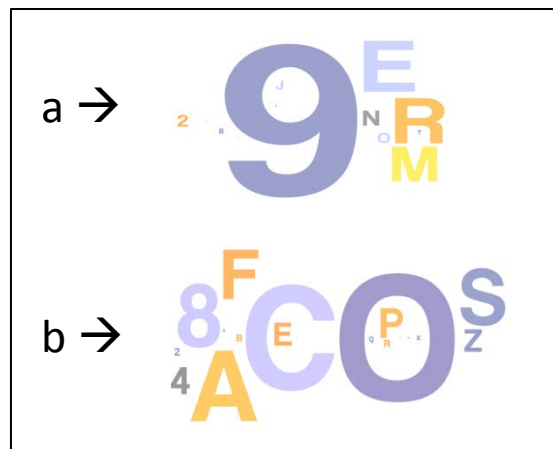


$P(\text{tag} \mid \text{letter})$

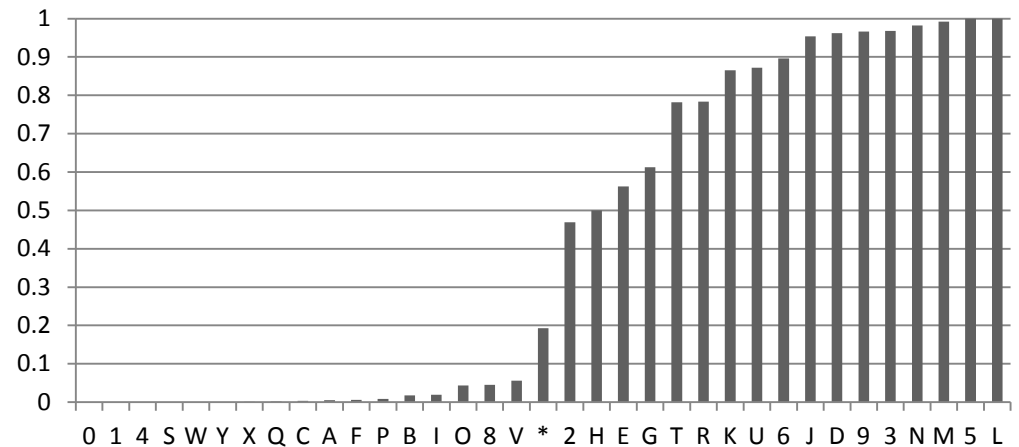
$P(a)$



Voynich



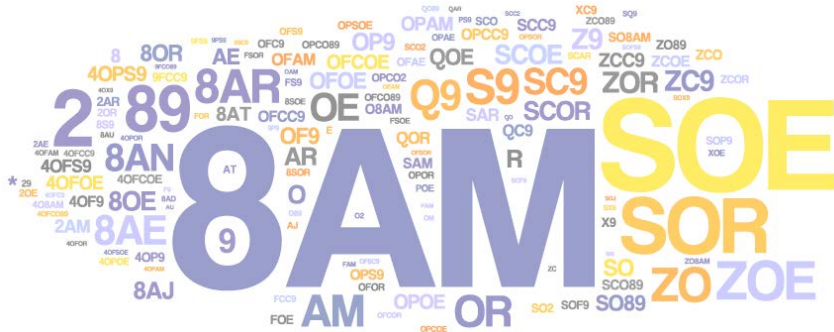
$P(a)$



Word Clustering

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

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



b →



VAS92 9FAE AR APAM Z0E Z0R9 QRC2 9 ...

Sample tagging with learned model:

a a a a a
VAS92 9FAE AR APAM ZOE

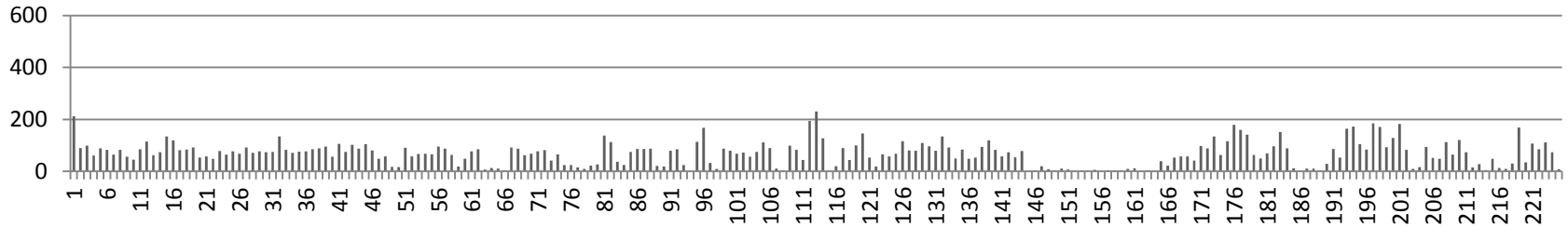
a	a	a	a	a	...
ZOR9	QRC2	9	FOR	ZOE89	...



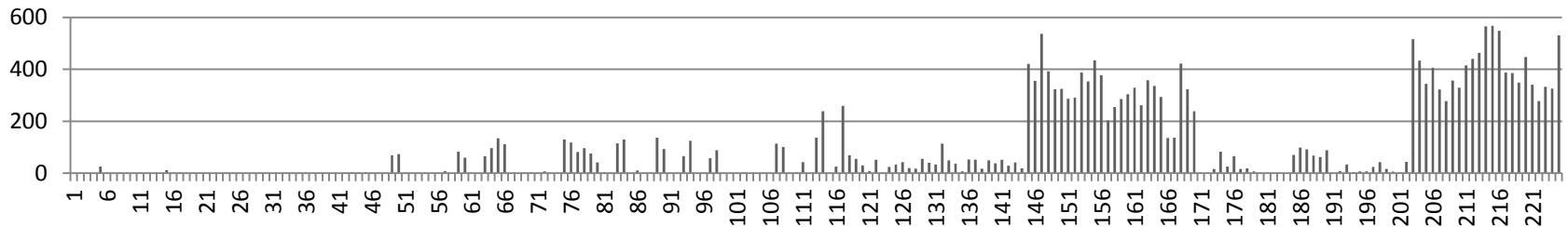
Word Clustering

Voynich words tagged as “a”

