Language Translation and Code-Breaking

Kevin Knight
Information Sciences Institute & Computer Science Department
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)

Stanford University November 4, 2014

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。

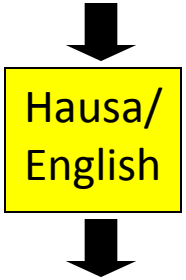
Kowane mutum na da hakkin ya sami yancin yin tunani da na sanin yakamata da na bin addini; saboda haka yana da yancin sake addini ko ra'ayin da ya bada gaskiya gare shi, da kuma yancin nuna addininsa ko ra'ayinsa, shi daya ko a cikin taro kuma a fili ko a boye ta hanyar koyarwa ko yin ibada, ko bauta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

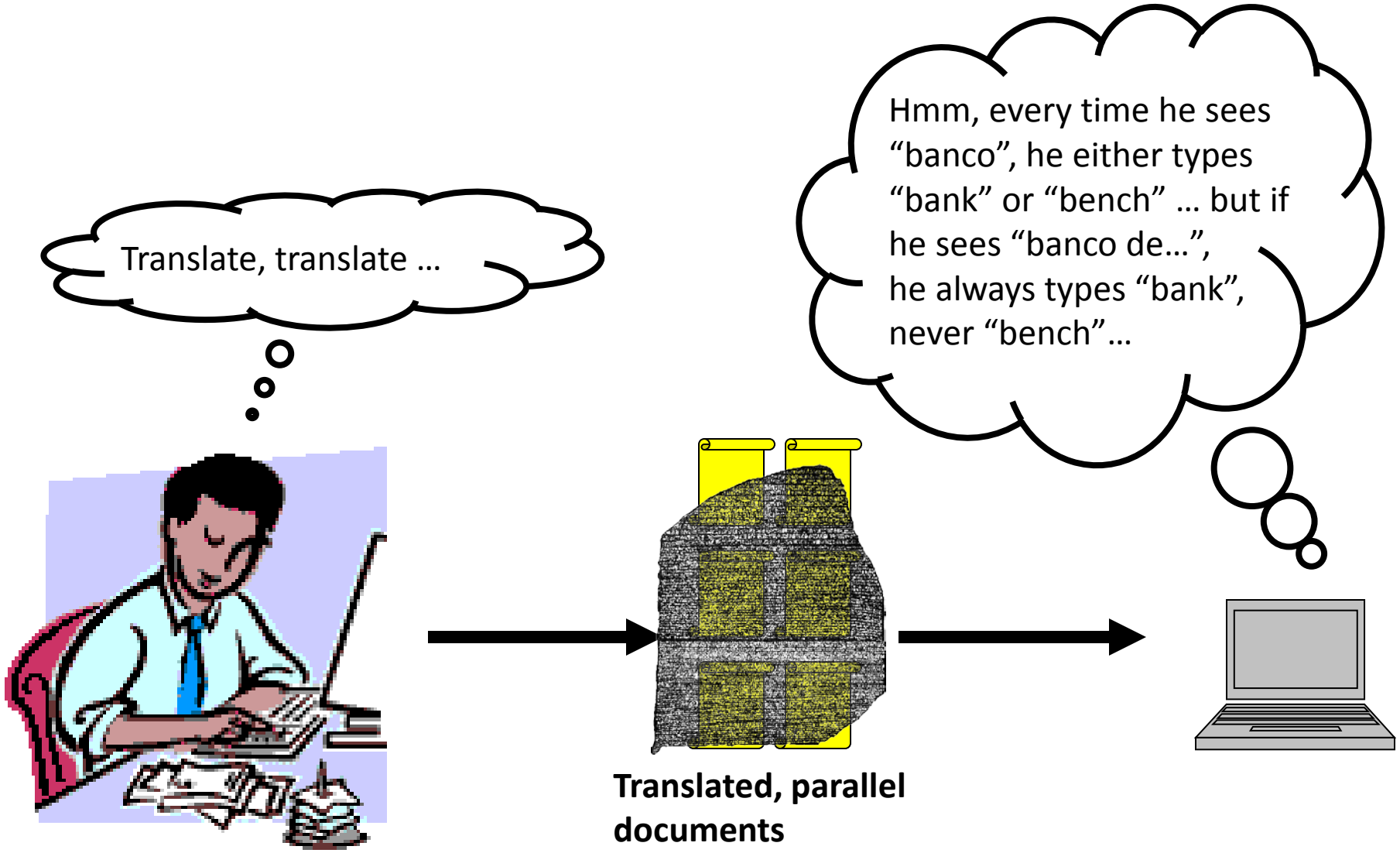
"When I look at an article in Russian, I say: this is really written in English, but it has been coded in some strange symbols. I will now proceed to decode." -- Warren Weaver (1947)

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 .

Parallel Corpus

12 English sentences in Centauri and Arcturan.

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 anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok 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 .

Centauri/Arcturan

Your assignment, translate this to Arcturan: farok crrrok 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 anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok 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 .

Centauri/Arcturan

Your assignment, translate this to Arcturan: farok crrrok 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 anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok **farok** izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok 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 .

Centauri/Arcturan

Your assignment, translate this to Arcturan: farok crrrok 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 anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok **farok** izok stok . 5b. totat **jjat** quat cat .	11a. lalok nok crrrok 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 .

Centauri/Arcturan

Your assignment, translate this to Arcturan: **farok** crrrok 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 anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok **farok** izok stok . 5b. totat jjat quat cat .	11a. lalok nok **crrrok** 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 .

Centauri/Arcturan

Your assignment, translate this to Arcturan: **farok** crrrok 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 anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok **hihok** ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok **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 .

Centauri/Arcturan

Your assignment, translate this to Arcturan: **farok** crrrok **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 anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok **hihok** ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok **yorok** ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok **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 .

Centauri/Arcturan

Your assignment, translate this to Arcturan: **farok** crrrok **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 anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok **clok** . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok 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 .

Centauri/Arcturan

Your assignment, translate this to Arcturan: **farok** crrrok **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 anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok **clok** . 10b. wat nnat gat mat bat hilat . ???
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok 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 .

Centauri/Arcturan

Your assignment, translate this to Arcturan: **farok** crrrok **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 anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok 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 .

Centauri/Arcturan

Your assignment, translate this to Arcturan: **farok** crrrok **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 anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat . process of elimination
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok 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

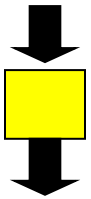
1999

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

2014

- 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".

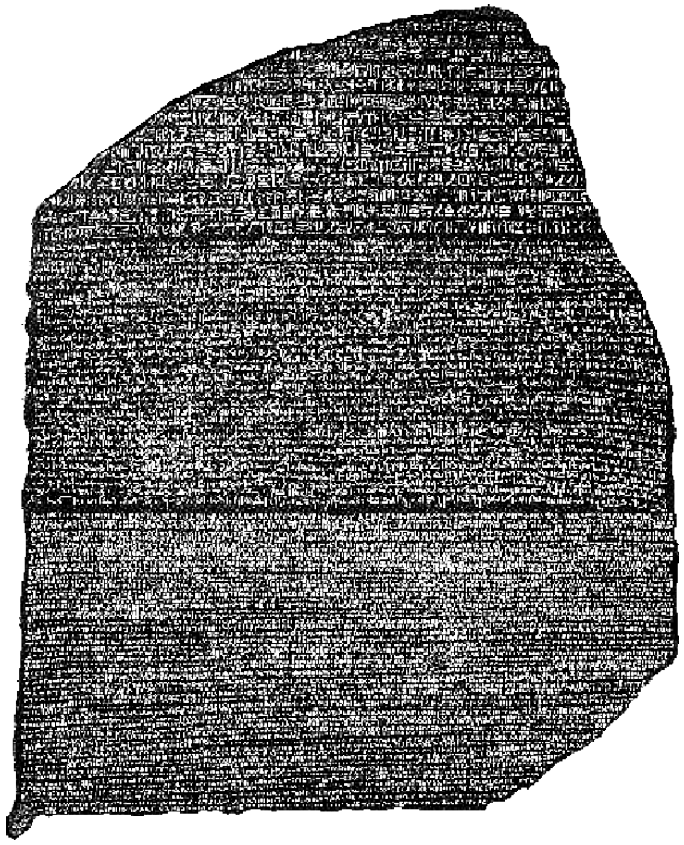


Somali President Sharif Sheikh Ahmed during a visit to sites of government forces and African in the district of Howden in Mogadishu that the current military campaign "will not stop until Somalia is liberated from the youth and al-Qaeda. "

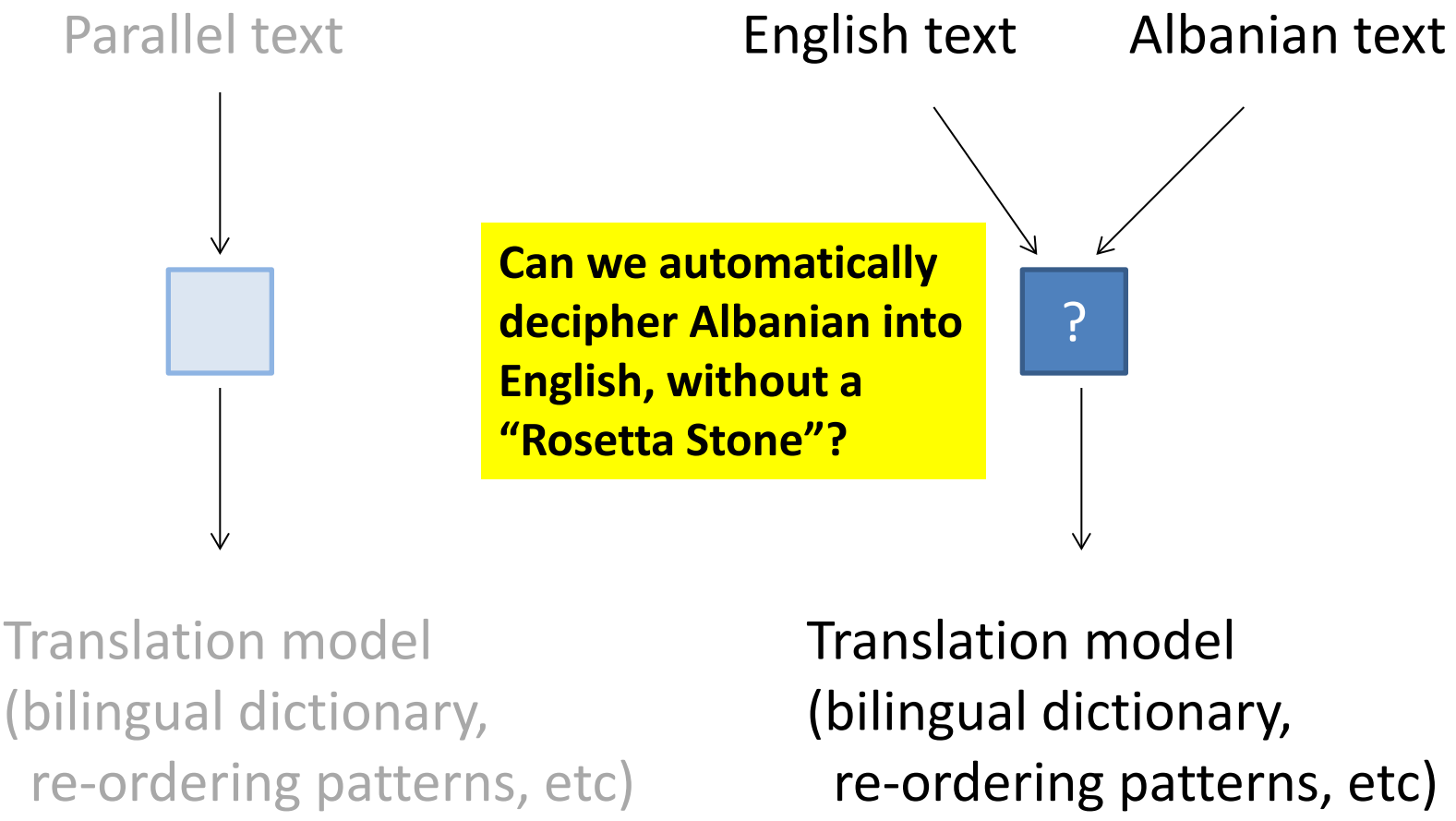
Learning Translation Knowledge from Parallel Text

Parallel text

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



Learning Translation Knowledge from Non-Parallel Text?

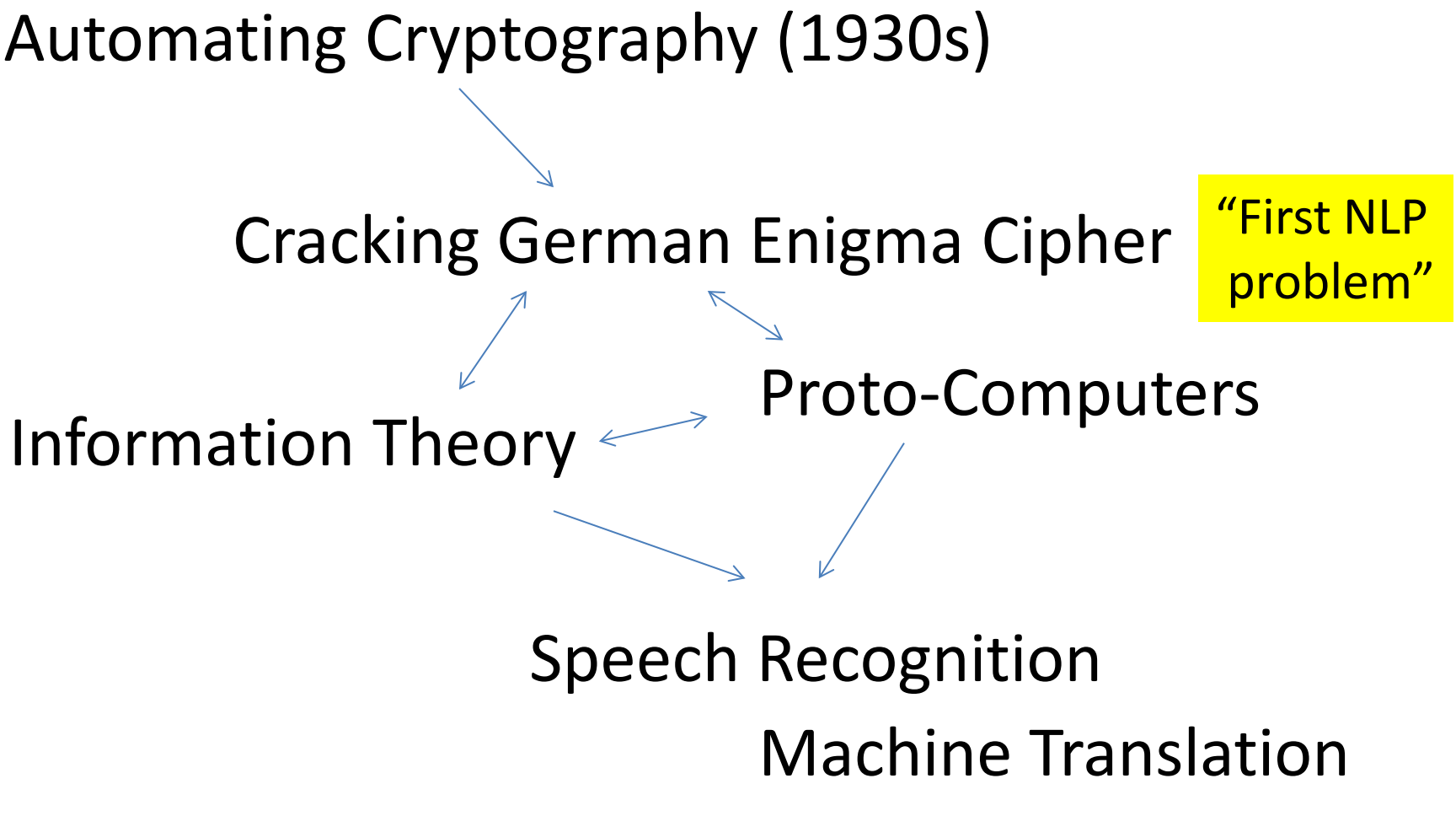


African Languages

Zero languages spoken

1000+ languages spoken, 40+ by 1m+ speakers

Translation as Decipherment? Our Cryptographic Roots



German Enigma Cipher Machine (1920s-1940s)

German text →  → over the radio:

DRKWI FKSLW
ORISJ DKSUE
IFKRP ...

→  → German text

input (intercepted ciphertext): **DRKWI FKSLW ORISJ DKSUE IFK …**
output (plaintext): **VASIS TDASH ERRCA PITAN RIC …**

German Enigma Cipher Machine

Substitution system

N → J

Substitution table **changes** with every keystroke:

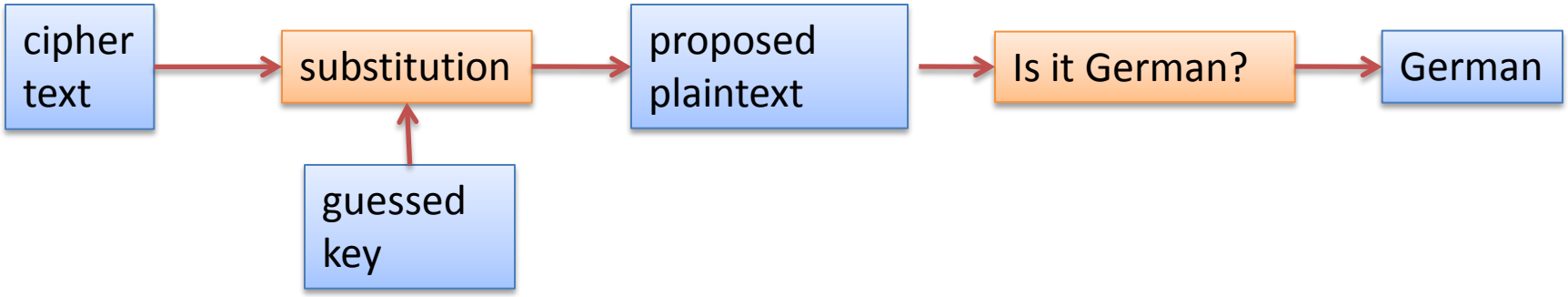
NNN → JTE

Secret key = initial rotor ordering and settings

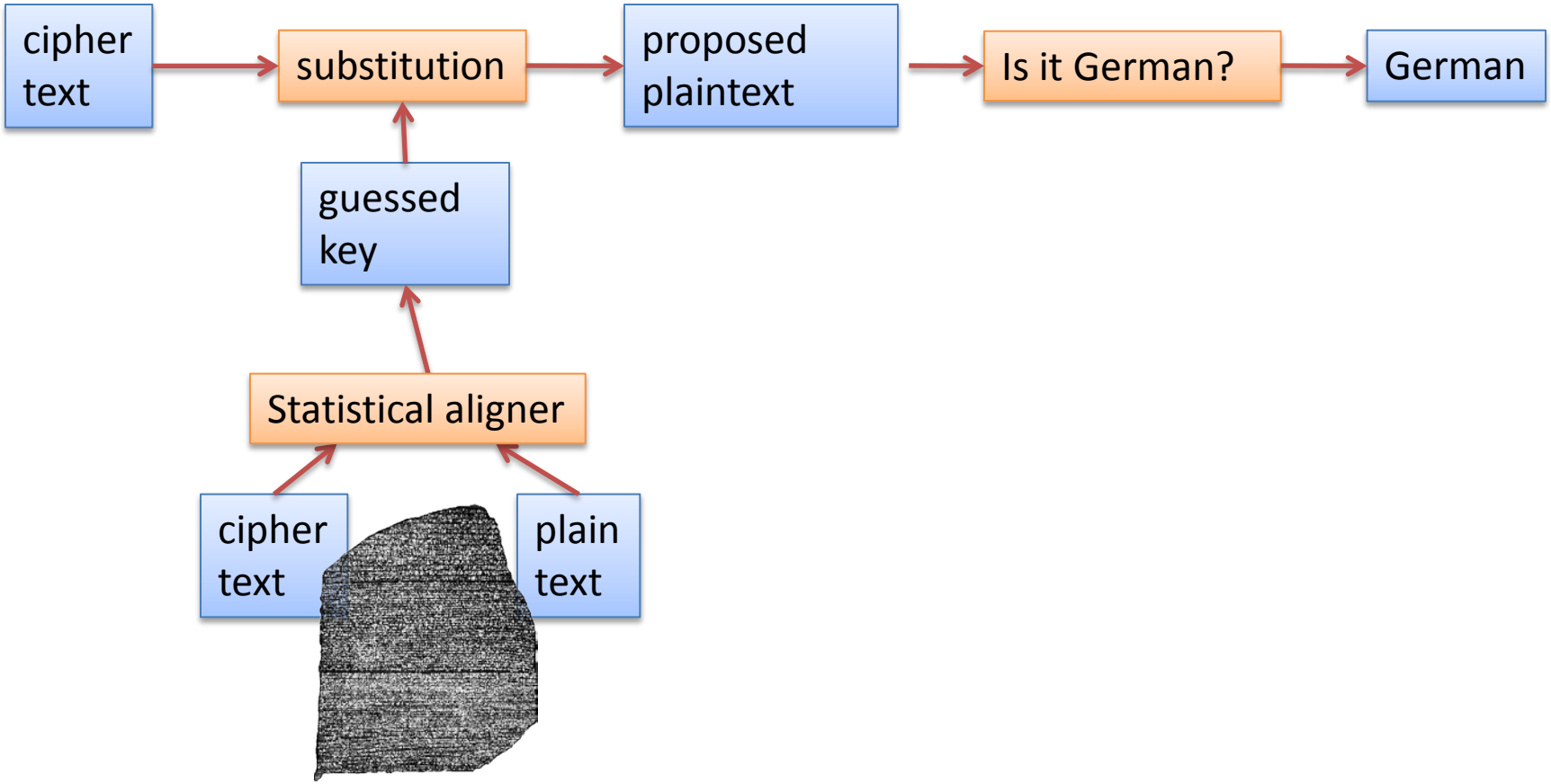
Reversible behavior

NNN → JTE → NNN

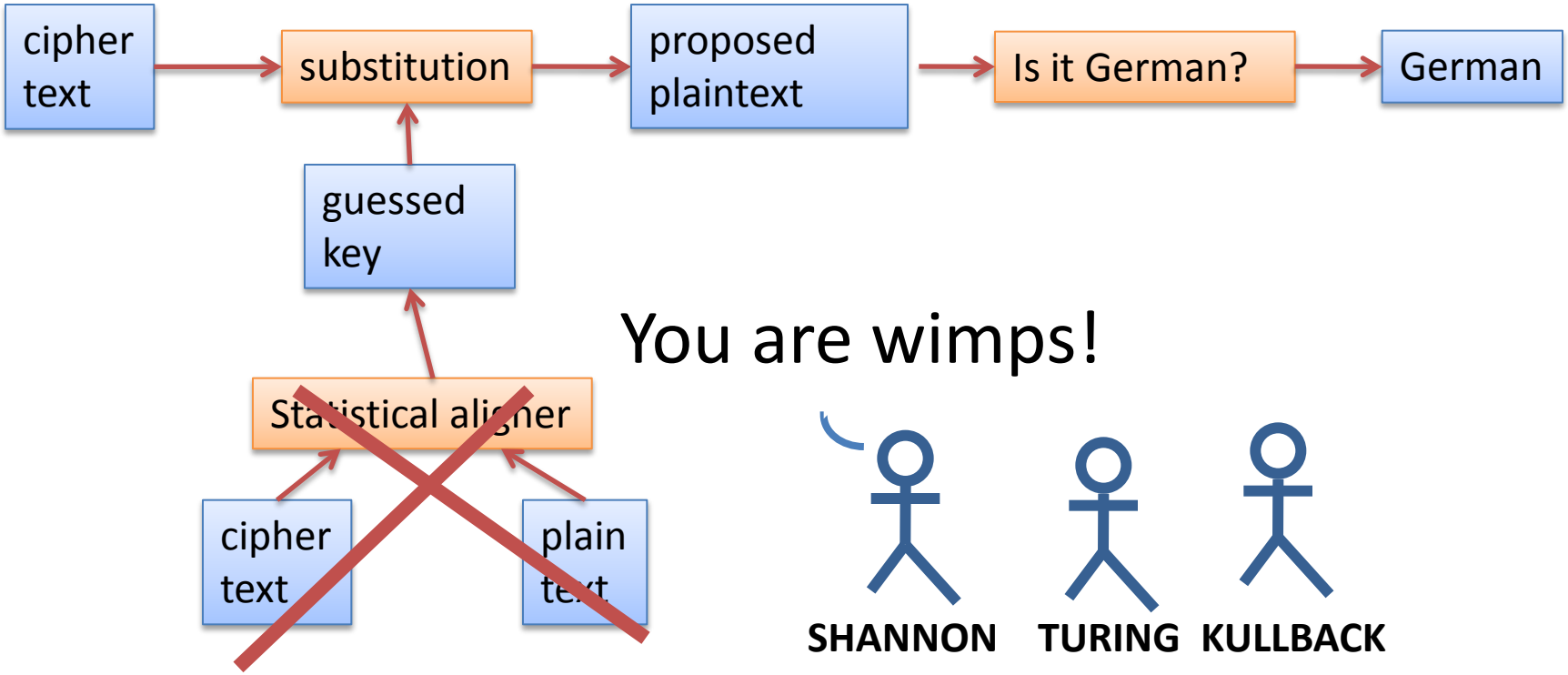
Breaking Enigma



Breaking Enigma

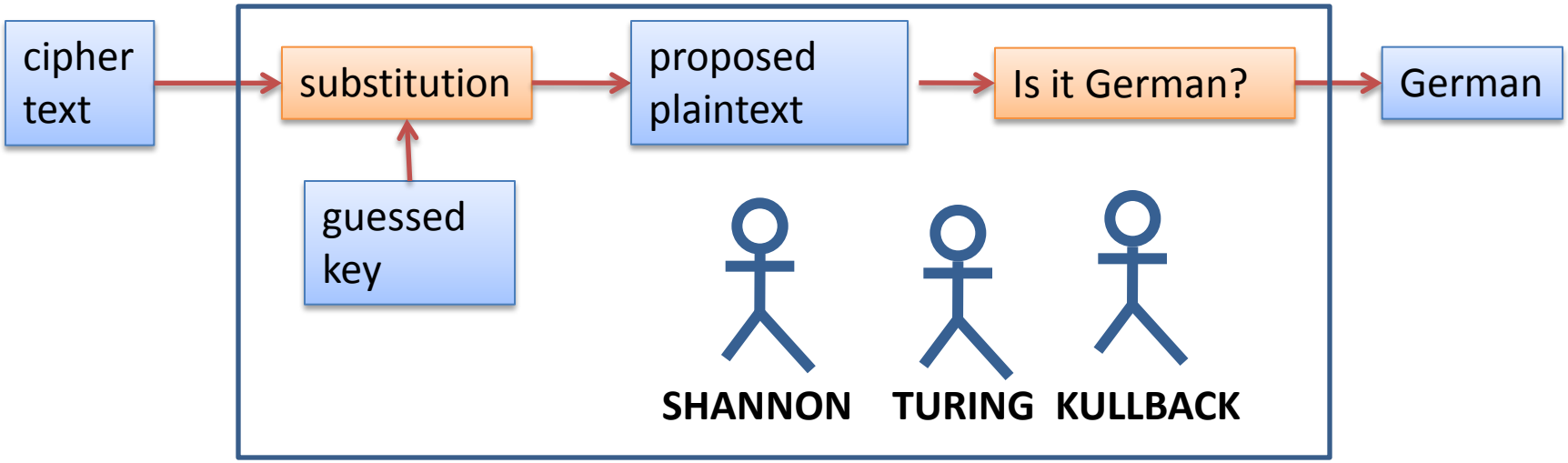


Breaking Enigma

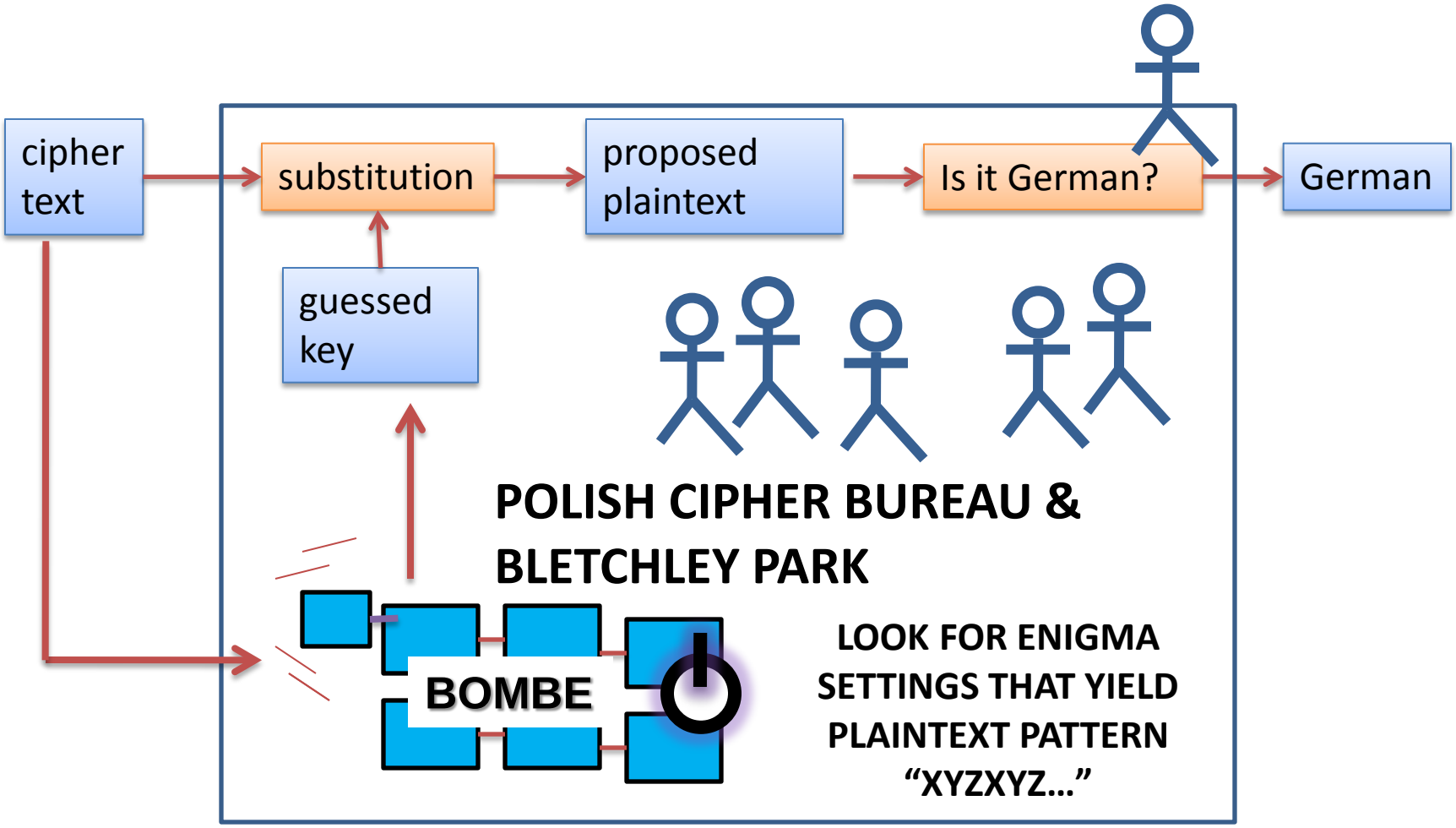


You are wimps!