← pages →



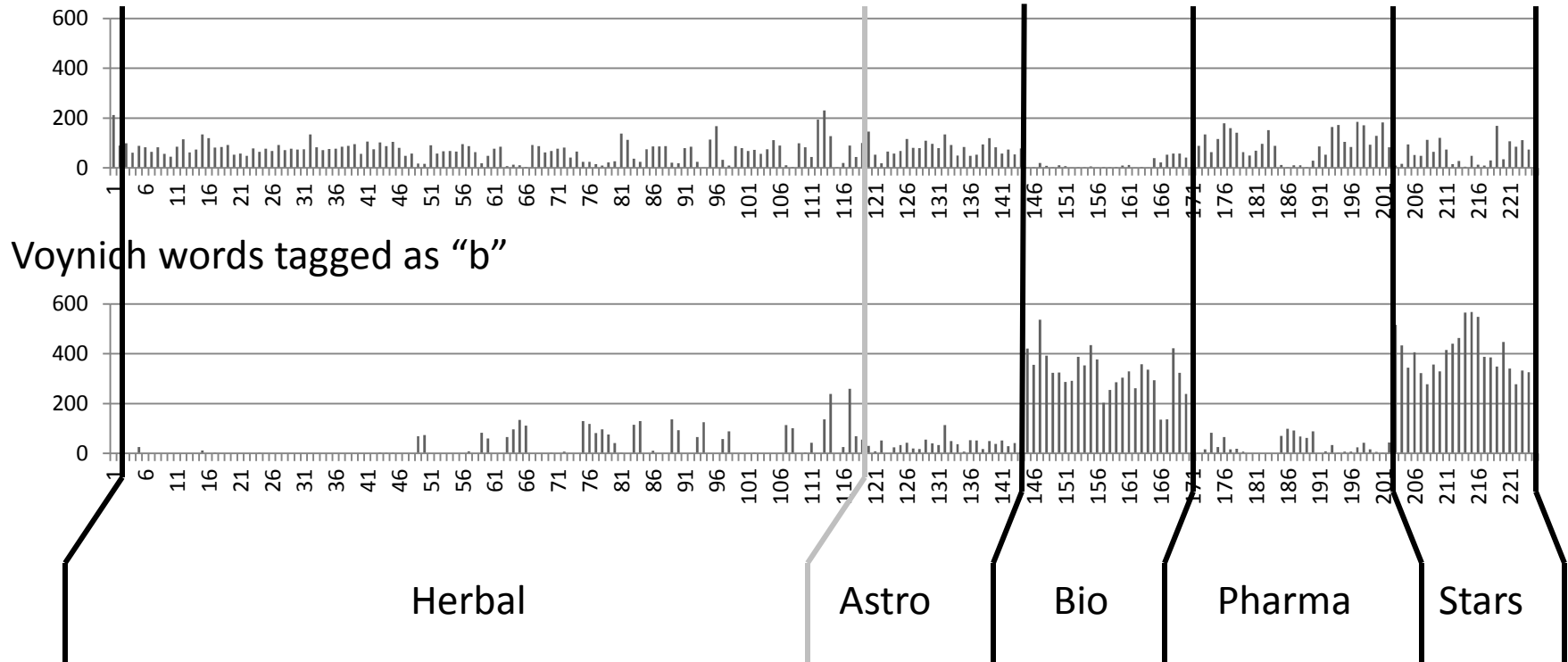
Voynich words tagged as “b”



Word Clustering

Voynich words tagged as “a”

← pages →



Voynich sections, per drawings observed.
Captain Currier’s “two languages” (1976).

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

An Application of PTAH to the Voynich Manuscript (U)

BY MARY E. D'IMPERIO

~~Top Secret Umbra~~

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

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

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

1970s NSA report
recently declassified!

National Security Agency

NSA applies statistics to ciphers, codes, and other language processing problems

NSA employs more mathematicians and linguists than any other organization.

NSA has more computers than any other organization.

Oh yeah -- we've been on Mars since 1962.



??

1970s paper on HMM Voynich

1950s

1960s

1970s

1980s

1990s

2000s

2010s

2020s

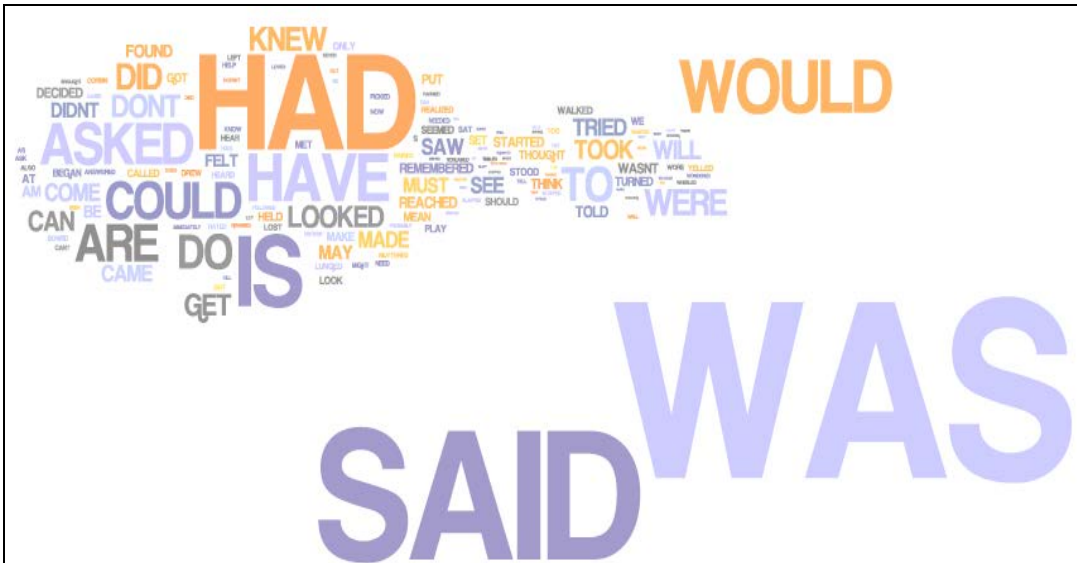
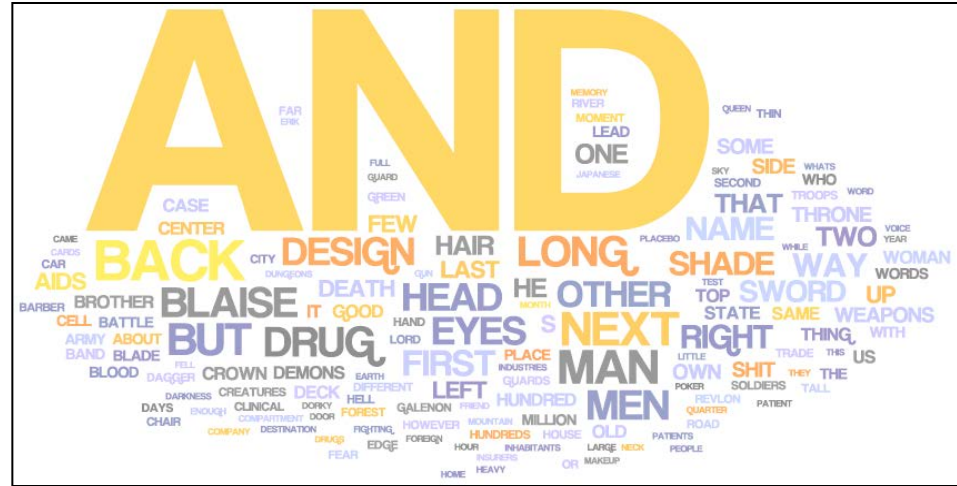
Association for Computational Linguistics