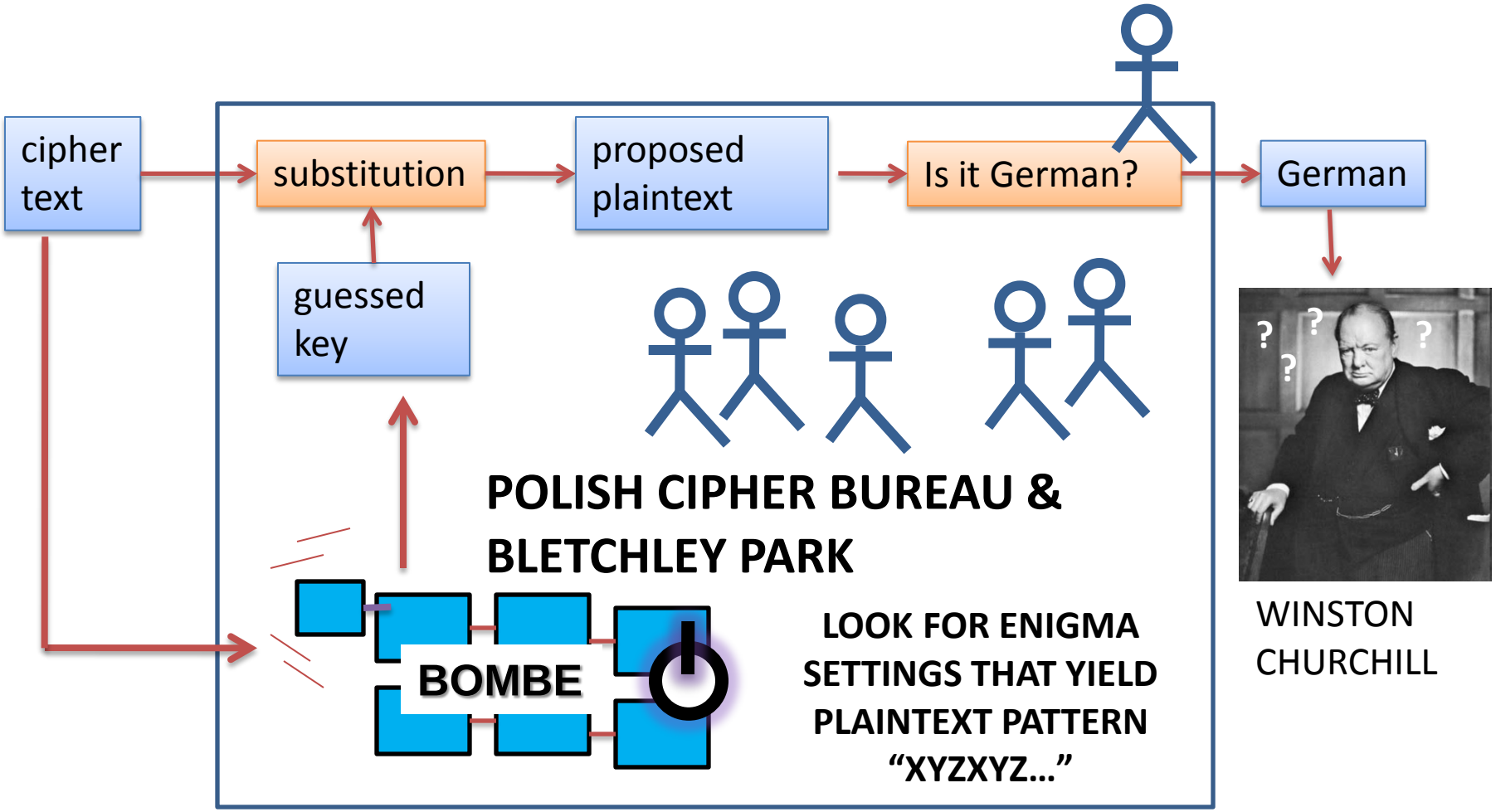
Breaking Enigma



Breaking Enigma



Breaking Enigma

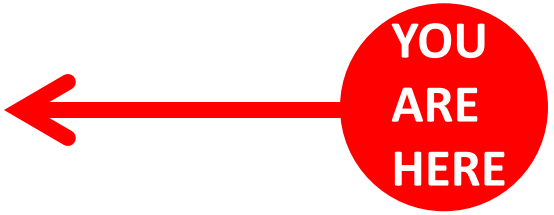


Plan for This Talk

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

Plan for This Talk

- Break a series of codes
  - 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:   PLOKMIJNUHBYGVTFCRDXESZAQW
```

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

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

KDCY LQZKTLJKX CY MDBCYJQL: “TR

HYD FKXC, FQ MKX RLQQIQ HYDL

MKL DXCTW RDCDLQ JQMNKXTMB

PTBMYEQL K FKH CY LQZKTL TC.”

KDCY LQZKTLJKX CY MDBCYJQL: “TR

HYD FKXC, FQ MKX RLQQIQ HYDL

MKL DXCTW RDCDLQ JQMNKXTMB

PTBMYEQL K FKH CY LQZKTL TC.”

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

. . . .

KDCY LQZKTLJKX CY MDBCYJQL: “TR

. . . . . .

HYD FKXC, FQ MKX RLQQIQ HYDL

. . . . .

MKL DXCTW RDCDLQ JQMNKXTMB

. . . . .

PTBMYEQL K FKH CY LQZKTL TC.”

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

. . . .

KDCY LQZKTLJKX CY MDBCYJQL: “TR

. . . . . .

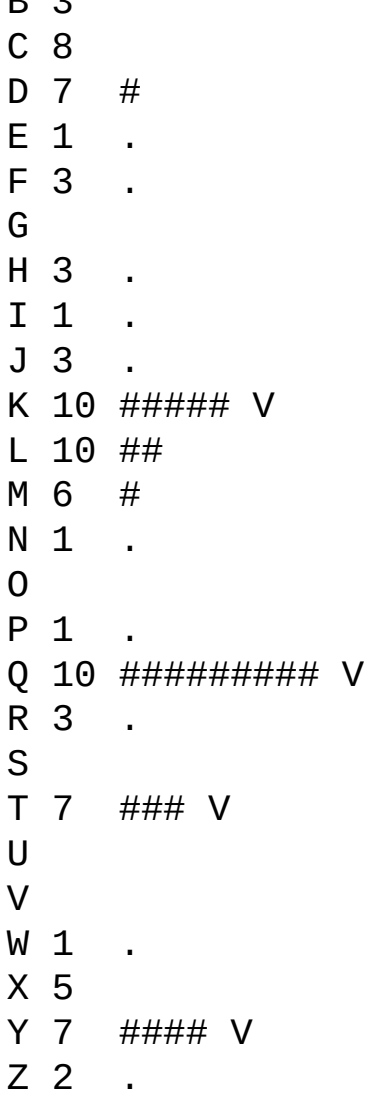
HYD FKXC, FQ MKX RLQQIQ HYDL

. . . . .

MKL DXCTW RDCDLQ JQMNKXTMB

. . . . .

PTBMYEQL K FKH CY LQZKTL TC.”



a .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  .
```

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
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
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  .
```

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 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 want we can freeze your

HYD FKXC, FQ MKX RLQQIQ HYDL

car until future mechanics

MKL DXCTW RDCDLQ JQMNKXTMB

discover a way to repair it

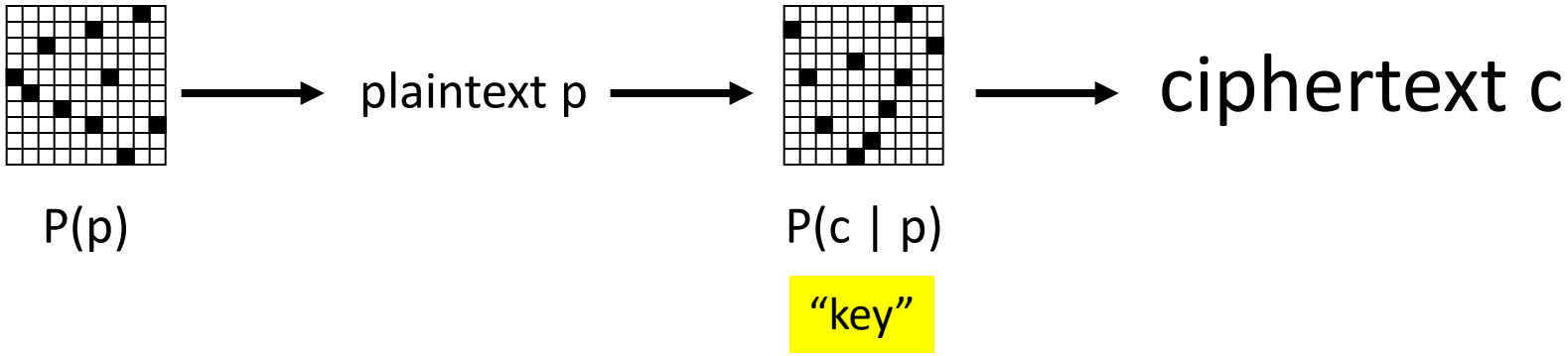
PTBMYEQL K FKH CY LQZKTL TC.”