1993 paper on Statistical Machine Translation

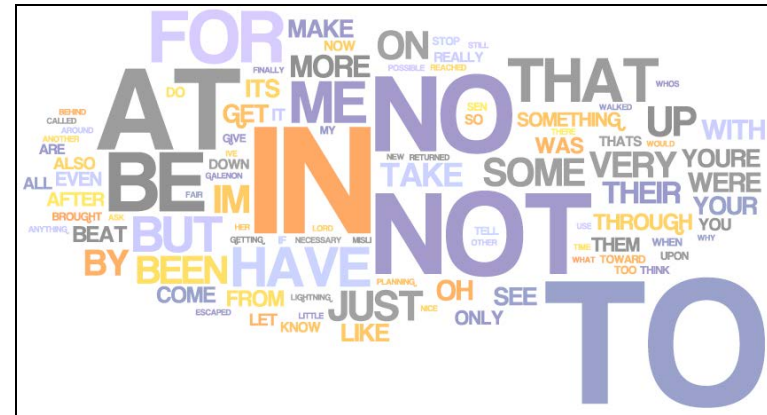
2011 paper on HMM Voynich

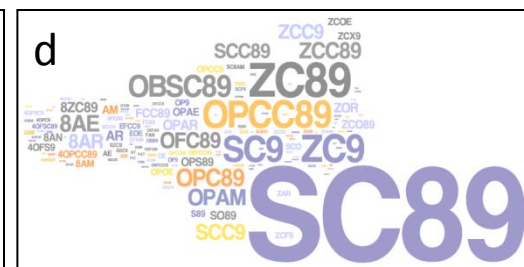
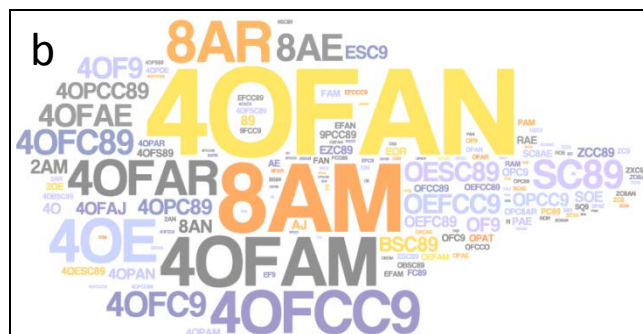
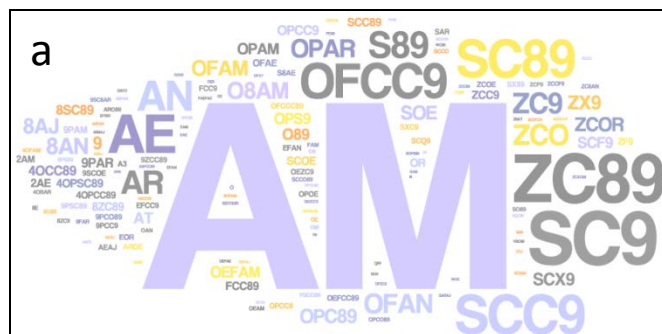
ACL applies statistics to language processing problems

etc

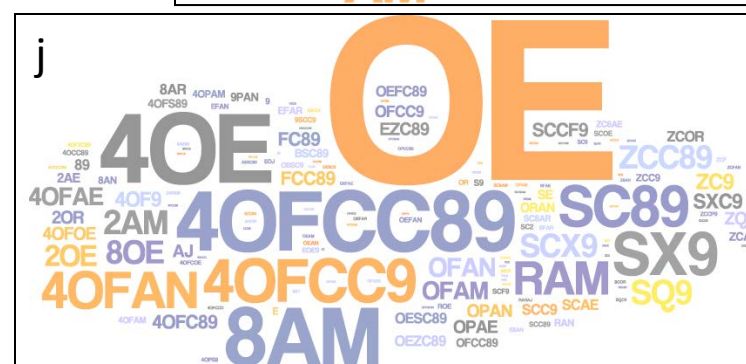
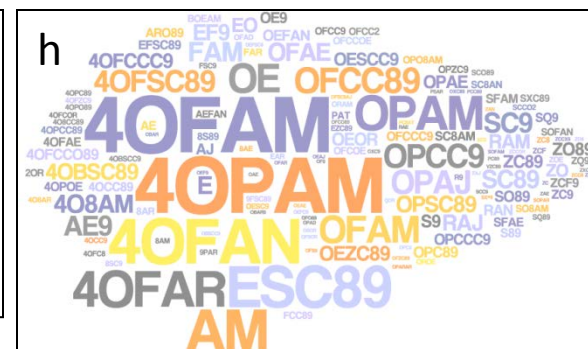
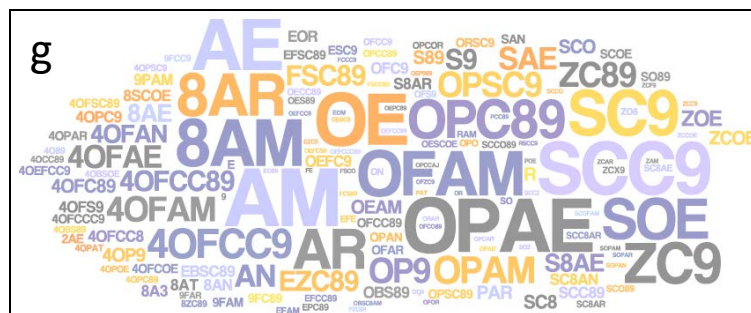
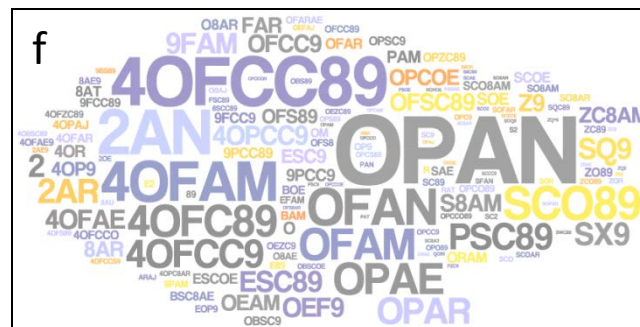
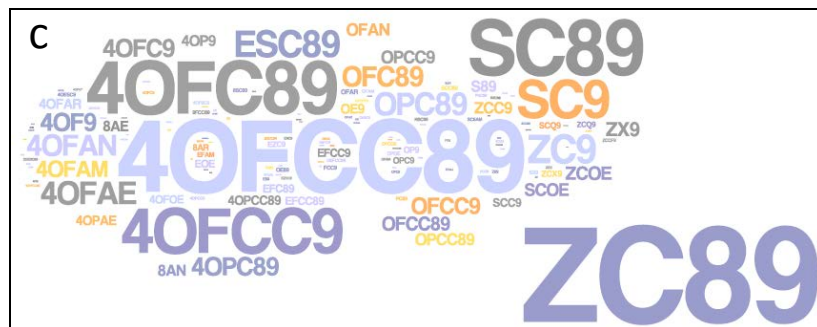


etc



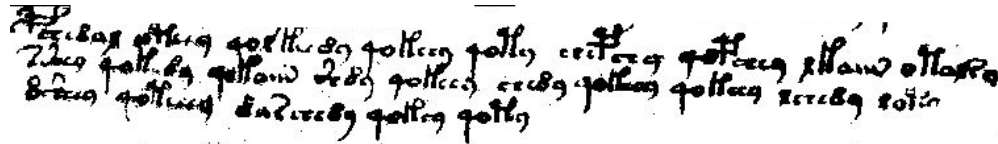


10 Classes:
Voynich-B



Other Unsolved Ciphers

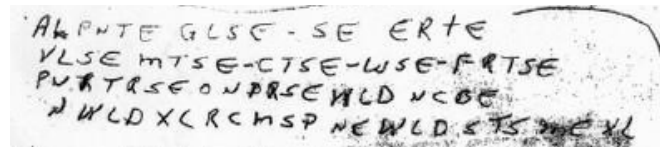
Voynich Manuscript (1400) Reddy & Knight 11



Zodiac Killer (1967) Ravi & Knight 11



FBI cipher (1999)



Kryptos (1990)

OBKR
UOXOGHULBSOLIFBBWFLRVQQPRNGKSSO
TWTQSQSSEKZZWATJKLUDIAWINFBNYP
VTTMZFPKWGDZXTJCDIGKUHUAUEKCAR

Lost
Language
Decipherment



Snyder, Barzilay, Knight 10

Unsupervised
Training for NLP

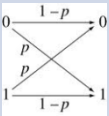
Machine
Translation ...

Plan for This Talk

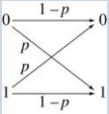
- Break a series of codes
 - Simple letter substitution
 - Phonetic substitution
 - archaeology
 - transliteration
 - Word substitution
 - Foreign language as cipher
- Bonus
 - Two historical ciphers
 - Final thought on translation and cryptography



Future Prospects for Translation

		Cryptography	Translation
Manual		Manual encoding	Human translation
Mechanical		1920s Mechanical encoding; intuition-based decryption	
Mathematical		1950s Computer decryption, based on information theory	
Higher math, deeper understanding		1980s Public-key systems, based on number theory	

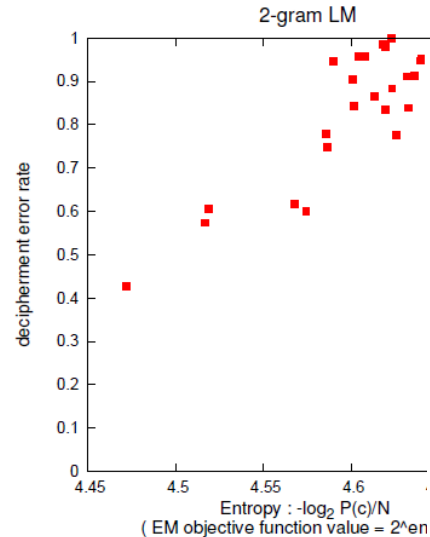
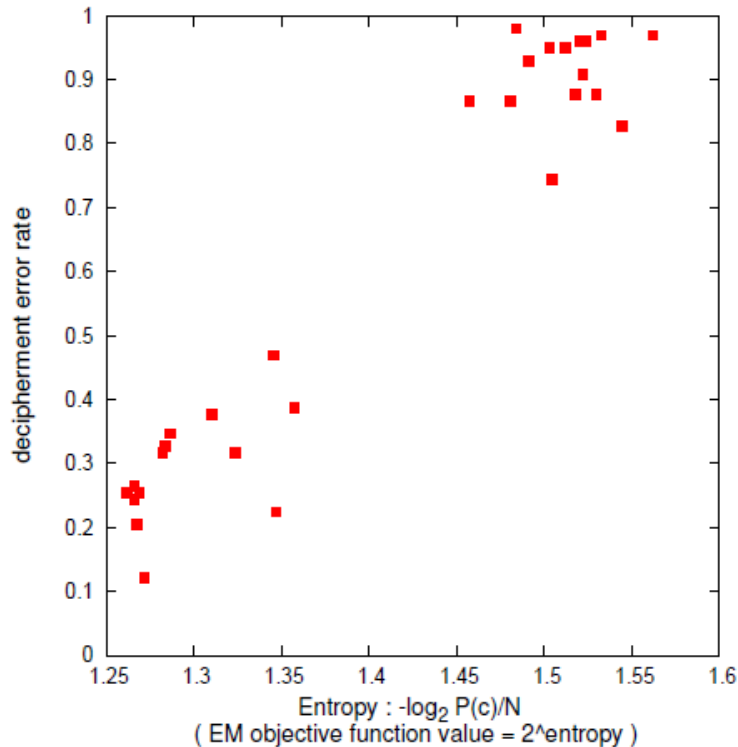
Future Prospects for Translation

		Cryptography	Translation
Manual		Manual encoding	Human translation
Mechanical		1920s Mechanical encoding; intuition-based decryption	1960s Rule-based MT
Mathematical		1950s Computer decryption, based on information theory	1990s Statistical MT
Higher math, deeper understanding		1980s Public-key systems, based on number theory	2020s ??? ??? ???

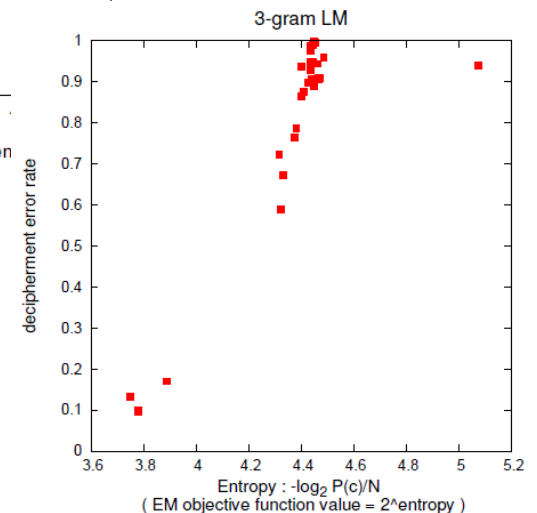
thanks

Random Restarts are Critical

English 98-letter cipher, 3-gram LM

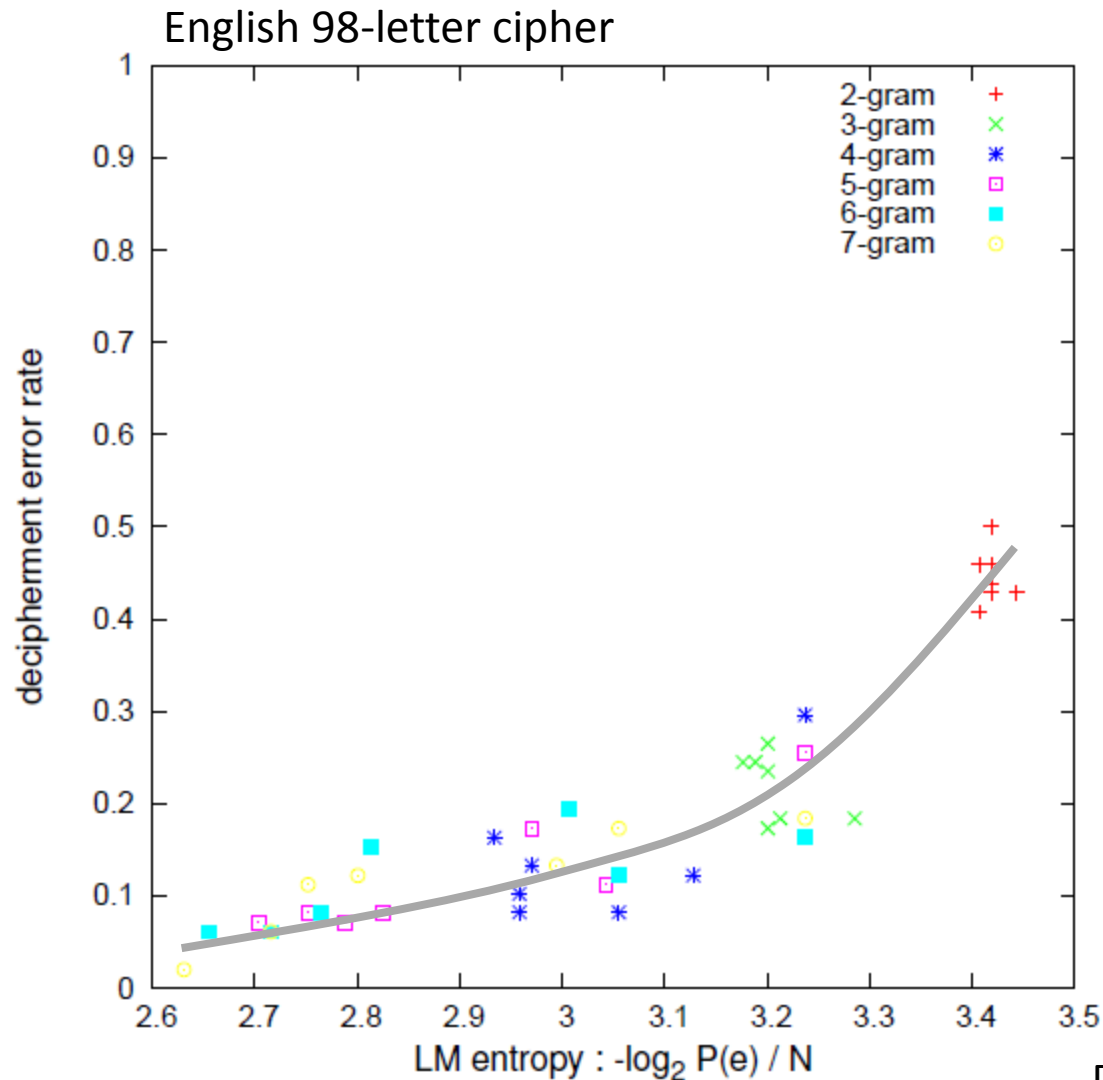


Japanese
syllable
cipher



even people do restarts!