```
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

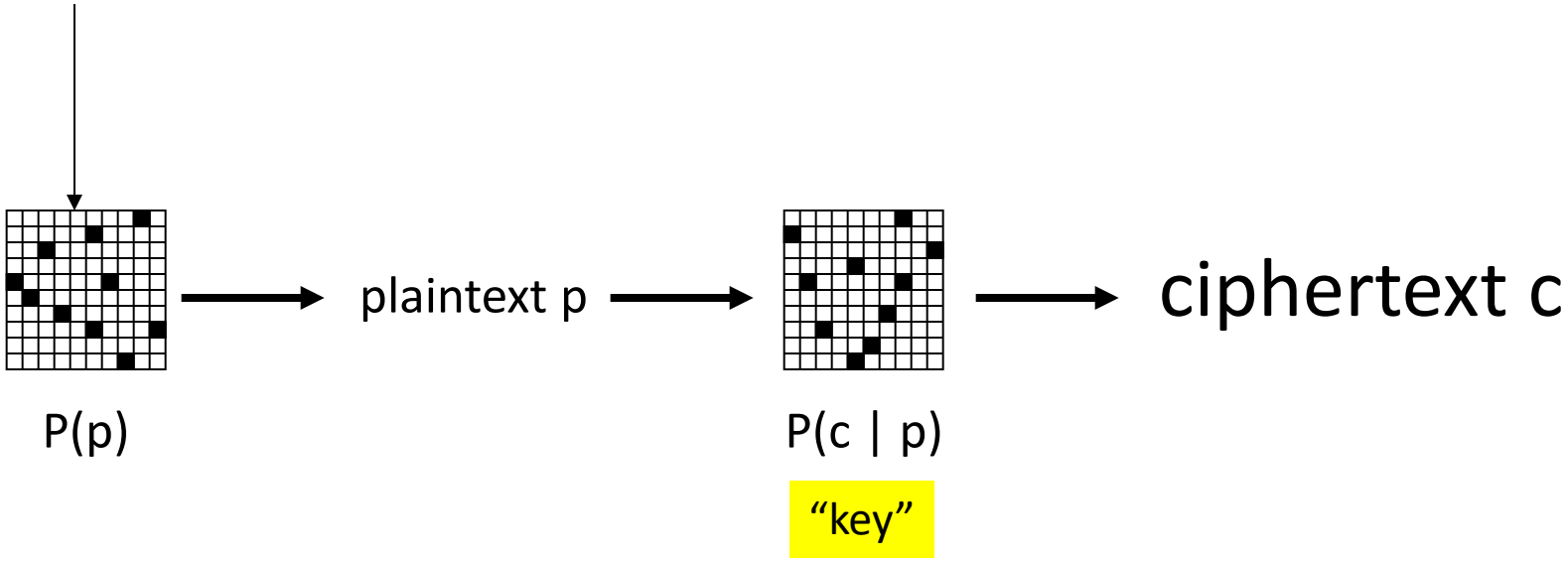
ciphertext c

Letter Substitution Cipher



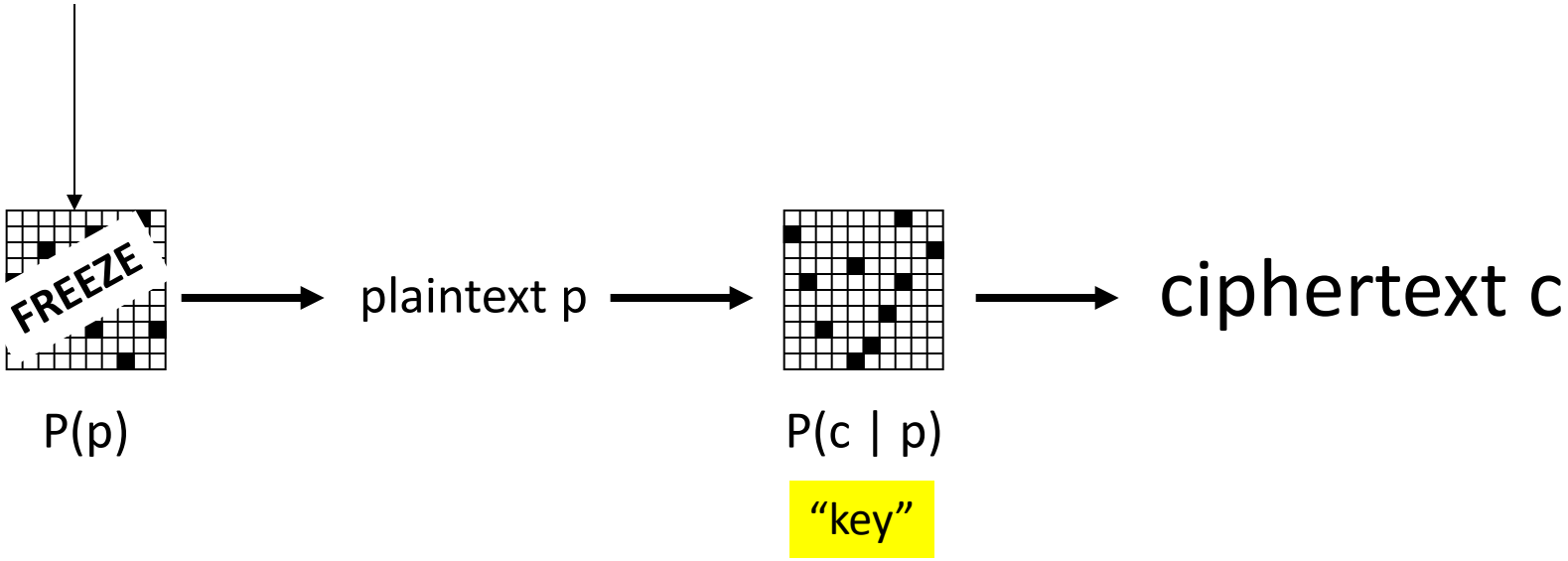
Letter Substitution Cipher

plaintext samples,
unrelated to ciphertext

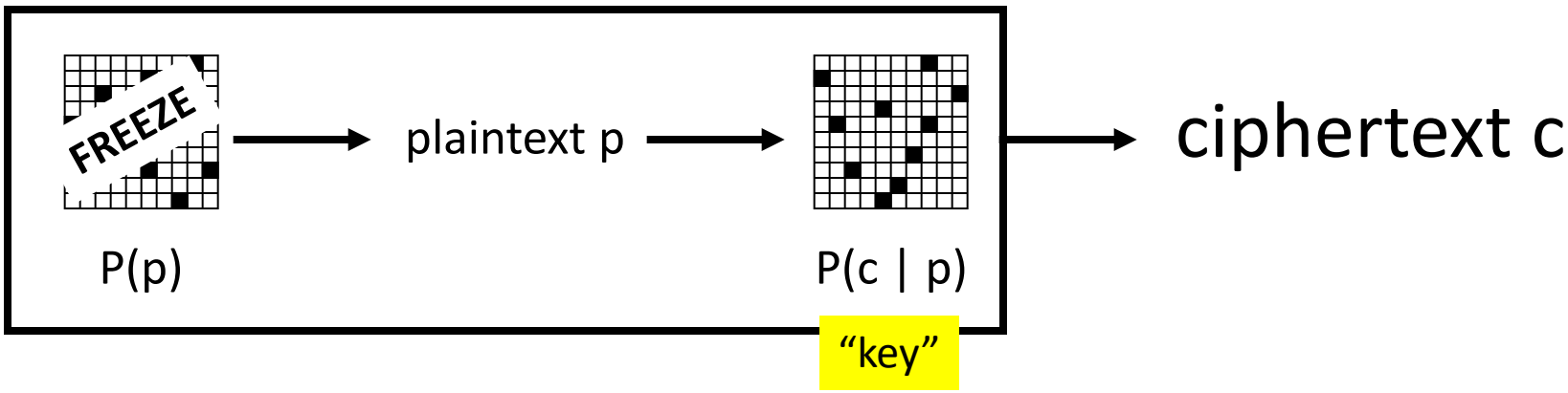


Letter Substitution Cipher

plaintext samples,
unrelated to ciphertext



Letter Substitution Cipher

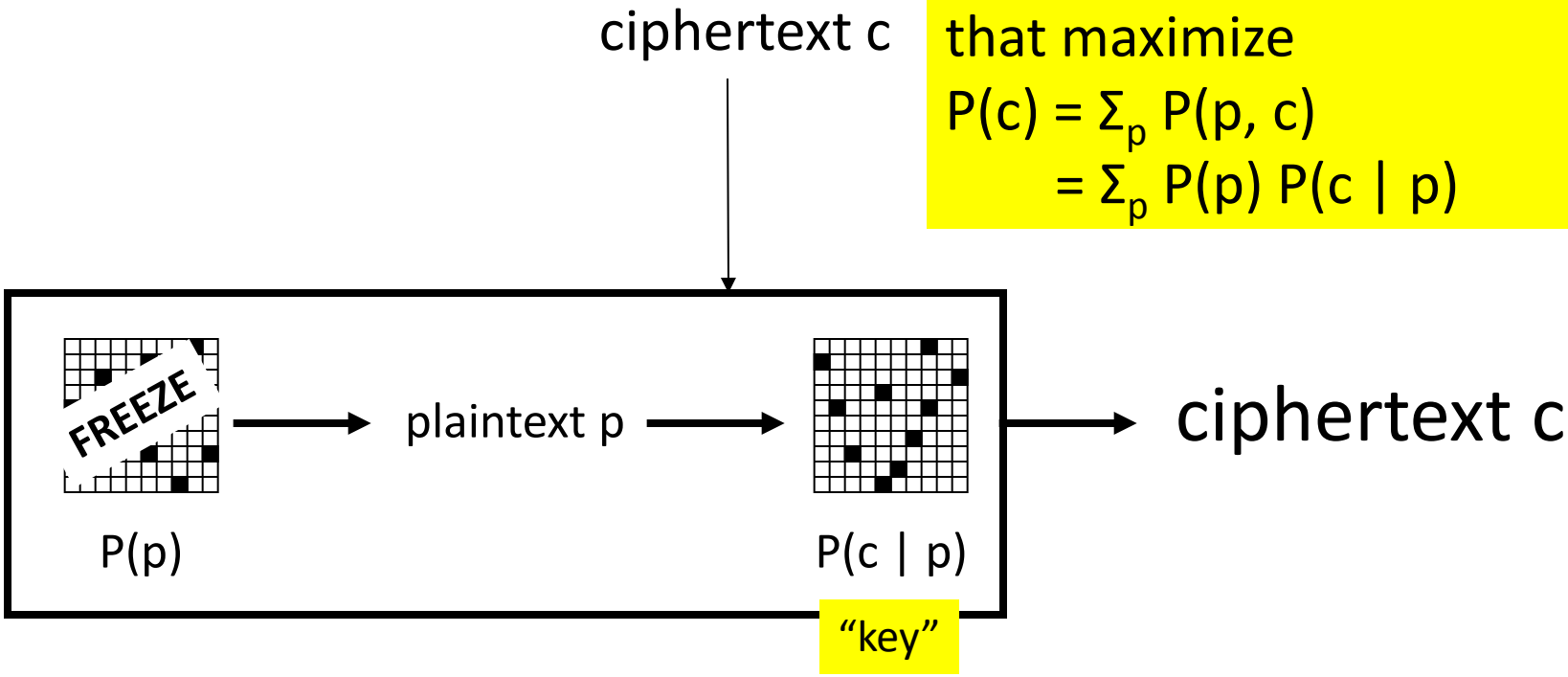


Letter Substitution Cipher

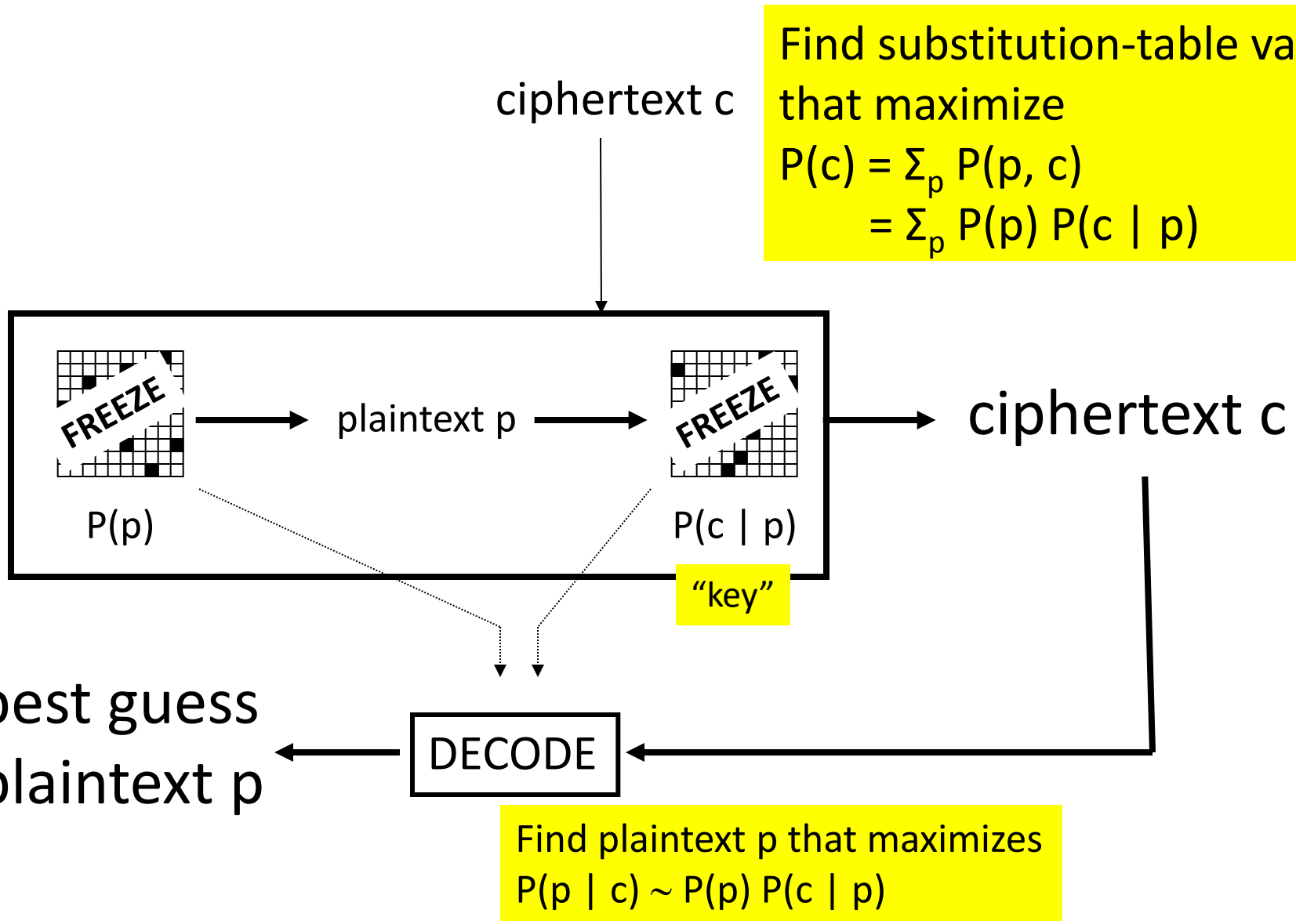
Find substitution-table values that maximize