Good Language Models are Critical



[Ravi & Knight 09b]

Deterministic Substitution Constraint

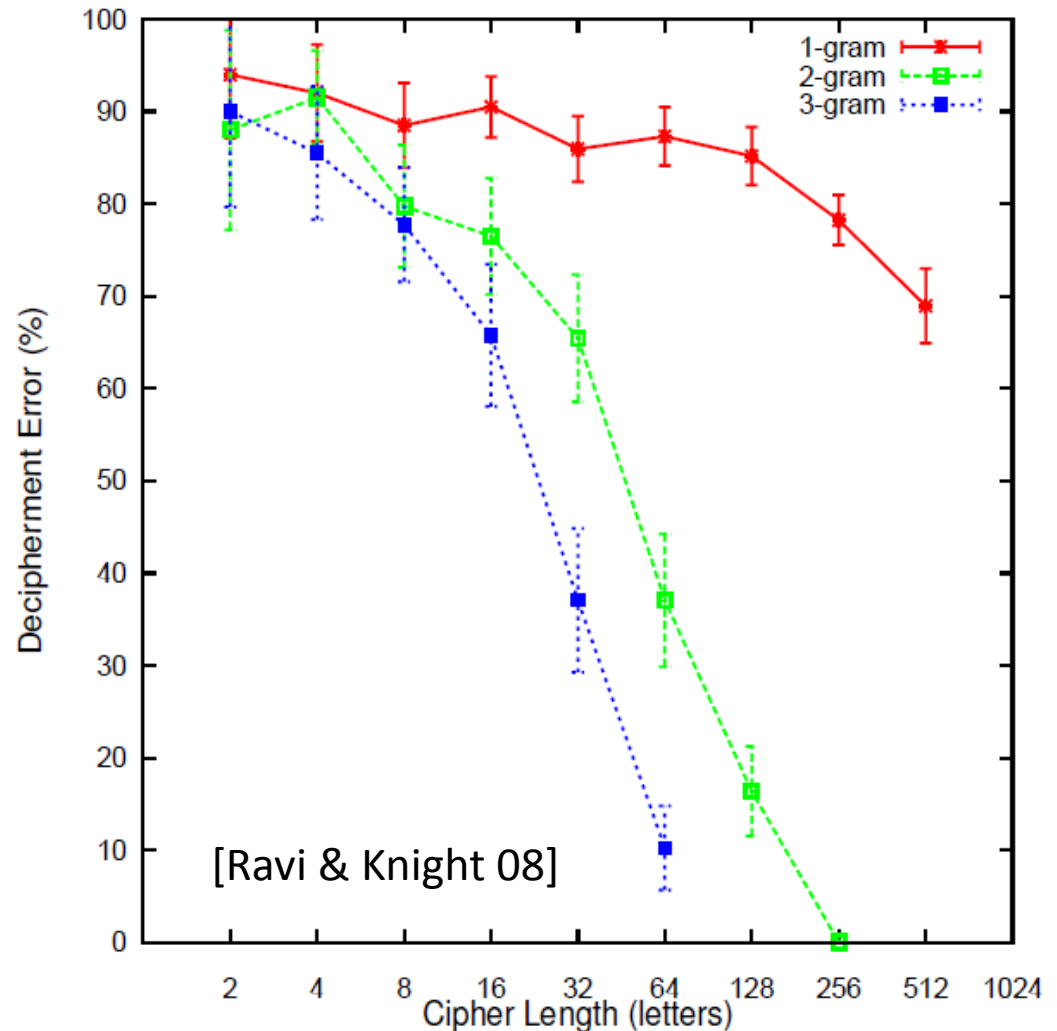
Using ILP instead of EM

- * Search only over deterministic keys.
- * Exact, no restarts.