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

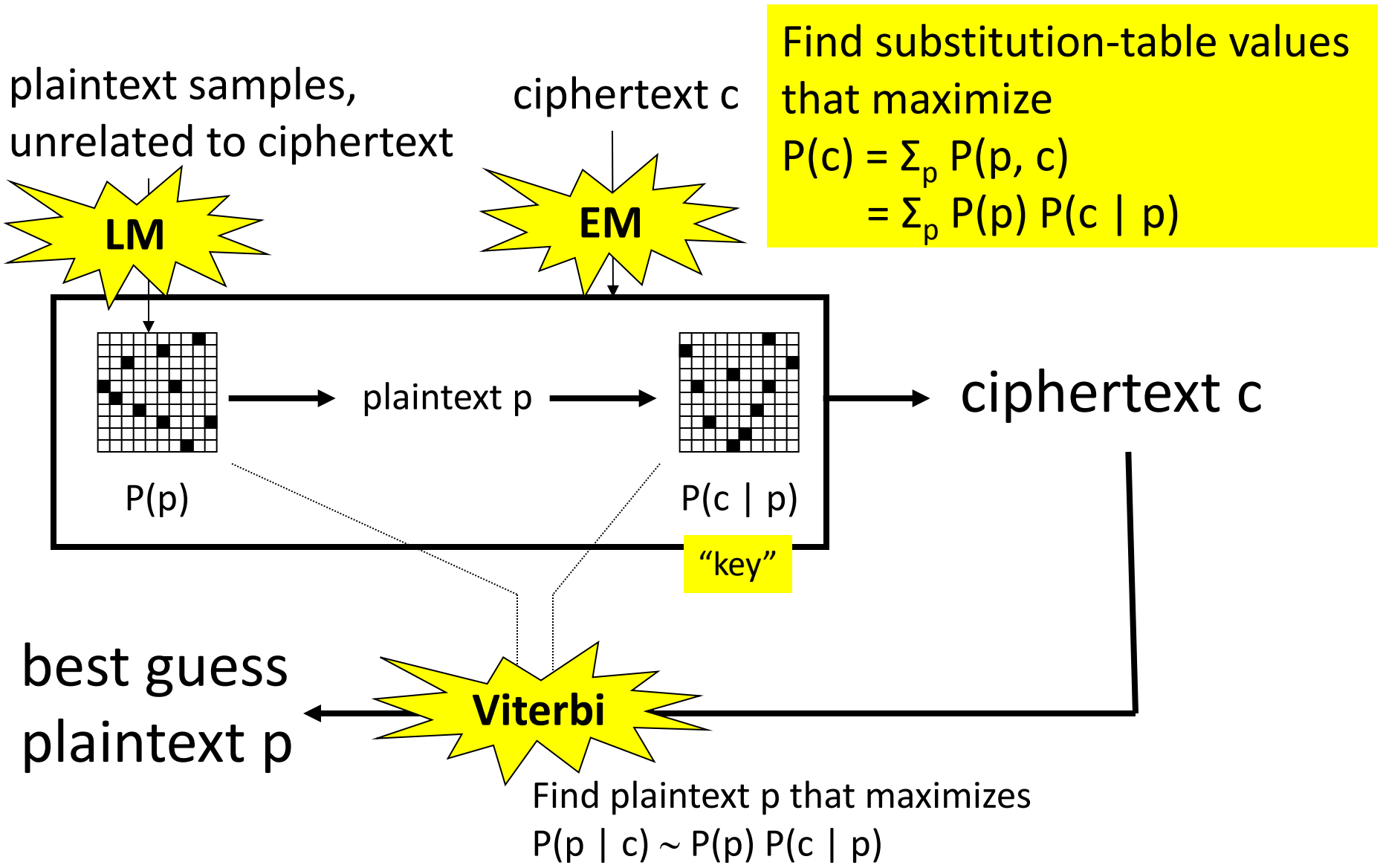
$= \Sigma_p P(p) P(c \mid p)$



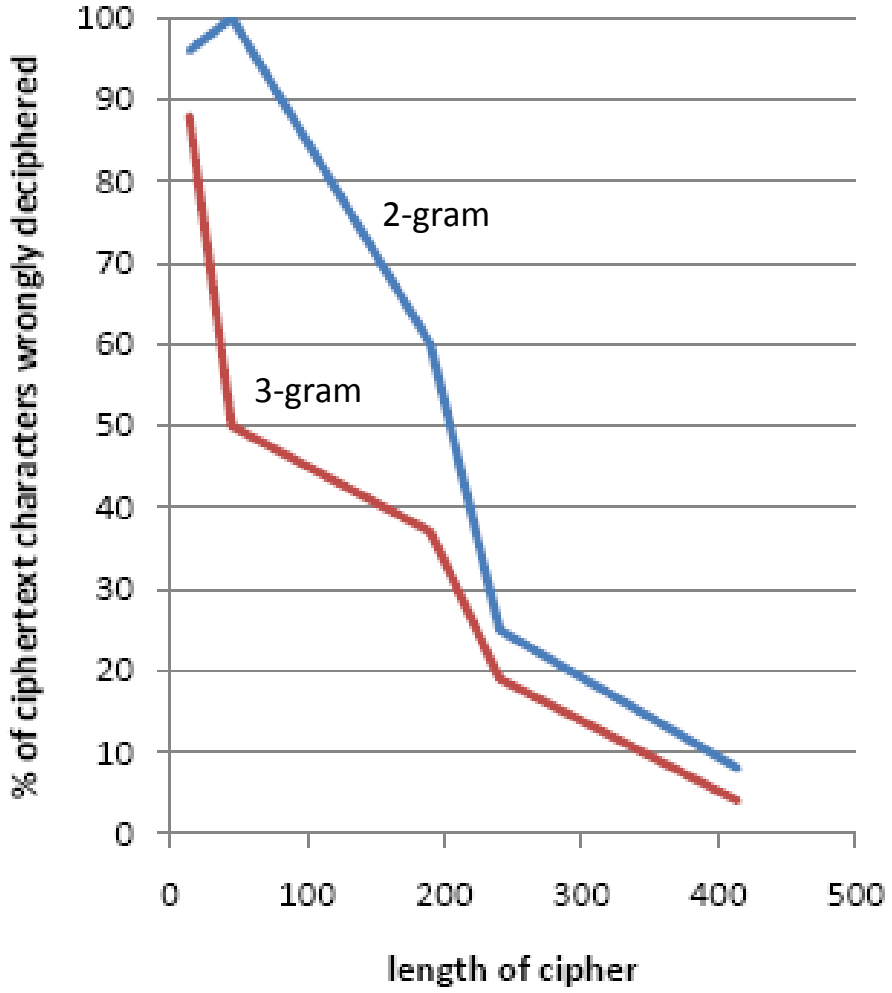
Letter Substitution Cipher



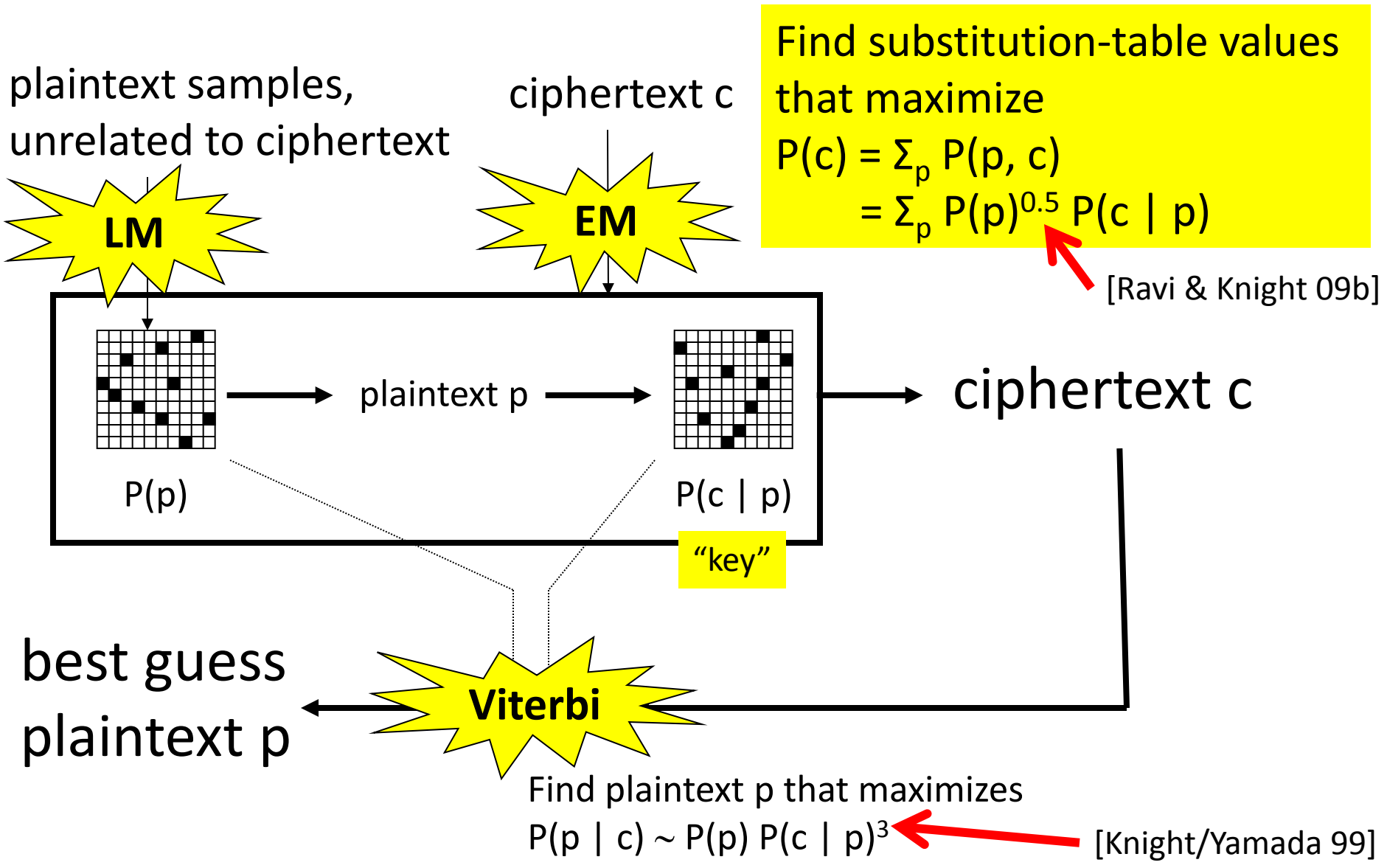
Letter Substitution Cipher



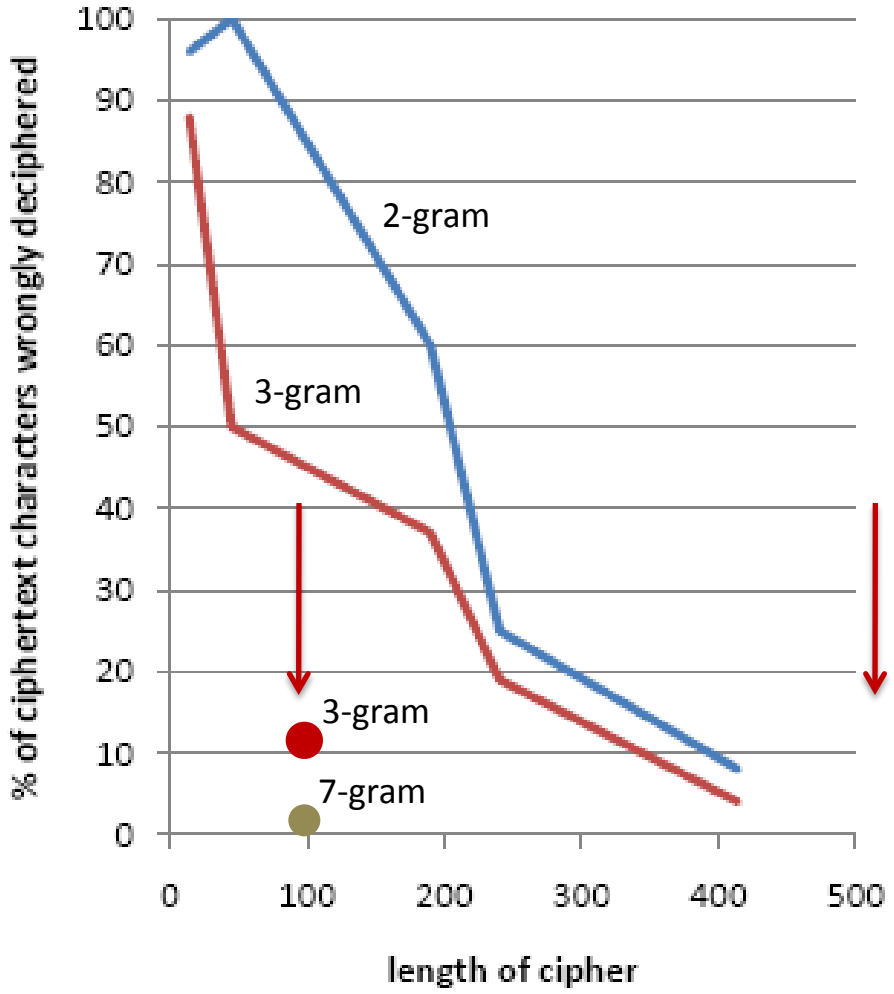
Decipherment Accuracy vs. Cipher Length



Letter Substitution Cipher



Reducing LM Weight During EM



Set EM to maximize

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

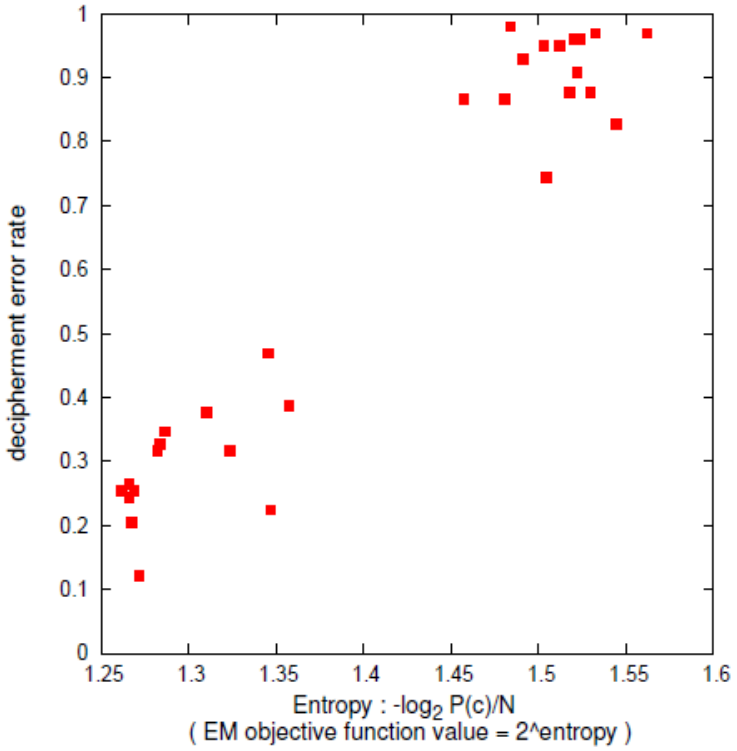
instead of

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

[Ravi & Knight 09b]

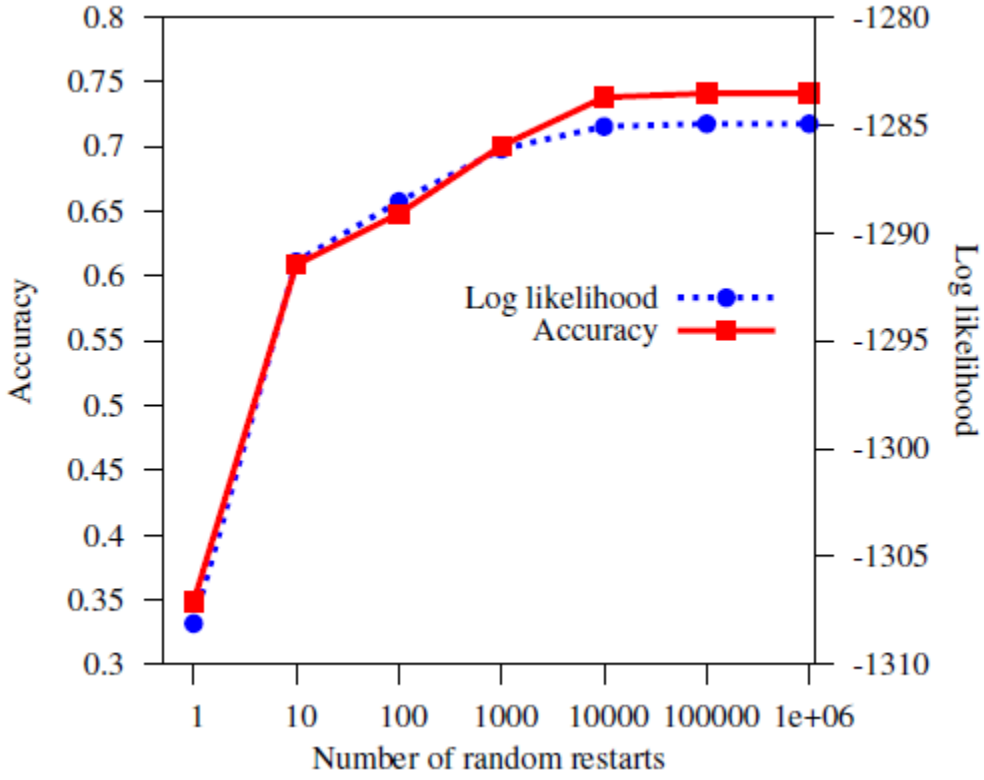
Random Restarts are Critical

English 98-letter cipher, 3-gram LM



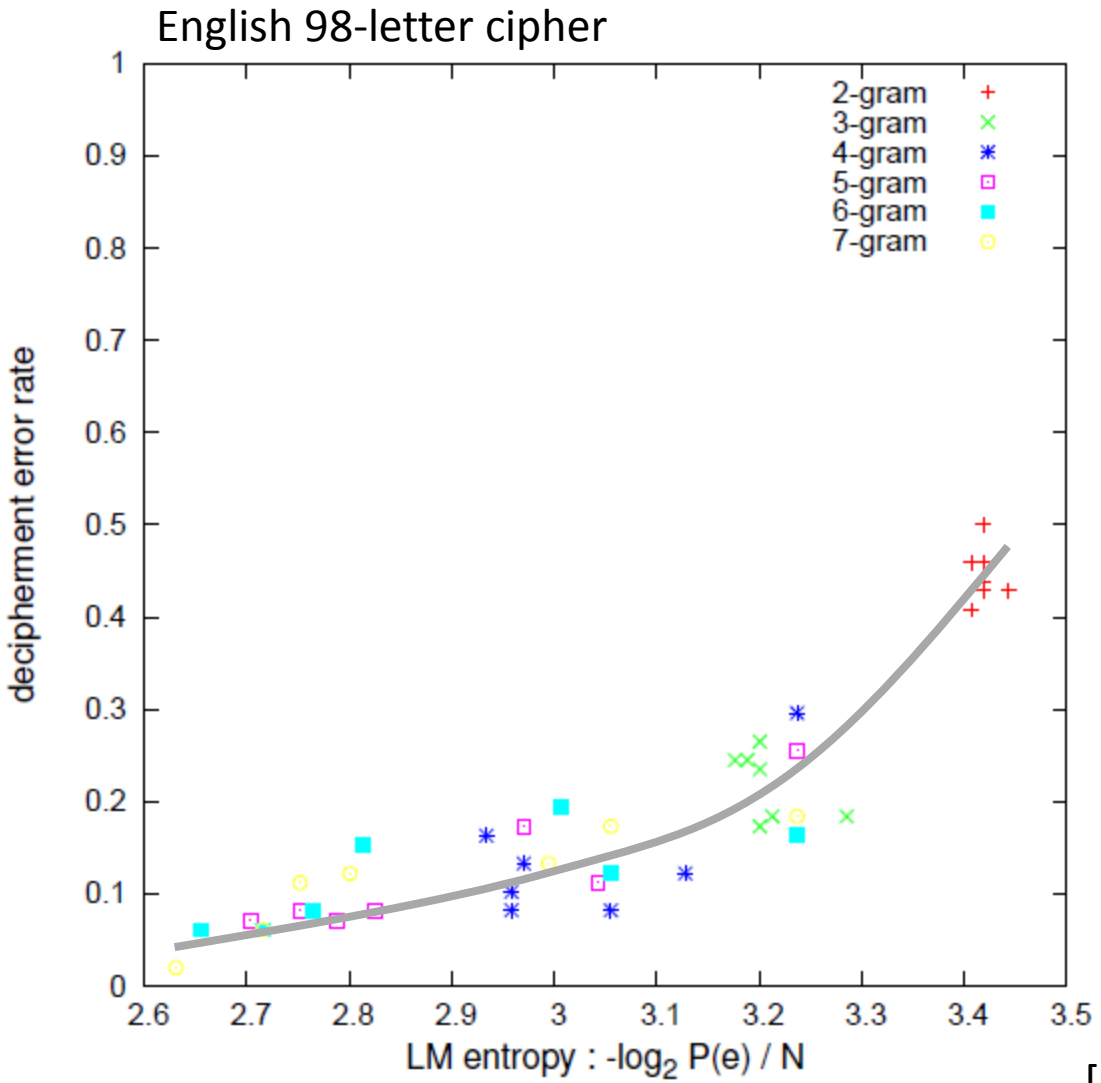
[Ravi & Knight 09b]

Zodiac killer 340 cipher



[Berg-Kirkpatrick & Klein 13]

Good Language Models are Critical



[Ravi & Knight 09b]