Cipher Length	EM error	ILP error
52	85 %	21 %
98	45 %	12 %
414	10 %	0.5 %

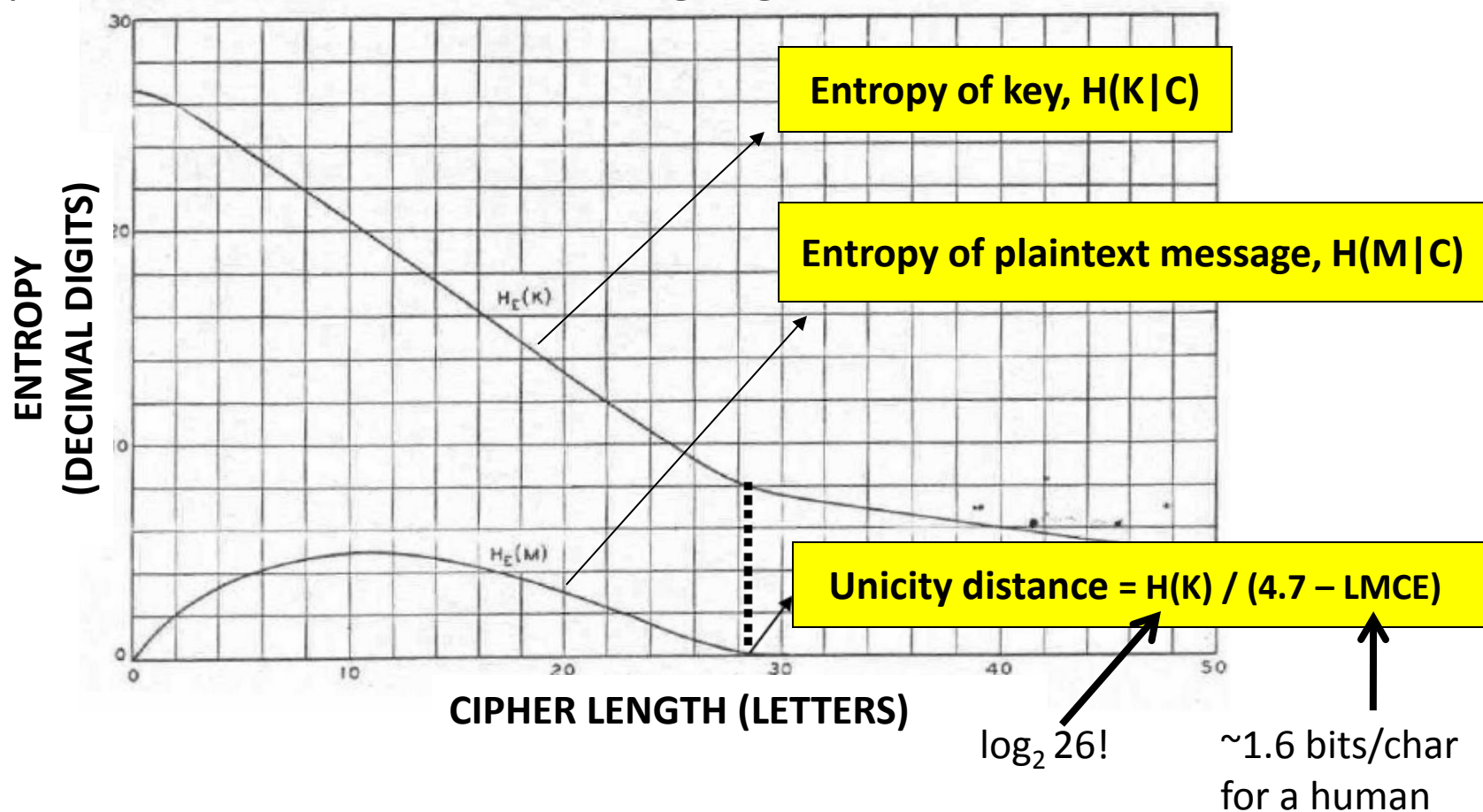
Using 2-gram letter-based LM



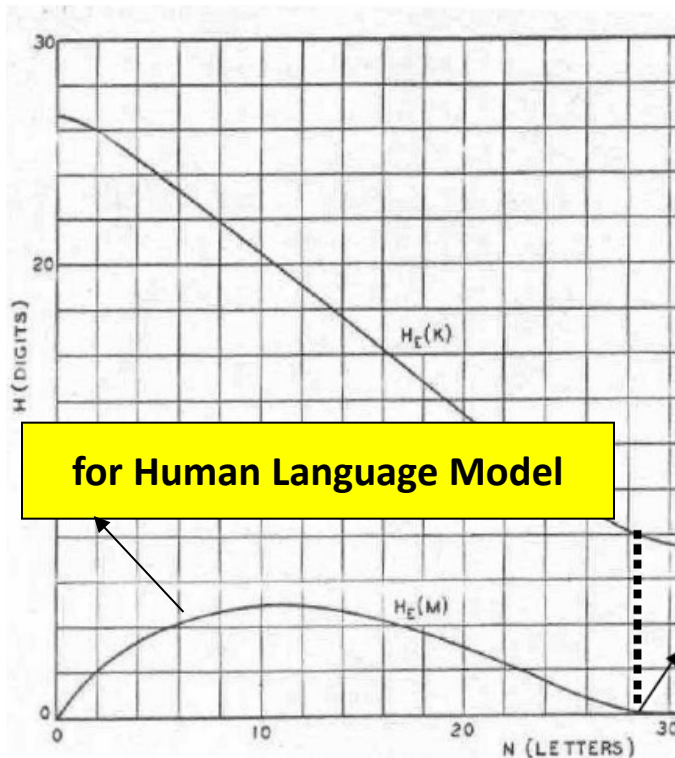
[Shannon 46, 49]

“Communication Theory of Secrecy Systems”

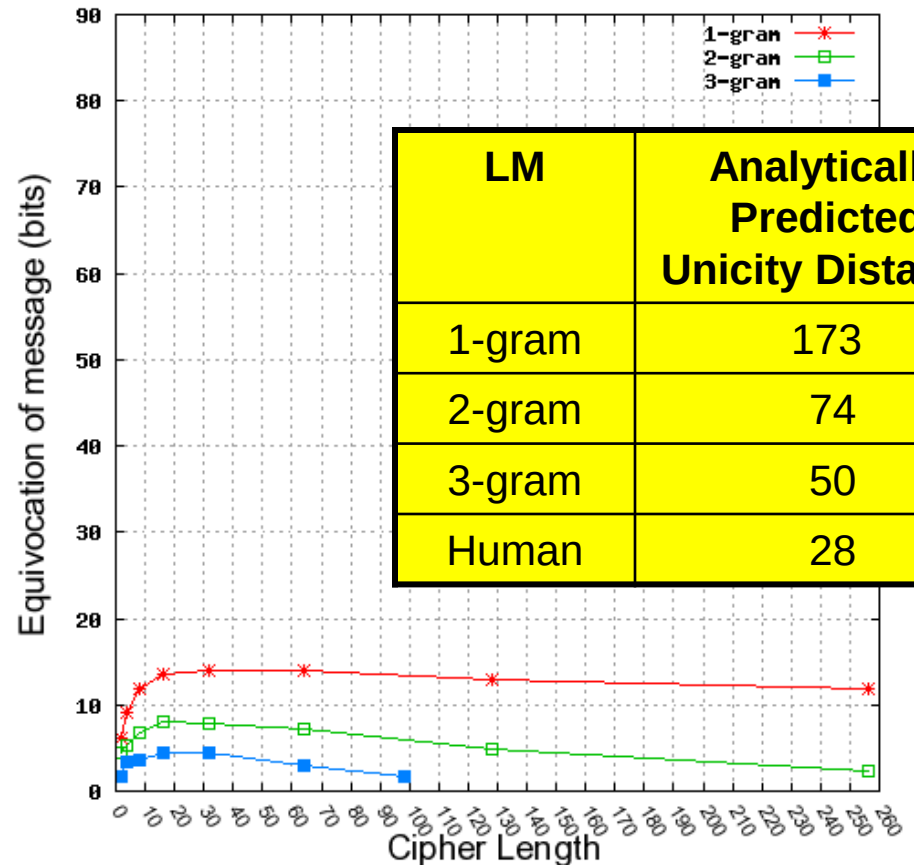
- Shannon analytically predicted uncertainty about key and message
- Graphed it for a human-level language model



Verifying Shannon's Prediction of Plaintext Message Uncertainty



ANALYTIC CURVES
(Shannon's)

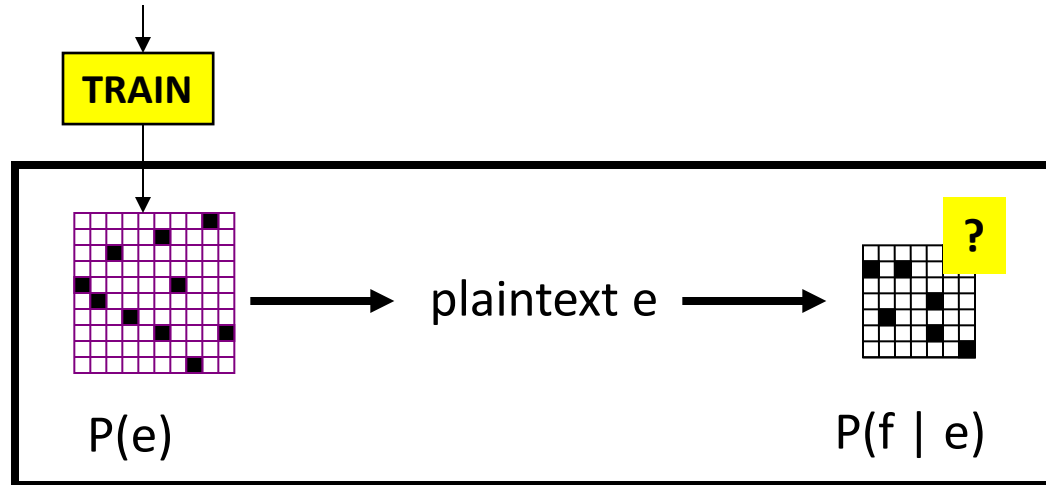


ACTUAL CURVES
(ours)

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".

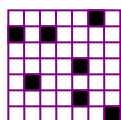
الخيار بكل ترحاب من جانبه شجب رئيس الحكومة الفلسطينية أحمد قريع الطابع الأحادي الجانب للانسحاب الإسرائيلي من غزة، وأكد أن إسرائيل تريد مغادرة هذه الأراضي لتعزيز سيطرتها على الضفة الغربية وقال قريع في كلمة له خلال مؤتمر نظمته وزارة الأوقاف في رام الله "سينسحبون من غزة ولكننا لا نعرف ما هو شكل هذا الانسحاب وماذا سيتركون، وما هو مصير المعابر والحدود، وكل ذلك غامض لأنه قرار أحادي الجان

Word Substitution Cipher

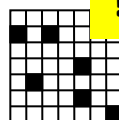
.....France.....Britain.....Canada...
Mexico.....Indonesia.....Malaysia...
Britain.....Canada.....Australia...
Britain.....France.....Indonesia.....
Mexico.....Australia.....France...
 ...Britain.....

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

TRAIN



plaintext e



?

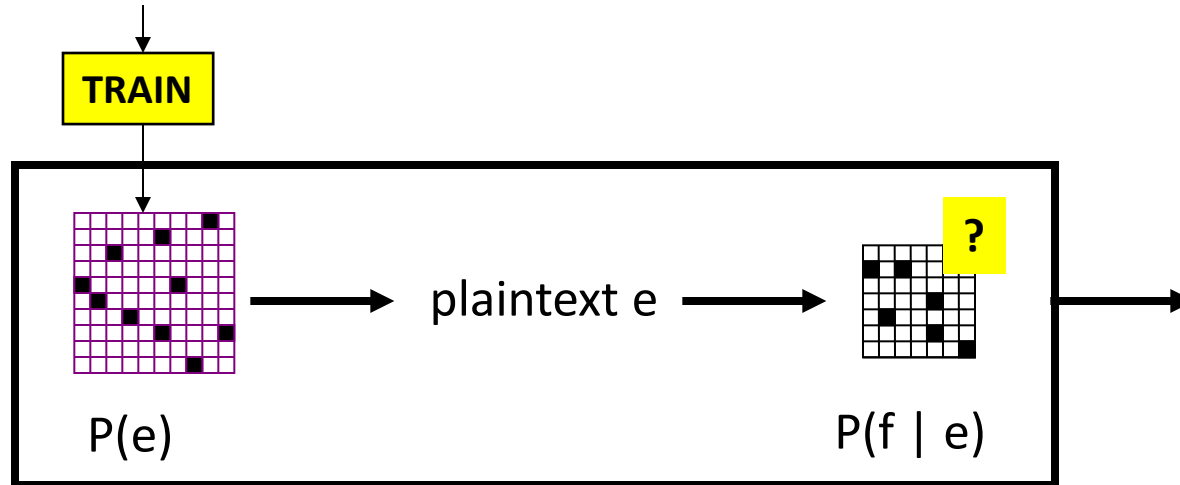
$P(\text{sentence has } w1 \mid \text{sentence has } w2)$

$P(f \mid e) =$
7 x 7 subst table

.....knd!.....bryT!ny!
knd!.....
 !lmksyk.....
 ...!ndwnysy!.....!lmksyk.....
bryT!ny!.....m!lyzy!...
bryT!ny!.....frns!.....
!str!ly!.....!ndwnysy!...
frns!.....frns!
frns!.....bryT!ny!.....
!str!ly!.....

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

Zodiac Killer Ciphers

Zodiac 408 (solved, 1969)

Δ ▣ P / Z / U B ▣ X O R π 9 X π B
W V + 3 G Y F O Δ H P ▣ K π ρ Y 3
M J Y Λ U I X Δ ρ T ⊥ N Q Y D ● ●
S ϕ / Δ ▣ B P O R A U ▣ 7 R J ρ E
X Λ L M Z J O R \ 9 F H V W 3 Δ Y
▣ + ρ G D Δ K I ● ● ρ X Δ ● ● S ϕ
R N ⊥ I Y E J O Δ ρ G B T Q S ▣ B
L O / P ▣ B ▣ X ρ E H M U Λ R R X

3 Z K ρ 9 I ● W ρ I Δ ● L M R Δ ▣
B P D R + τ π ρ \ N ϕ 3 E U H K F
Z 3 9 O V W I ● + ⊥ L ● J Λ R O H
I Δ D R ▣ T Y R \ O 3 / ▣ X J Q A
P ● M Δ R U ⊥ ▣ L ● N V E K H π 6
R I I J X ● Δ Δ L M J N A ● Z ϕ P
ϕ U 9 X A Δ ▣ B V W \ + V T ⊥ O P
Λ π S ρ J F U 3 O Δ ρ ϕ G ▣ ▣ I M

N K ● S 3 E / Δ ▣ ▣ Z 7 A P ▣ B V
9 3 X ρ W ρ ▣ F ▣ Δ 3 + ▣ Δ A Δ B
▣ O T ● R U 3 + ▣ Q Y ρ ▣ Λ S ρ W
V Z 3 G Y K E ▣ T Y A Δ ▣ ▣ L ⊥ ▣
H I F B X Δ ϕ X A D ▣ \ Δ L I π ρ
▣ 3 ▣ ▣ ● 3 ● P O R X Q F ▣ G 3
Z ▣ J T ⊥ ρ ▣ Δ J I + R B P Q W O
V E X R Δ W I ρ ρ E H M ρ π U I K



Zodiac 340 (still unsolved)