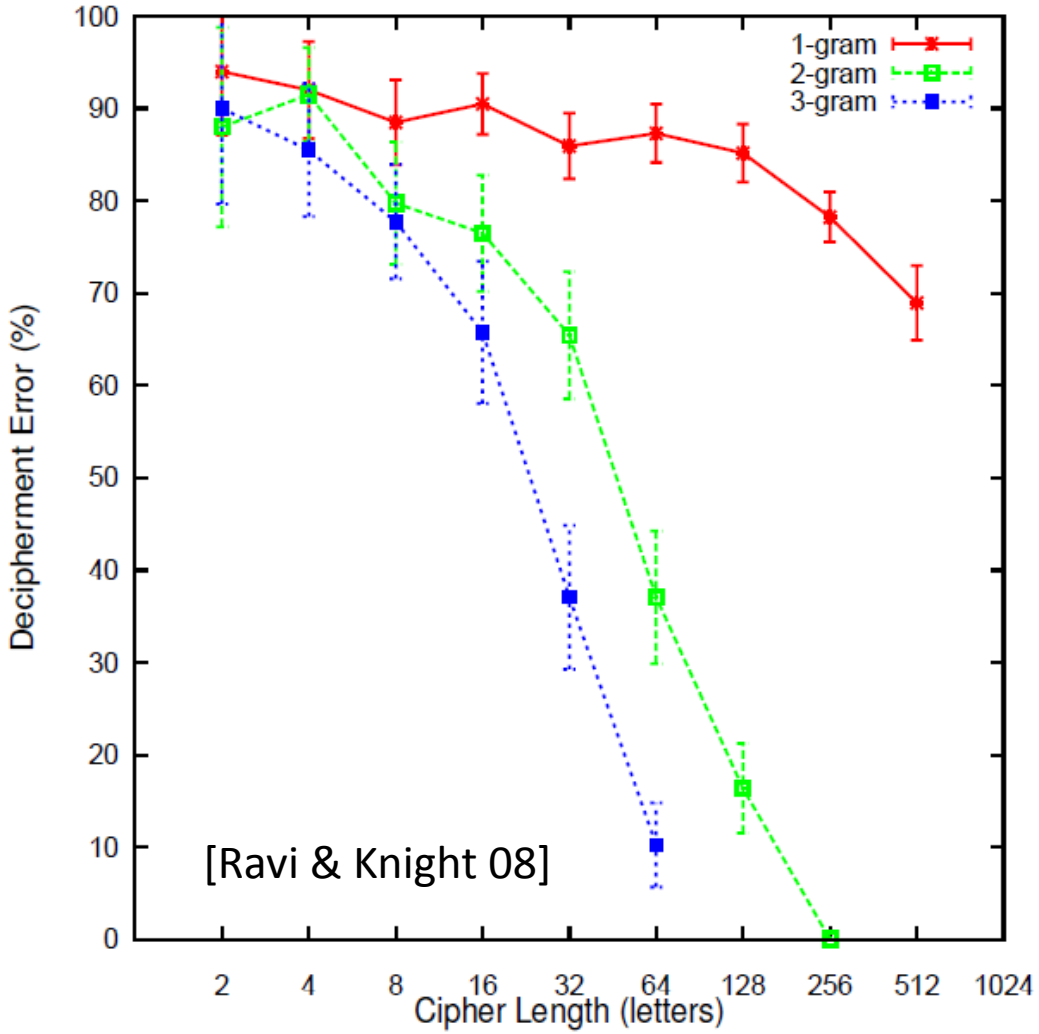
Deterministic Substitution Constraint

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 %**

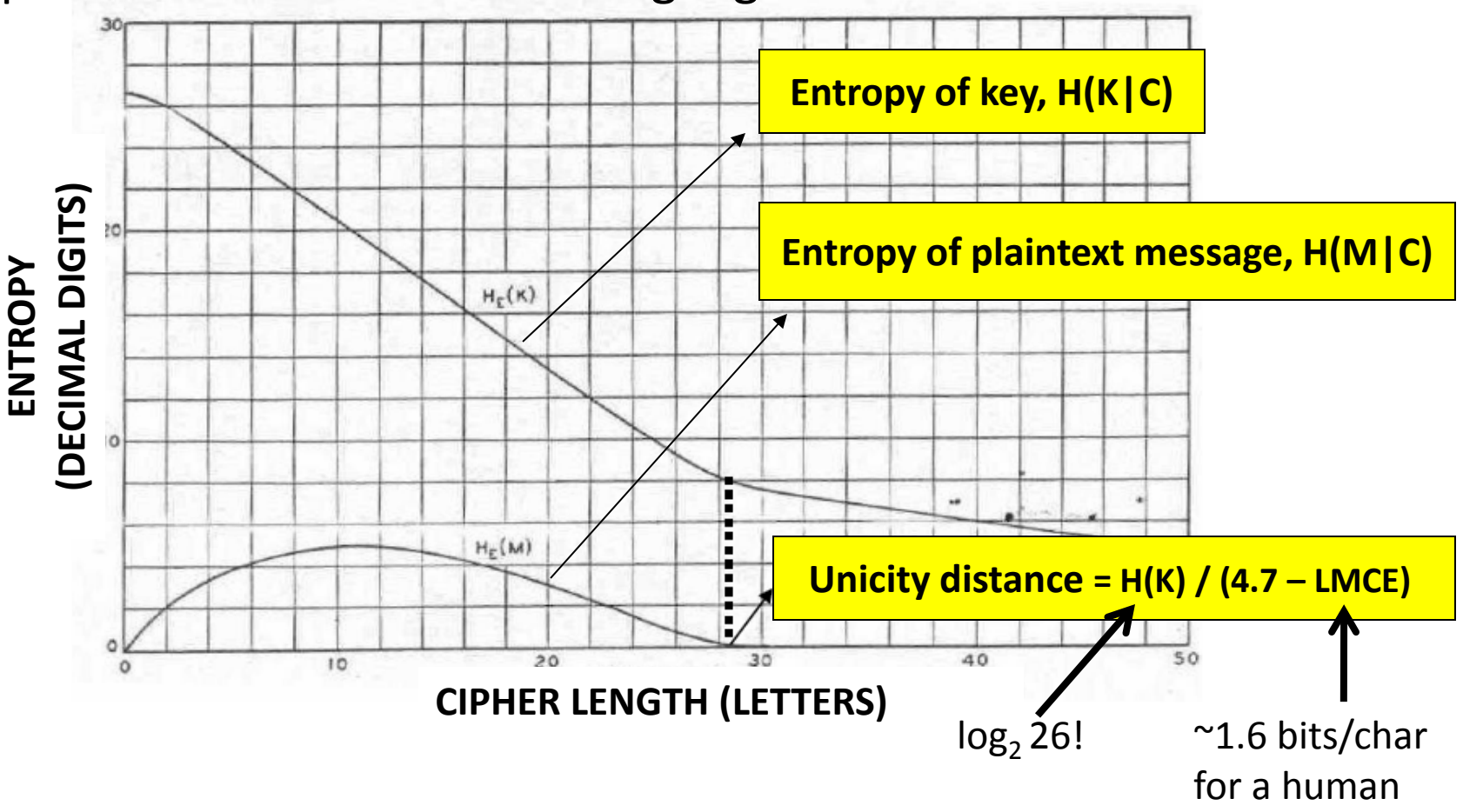
Using 2-gram letter-based LM



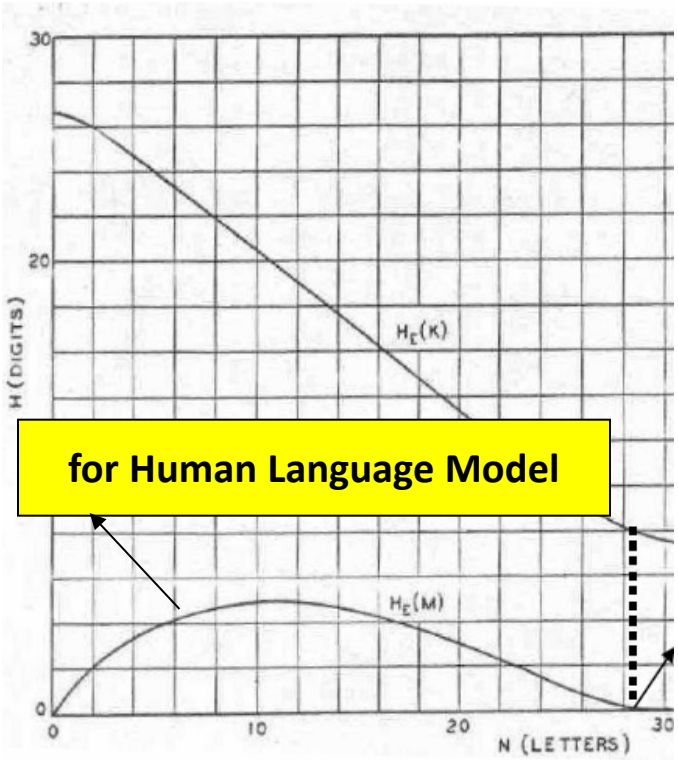
see also [Nuhn, Ney, Schamper] on beam search

[Shannon 46, 49]
"Communication Theory of Secrecy Systems"

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

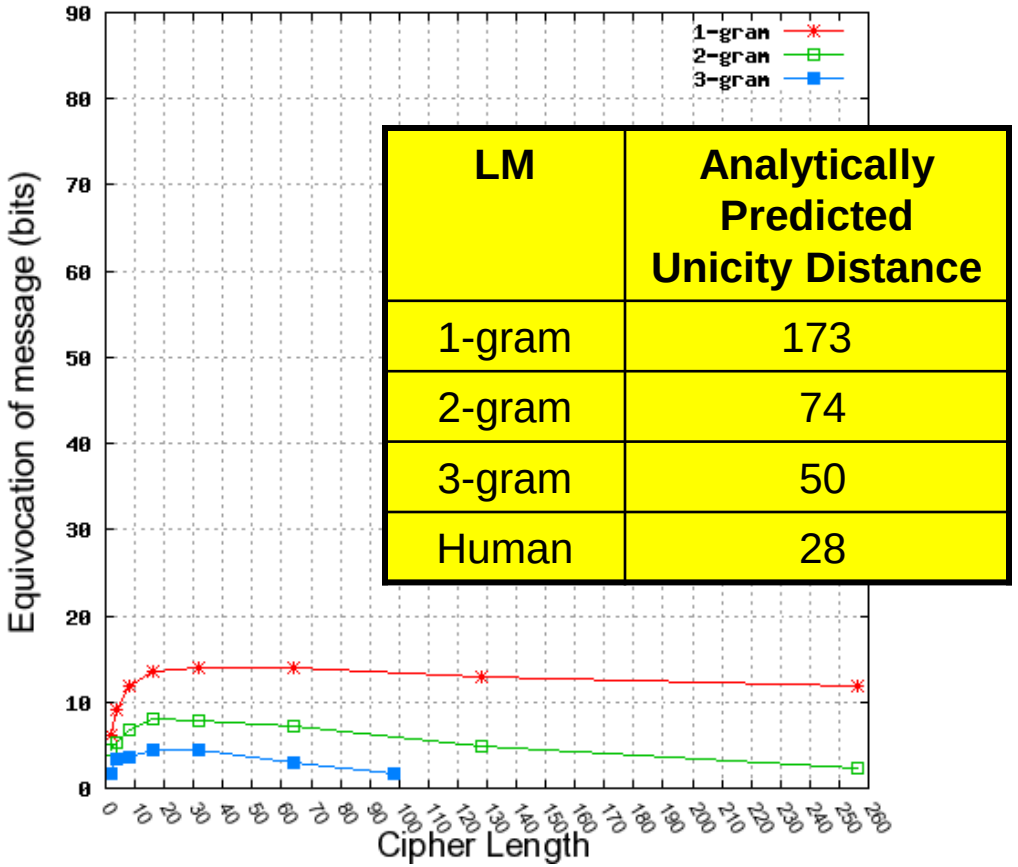


Verifying Shannon's Prediction of Plaintext Message Uncertainty



for Human Language Model

ANALYTIC CURVES
(Shannon's)



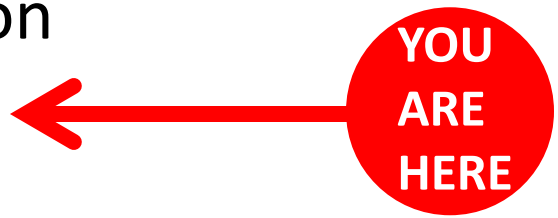
LM	Analytically Predicted Unicity Distance
1-gram	173
2-gram	74
3-gram	50
Human	28

ACTUAL CURVES
(ours)

[Ravi & Knight 08]

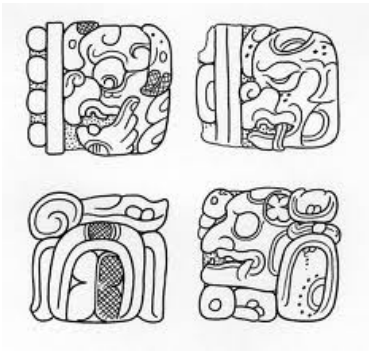
Plan for This Talk

- 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



[Knight & Yamada 99]

Phonetic Decipherment

ciphertext

primera parte
del ingenioso
hidalgo don …

[Knight & Yamada 99]

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

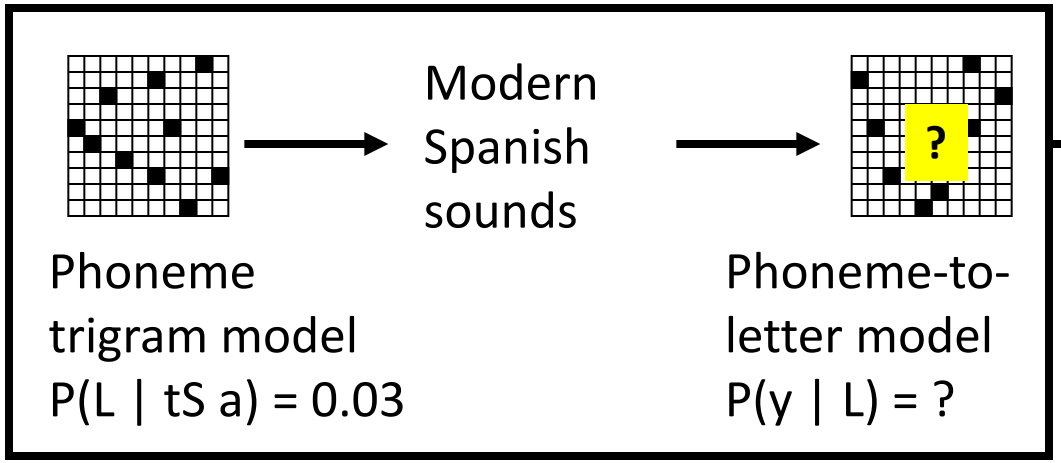
ciphertext

primera parte del ingenioso hidalgo don …

(Don Quixote)

[Knight & Yamada 99]

Phonetic Decipherment



ciphertext

primera parte del ingenioso hidalgo don …

(Don Quixote)

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

↔ ? ↔

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

EM approach = 93% accurate phonetic decipherment

[Knight & Yamada 99]

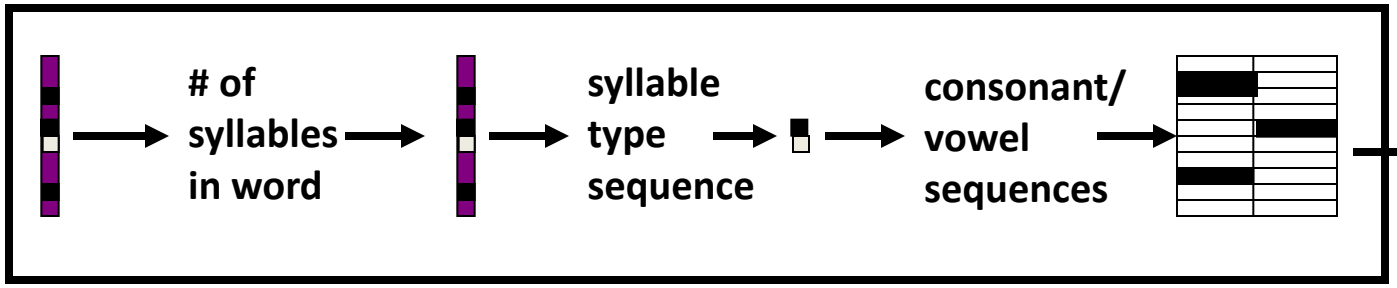
What if Spoken Language Behind Script is Unknown?

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

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

[Knight et al 06]

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.

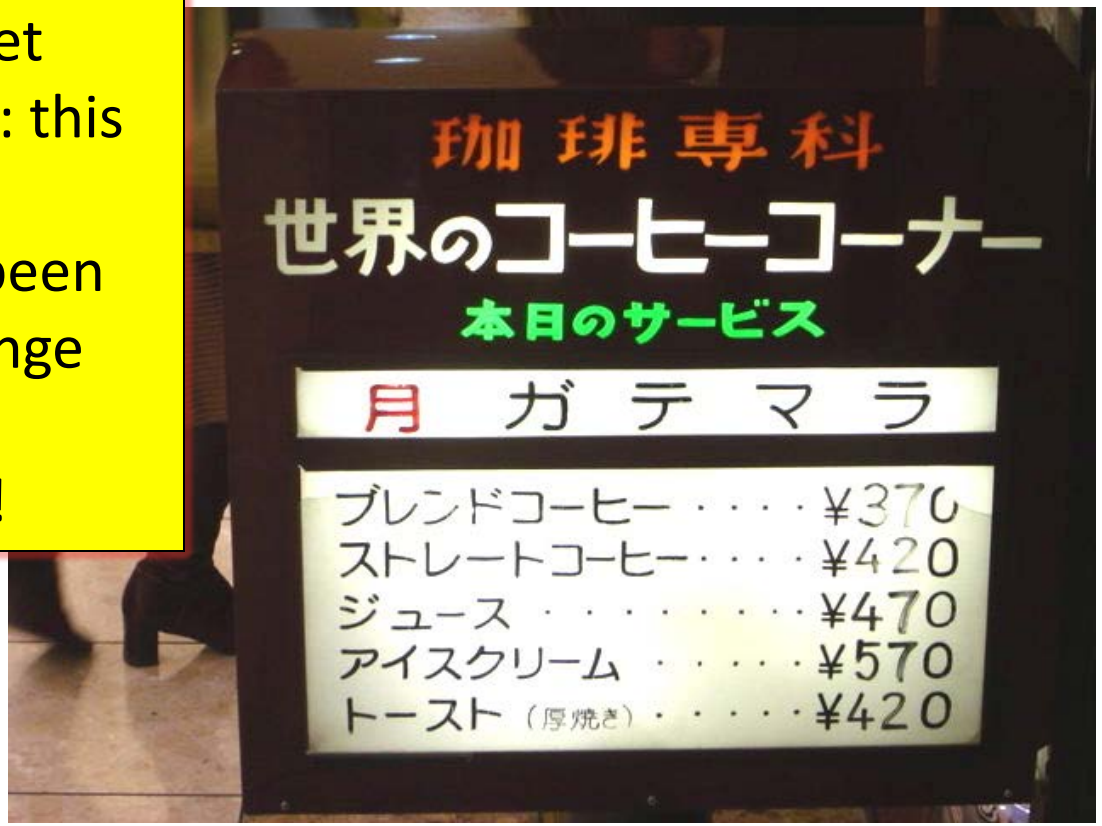
Input: **primera parte del ingenioso …**
Output: **NSV.NV.NV NVS.NV NVS VS.NV.SV.V.NV …**

S = sonorous consonant phoneme
N = non-sonorous consonant phoneme
V = vowel phoneme

[Knight et al 06]
see also
[Kim & Snyder 13]

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!



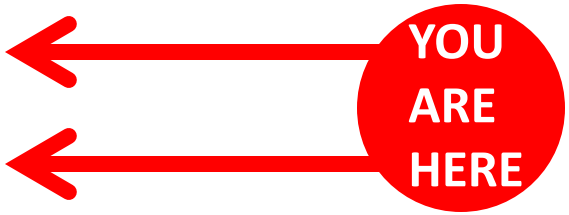
OUR HERO

Parallel data: [Knight & Graehl 97]

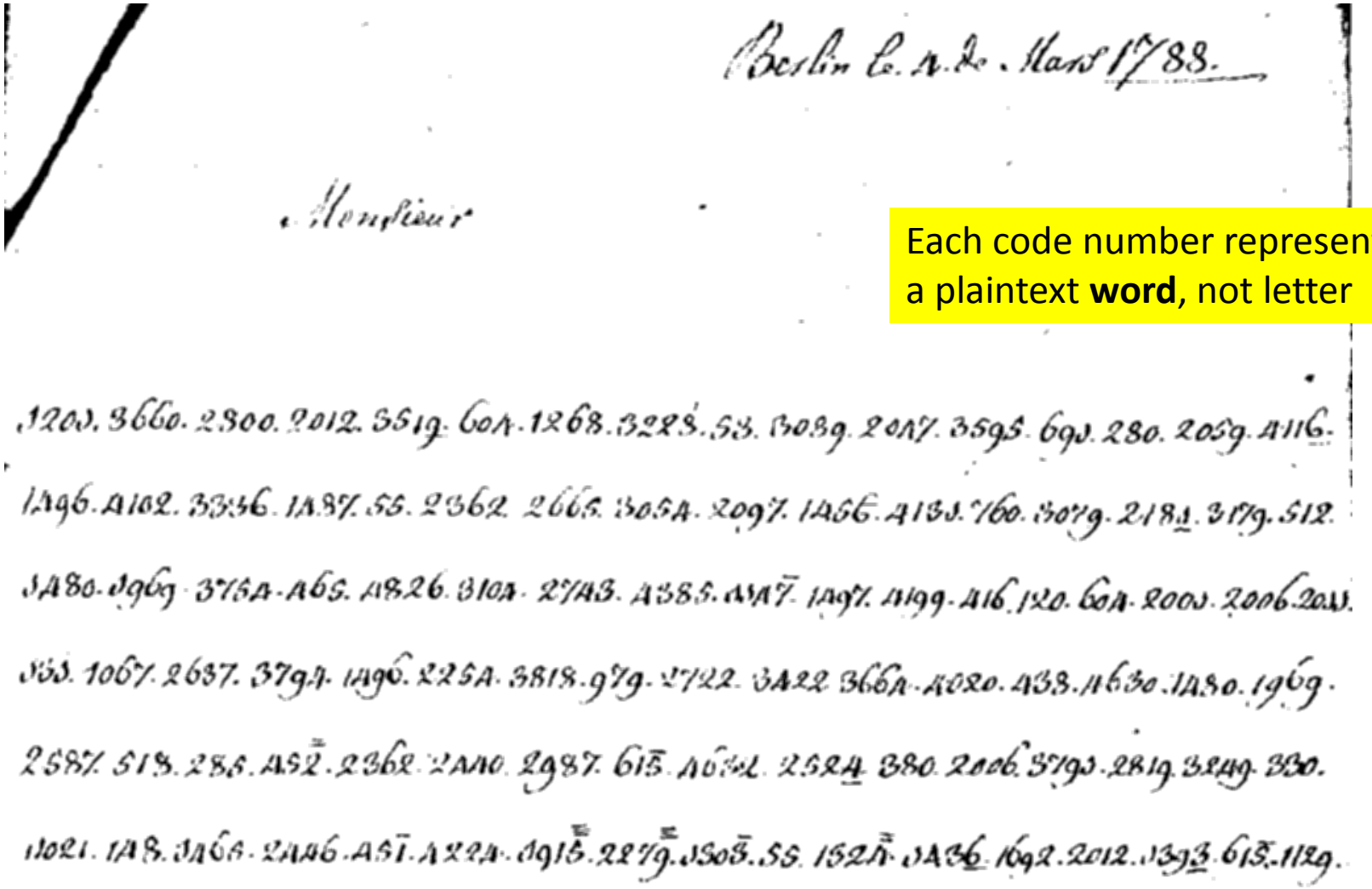
Non-parallel data: [Ravi & Knight 09a]

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



Berlin le 2. Mars 1788.

Monsieur

Each code number represents a plaintext **word**, not letter

1203. 3660. 2300. 2012. 3519. 60A. 1268. 3228. 53. 3089. 2017. 3595. 693. 280. 2059. 4116.

1A96. A102. 3336. 1A87. 55. 2362. 2665. 305A. 2097. 1A5E. A133. 760. 3079. 2181. 3179. 512.

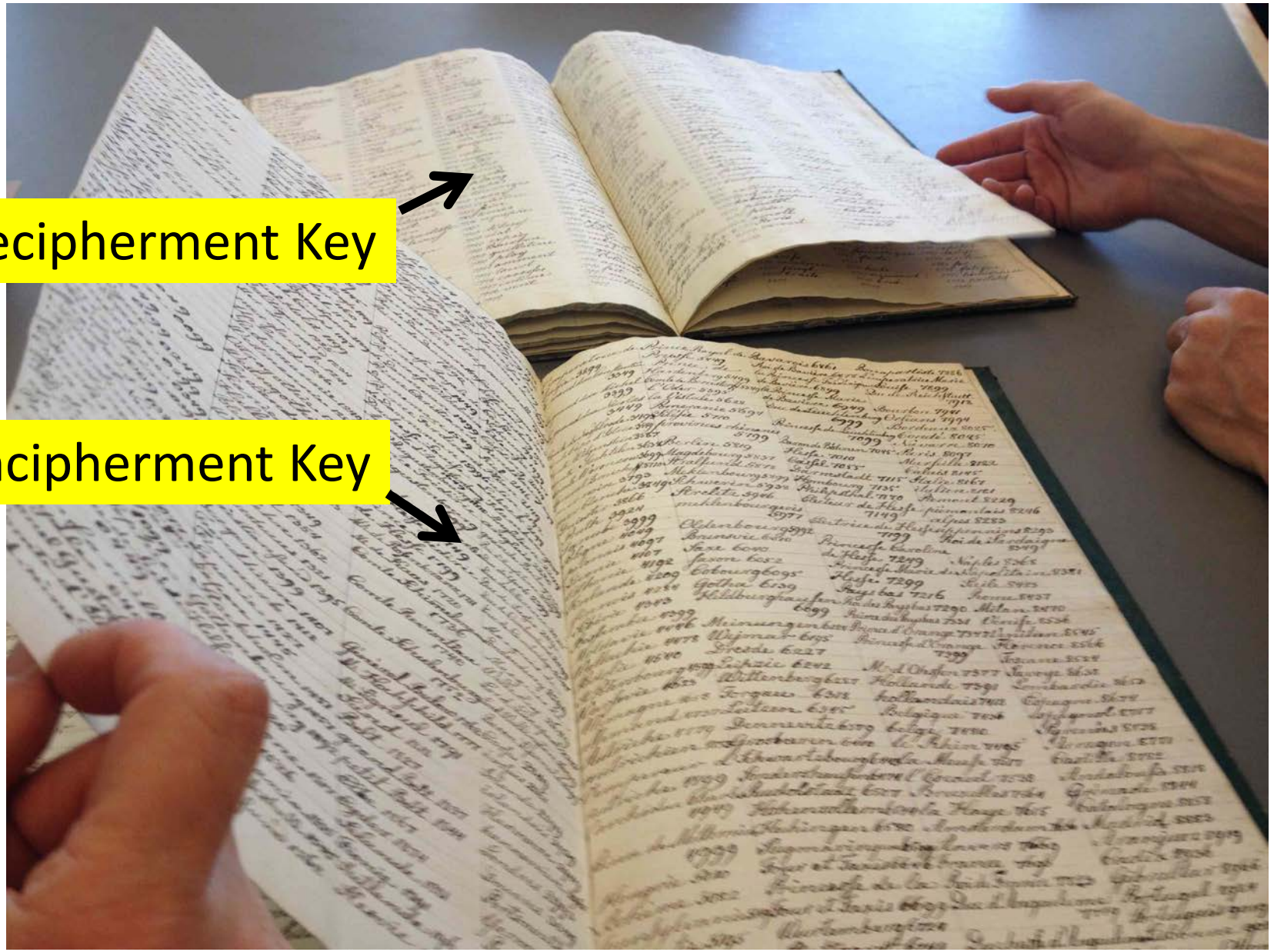
3A80. 3969. 375A. A65. A826. 3102. 27A3. A385. AA7. 1A97. A199. 416. 120. 60A. 2003. 2006. 203.

333. 1067. 2637. 3794. 1A96. 225A. 3818. 979. 2722. 3A22. 366A. A020. A33. A630. 1A80. 1969.

2587. 518. 286. A52. 2362. 2A10. 2987. 615. A0?2. 2524. 380. 2006. 3793. 2819. 3249. 330.

1021. 1A8. 3A66. 2A46. A51. A224. 1915. 2279. 3503. 55. 152A. 3A36. 1692. 2012. 393. 615. 1129.

Word Substitution Keys



Foreign Language as a Cipher

رفض رئيس السلطة الفلسطينية محمود عباس مجددا تصريحات وزير الخارجية الإسرائيلي سيلفان شالوم التي قال فيها إنه يتعين على إسرائيل إعادة النظر في انسحابها من غزة، المقرر أن يتم الصيف المقبل إذا فازت حركة المقاومة الإسلامية حماس في الانتخابات التشريعية.

وقال عباس في مؤتمر صحفي على هامش مشاركته في القمة العربية-اللاتينية الأولى إنه يتعين على إسرائيل احترام خيار الشعب الفلسطيني حتى لو فازت حماس بالانتخابات، وأضاف "إذا نجحت حماس أو فتح سيكون هذا خيار الشعب الفلسطيني، وعلى الجميع قبول هذا الخيار بكل ترحاب".

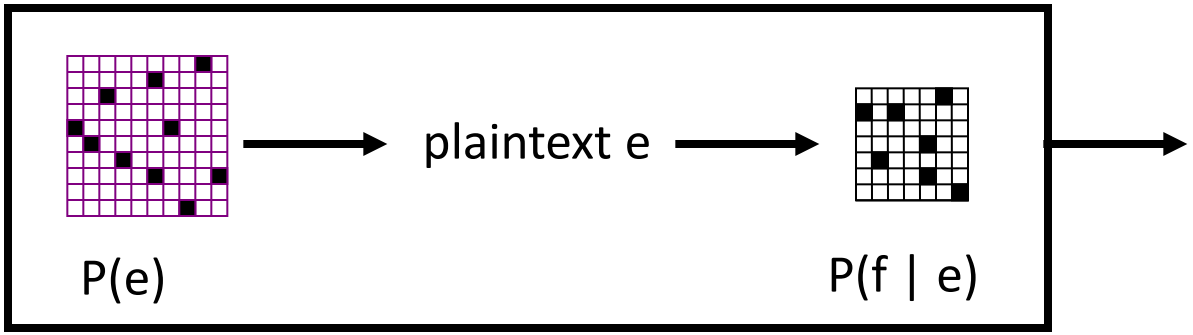
من جانبه شجب رئيس الحكومة الفلسطينية أحمد قريع الطابع الأحادي الجانب للانسحاب الإسرائيلي من غزة، وأكد أن إسرائيل تريد مغادرة هذه الأراضي لتعزيز سيطرتها على الضفة الغربية.

وقال قريع في كلمة له خلال مؤتمر نظمته وزارة الأوقاف في رام الله "سينسحبون من غزة ولكننا لا نعرف ما هو شكل هذا الانسحاب وماذا سيتركون، وما هو مصير المعابر والحدود, وكل ذلك غامض لأنه قرار أحادي الجان

Foreign Language as a Cipher

"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 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

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

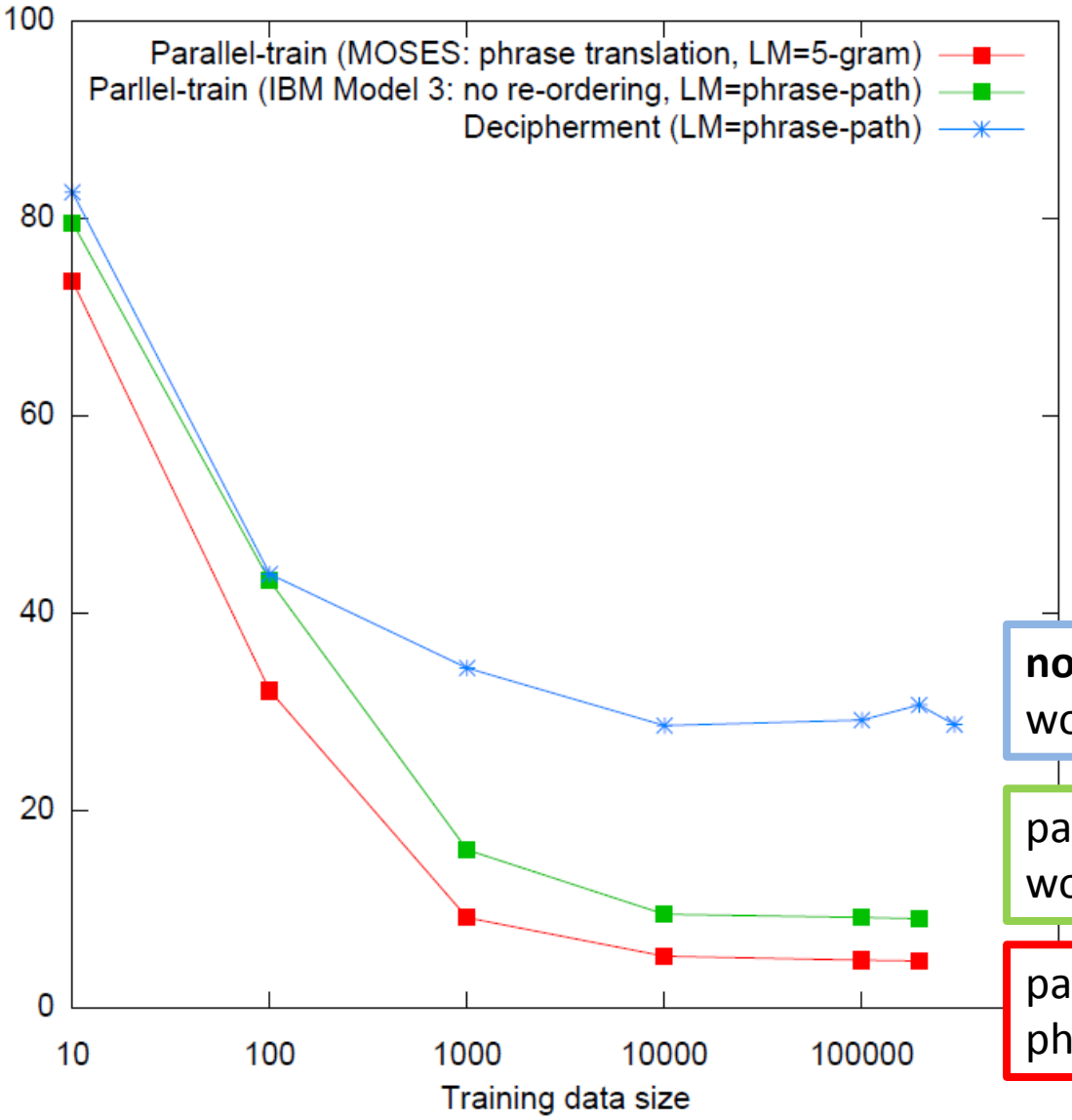
Hzyr!n

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)



no parallel data, word-based MT

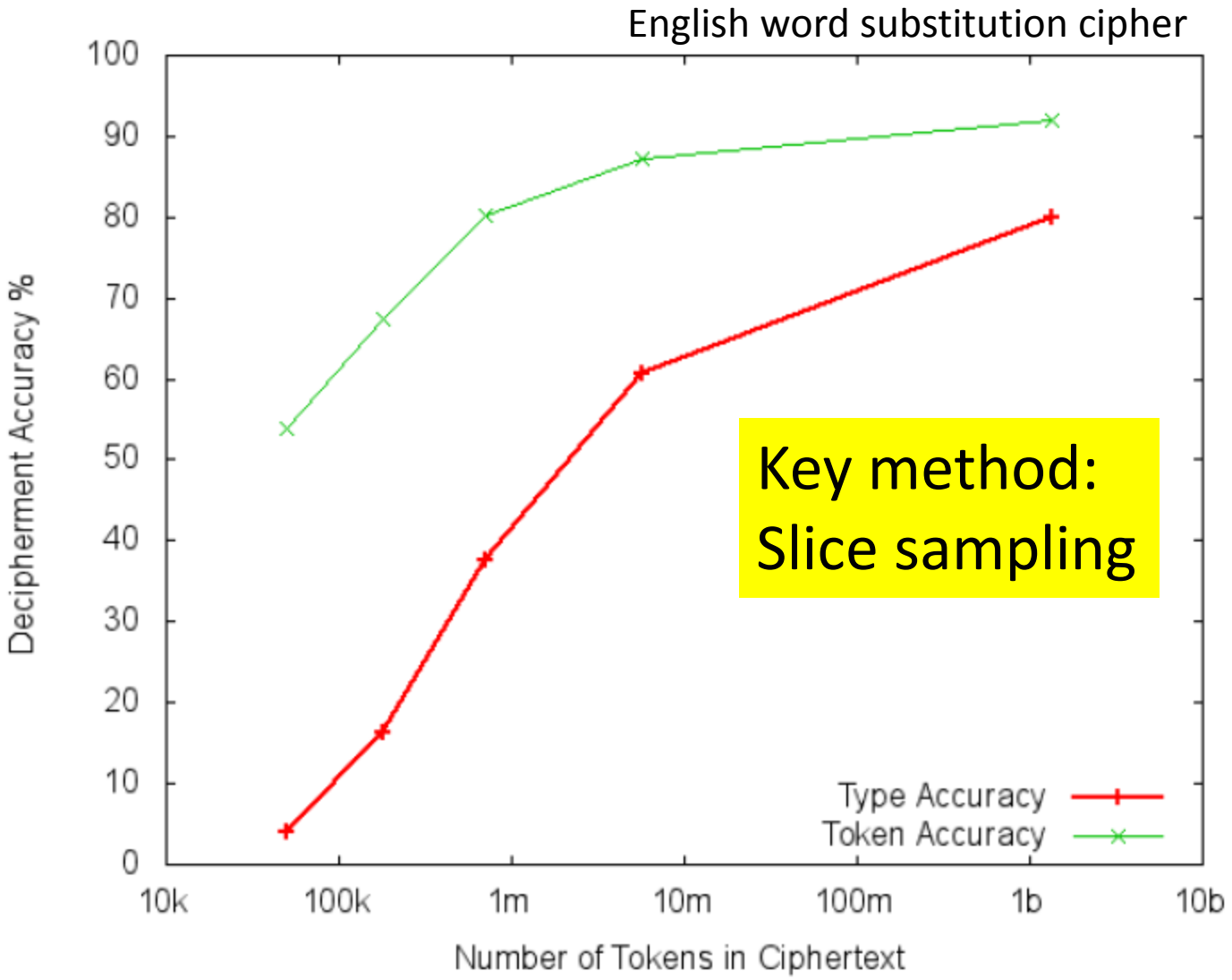
parallel data, word-based MT

parallel data, phrase-based MT

local re-ordering

[Ravi & Knight 11]