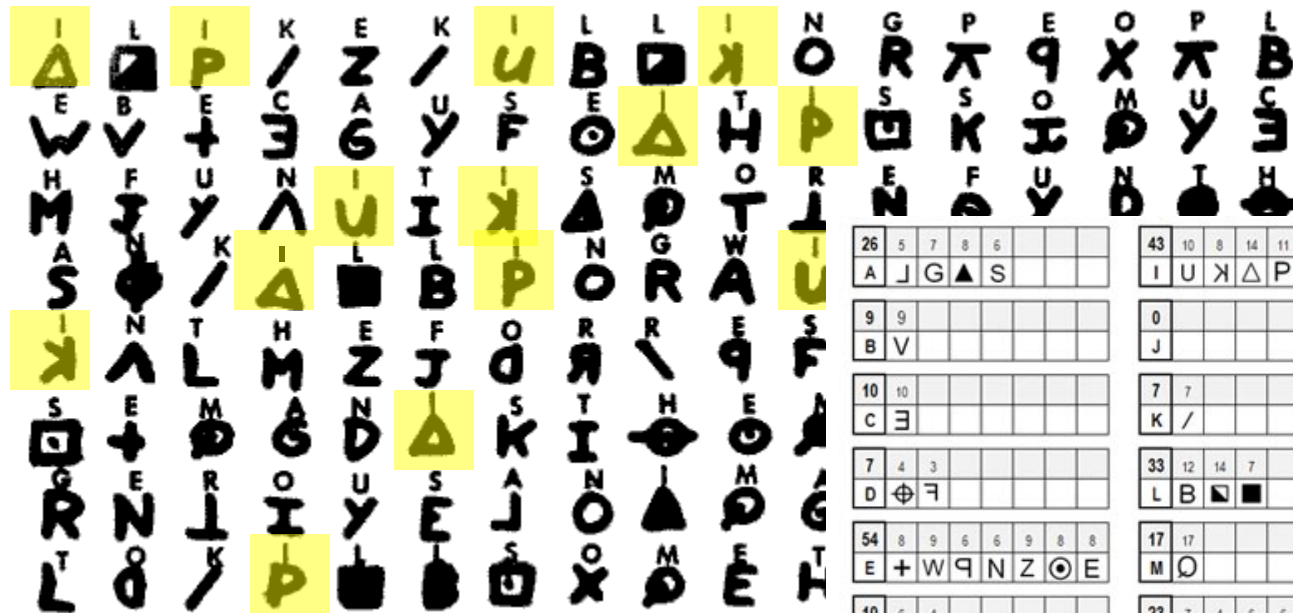
HER > 9 J Λ V P X I ● L T G ● Q
N 9 + B ϕ ▣ O ▣ D W Y · < ▣ K 7 ρ
B X π 3 M + u z G W ϕ ● L ▣ ϕ H J
S 9 9 Δ Λ J Δ ▣ V O 9 O + + R K ●
▣ Δ M + ϕ ⊥ T Q I ● F P + ρ X /
9 Δ R Λ F J O - ▣ Q C 7 F > ● ρ ϕ
▣ + K ρ ▣ π ● u 3 X G V · ϕ L I
ϕ G ● J 7 T ▣ O + ▣ N Y ϕ + ▣ L Δ
O < M + 8 + Z R ● F B 3 Y A O ● K
- ϕ J U V + Λ J + O 9 Δ < F B Y -
U + R / ● ⊥ E I D Y B 9 8 T M K O
● < 3 J R J I ▣ ● T ● M · + P B F
ϕ O Δ S Y ▣ + N I ● F B 3 ϕ π Δ R
J G F N Λ 7 ● ● ● 8 · 3 V ● ⊥ + +
Y B X ● ▣ π ● Δ C E > V U Z ● - +
I 3 · O ϕ B K ϕ O 9 Λ · 7 M ρ G ●
R 3 T + L ● ● C < + F J W B I ● L
+ + ϕ W C ϕ W 3 P O S H T / ϕ ϕ 9
I F X Q W < Δ ⊥ B ρ Y O B ▣ - C 3
> M D H N 9 X S ϕ Z O Δ A I K π +

COOP-SFPD
1596-78
7-14-78 GVL
7-27

#2 11-9-67

Zodiac Serial Killer

408-letter cipher (solved):



(plus two more sections)

26	5	7	8	6	
A	J	G	▲	S	

9	9				
B	V				

10	10				
C	☐				

7	4	3			
D	⊕	☐			

54	8	9	6	6	9	8	8
E	+	W	☐	N	Z	⊙	E

10	6	4				
F	J	Q				

11	11					
G	R					

16	8	8				
H	M	⊖				

43	10	8	14	11
I	U	X	△	P

0				
J				

7	7			
K	/			

33	12	14	7	
L	B	■	■	

17	17			
M	⊙			

23	7	4	6	6
N	○	Φ	Λ	D

28	6	7	9	6
O	⊠	T	X	⌚

7	7			
P	⌚			

0				
Q				

19	7	7	5	
R	⌚	Я	\	

22	7	6	3	6
S	F	K	△	⊠

33	7	8	8	10
T	●	L	H	I

10	10			
U	Y			

6	6			
V	⊙			

8	8			
W	A			

1	1			
X	⌚			

8	8			
Y	⊠			

0				
Z				

408				
-----	--	--	--	--

Zodiac Serial Killer

Plaintext solution

“ I LIKE KILLING PEOPLE BECAUSE IT IS SO
MUCH FUN IT IS MORE FUN THAN KILLING WILD
GAME IN THE FORREST BECAUSE MAN IS THE
MOST DANGEROUE ANAMAL OF ALL TO KILL
SOMETHING GIVES ME THE MOST THRILLING
EXPERENCE IT IS EVEN BETTER THAN GETTING
YOUR ROCKS OFF WITH A GIRL THE BEST
PART OF IT IS THAE WHEN I DIE I WILL BE
REBORN IN PARADICE AND THEI HAVE KILLED
WILL BECOME MY SLAVES I WILL NOT GIVE
YOU MY NAME BECAUSE YOU WILL TRY TO
SLOI DOWN OR ATOP MY COLLECTIOG OF
SLAVES FOR MY AFTERLIFE ”
EBEOR IETEMETHHPITI

Plaintext has many misspellings

Final 18 plaintext characters of 408 are “junk”

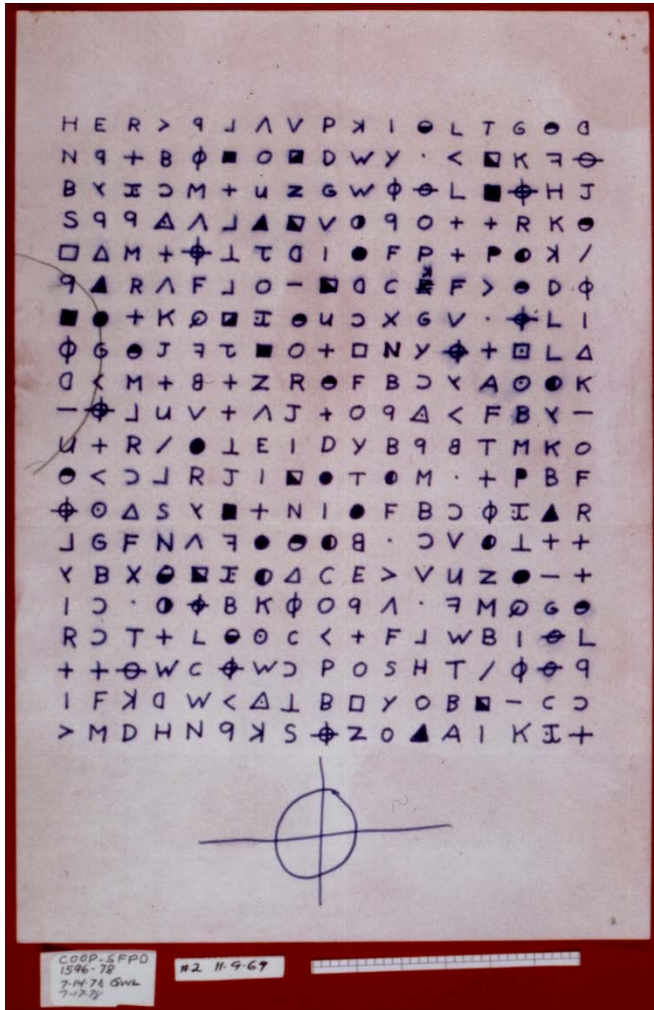
Deciphering Zodiac 408

Bayesian models

Extended Carmel finite-state toolkit to do Bayesian inference.
[Chiang et al 10]

Language Model	Initial Sample	Decipherment Error
3-gram	Random	62.3 / 48.5 / 47.4
5-gram	Random	all wrong!
“	3-gram solution	42.6
Word 1-gram	Random	all wrong!
<i>Interpolated</i> 5-gram and word 1-gram	Random	79.2
“	5-gram solution	3.3 / 2.6

Unsolved Zodiac 340



Has no obvious reading order bias:

% cipher bigram types that repeat (freq > 1)	Left/Right order	Up/Down order	Diag. North-East	Diag. South-East
Zodiac 408 (solved)	13 %	5	7	5
Zodiac 340 (unsolved)	7	6	8	5

Could be nonsense ... or maybe bigrams are smoothed out via more careful substitutions.