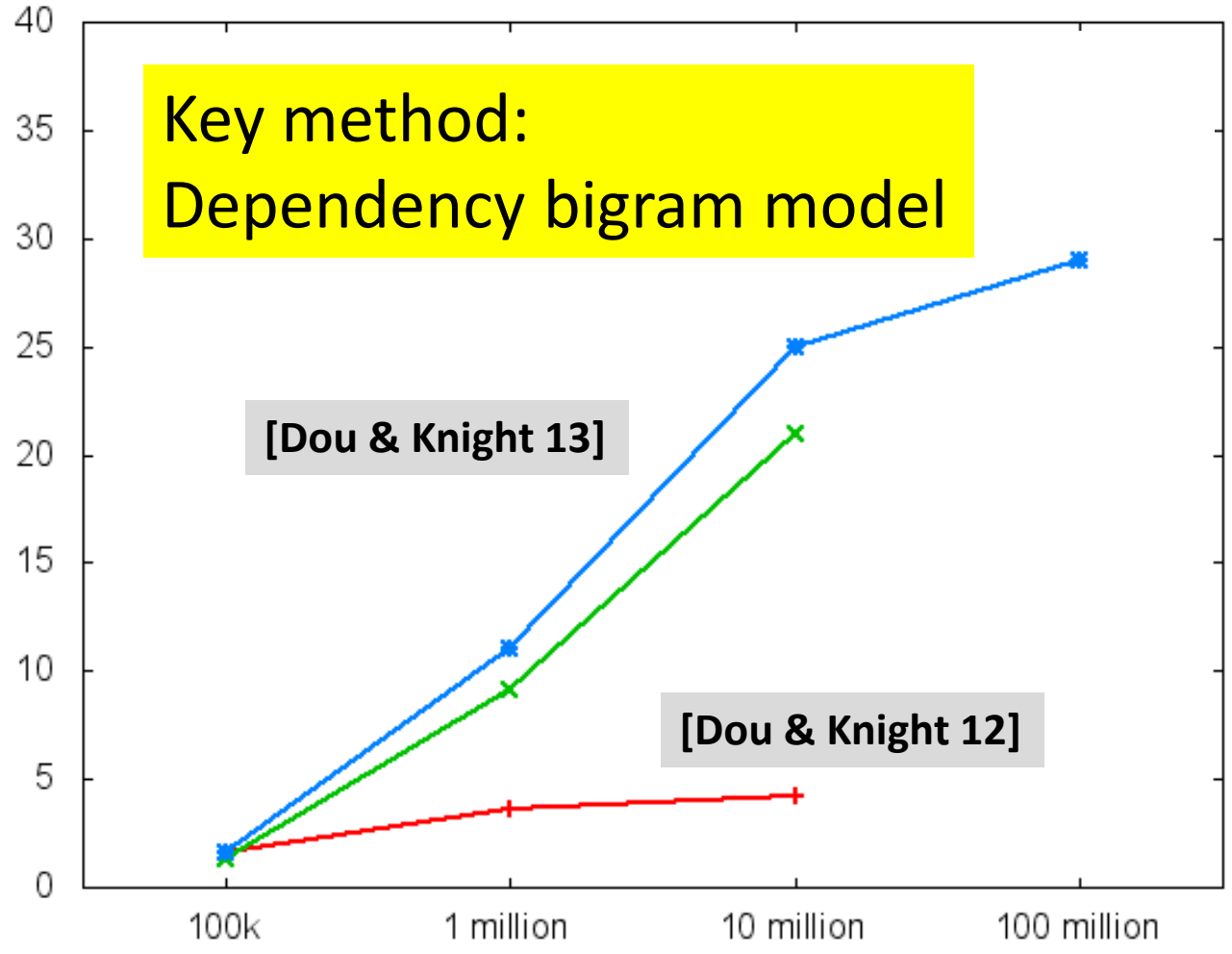
Scaling Up: Word Substitution



[Dou & Knight 12]

Scaling Up: Translation Dictionaries

Spanish as a cipher for English

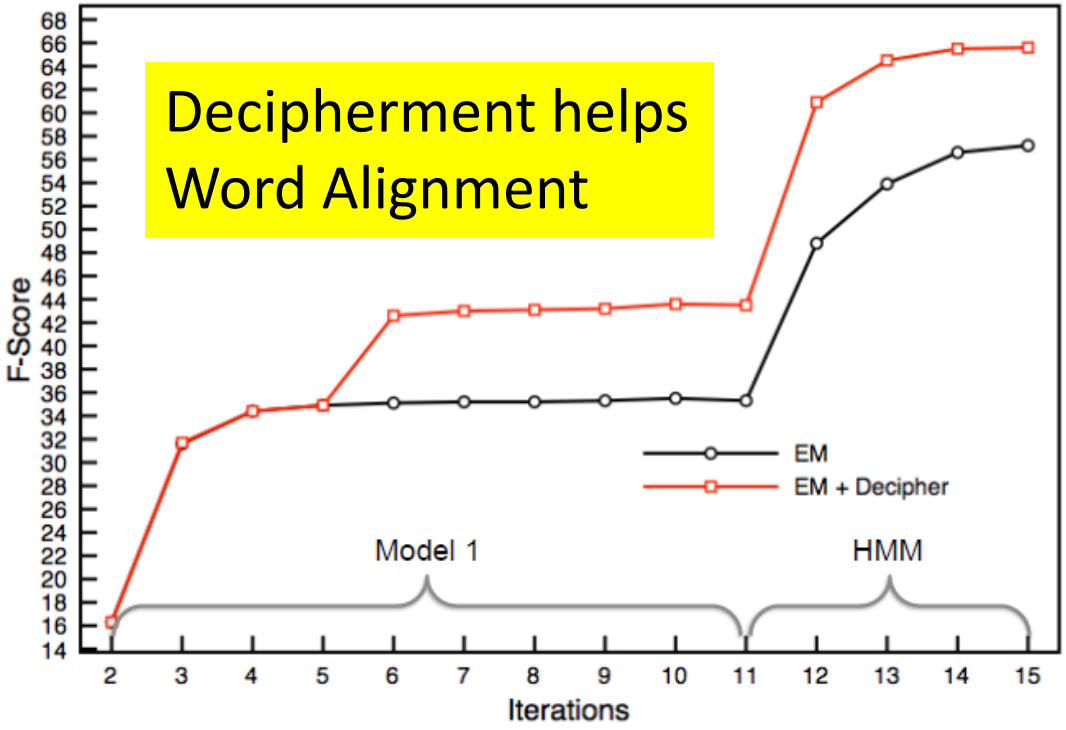


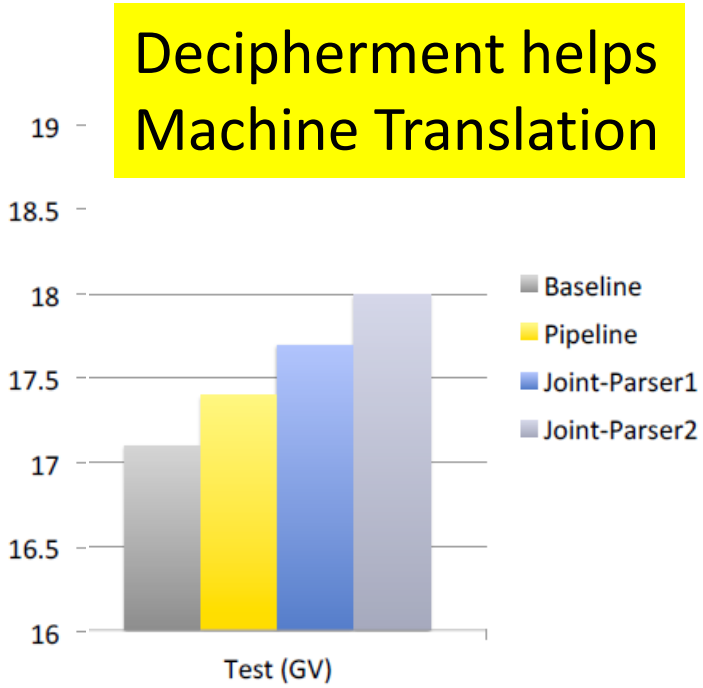
Accuracy of learned bilingual dictionary

Monolingual Spanish text (running words)

Scaling Up: Malagasy Translation

Small bilingual Malagasy/English text (need to **align** words [Brown et al 93])  Large Malagasy monolingual text (need to **decipher** [Dou & Knight 13])

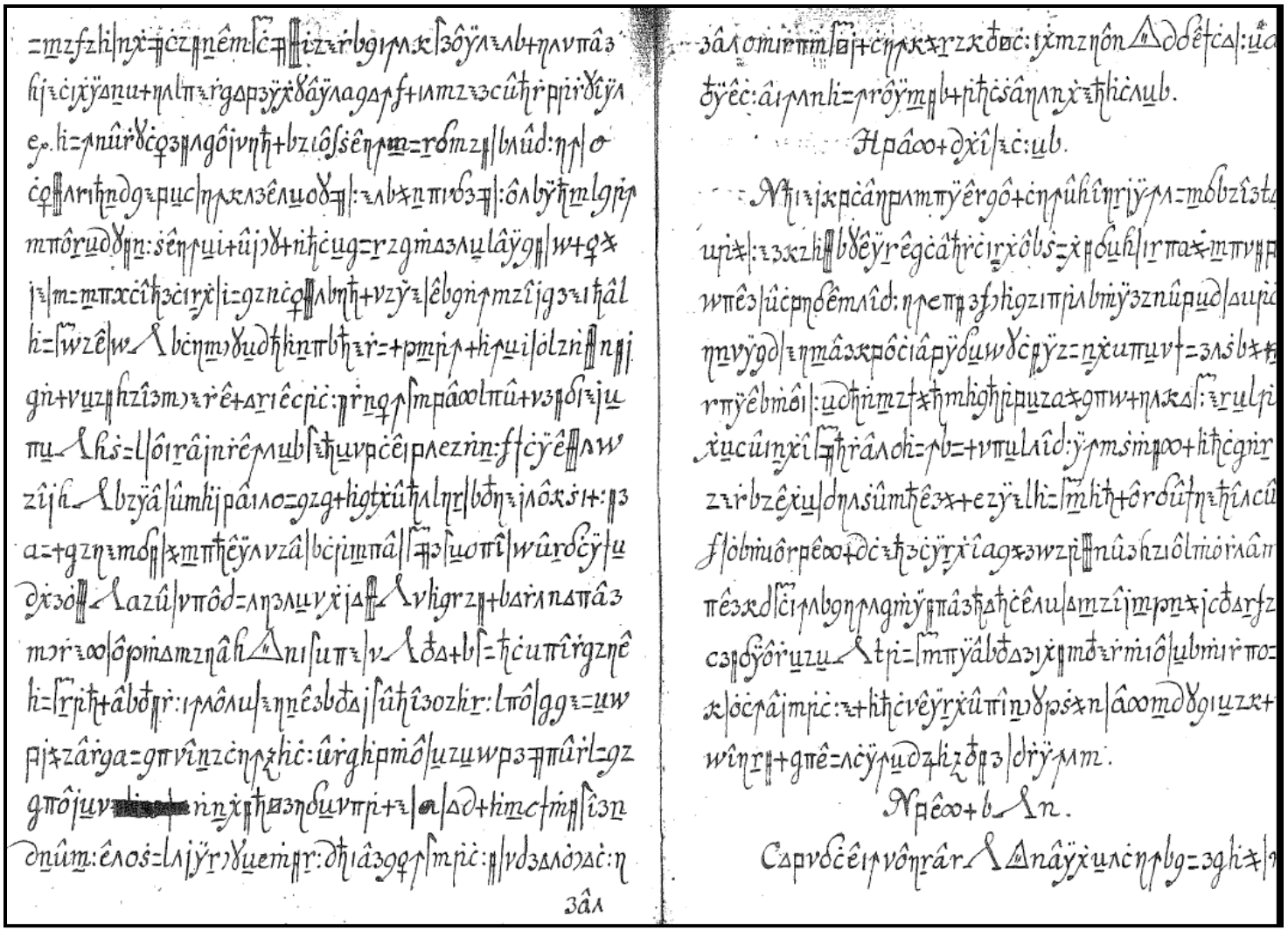




Plan for This Talk

- 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

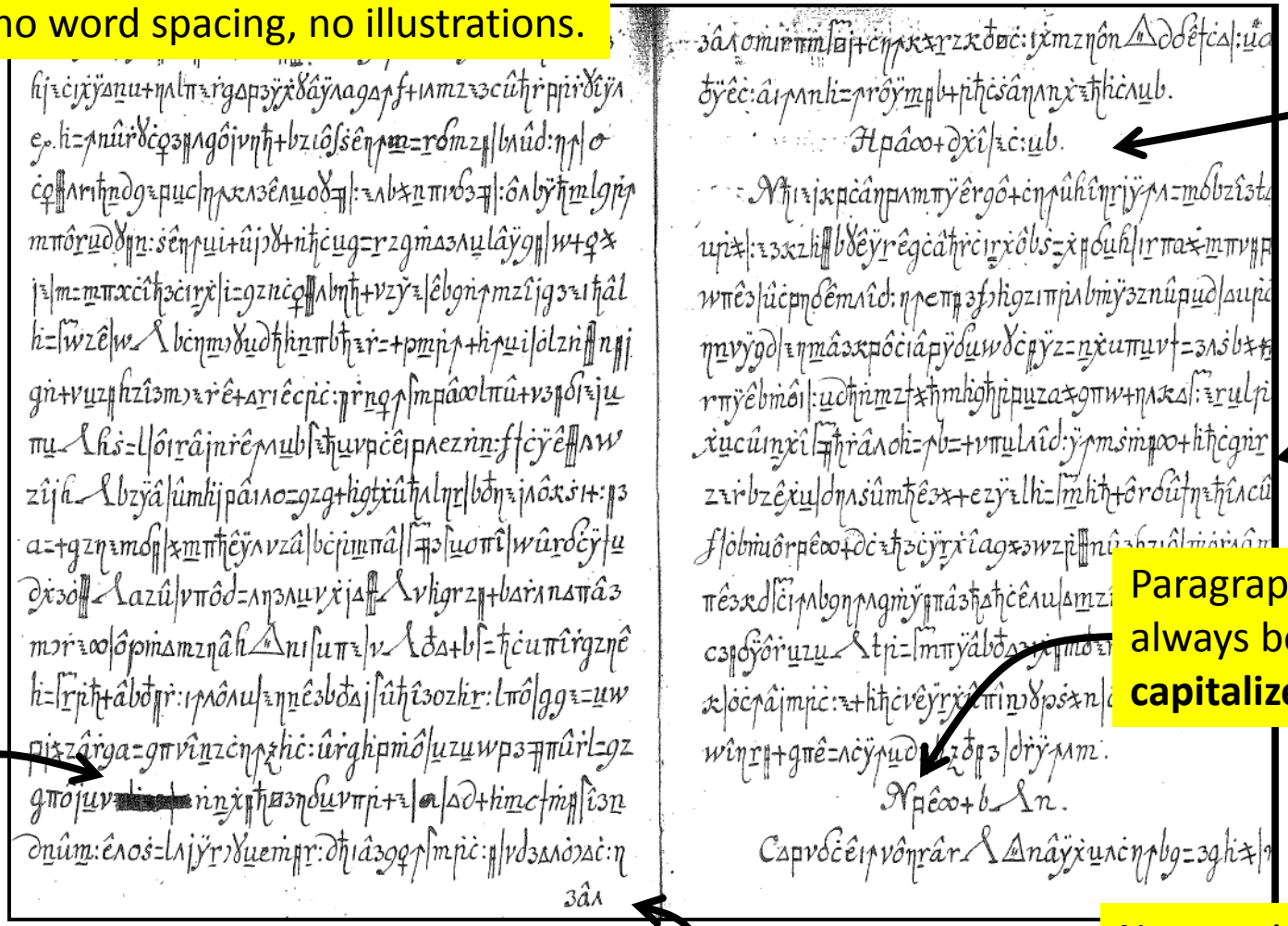
Copiale Cipher



[Knight, Megyesi, Schaefer 11]

Copiale Cipher

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



Section headers

Lines ≈ equal length

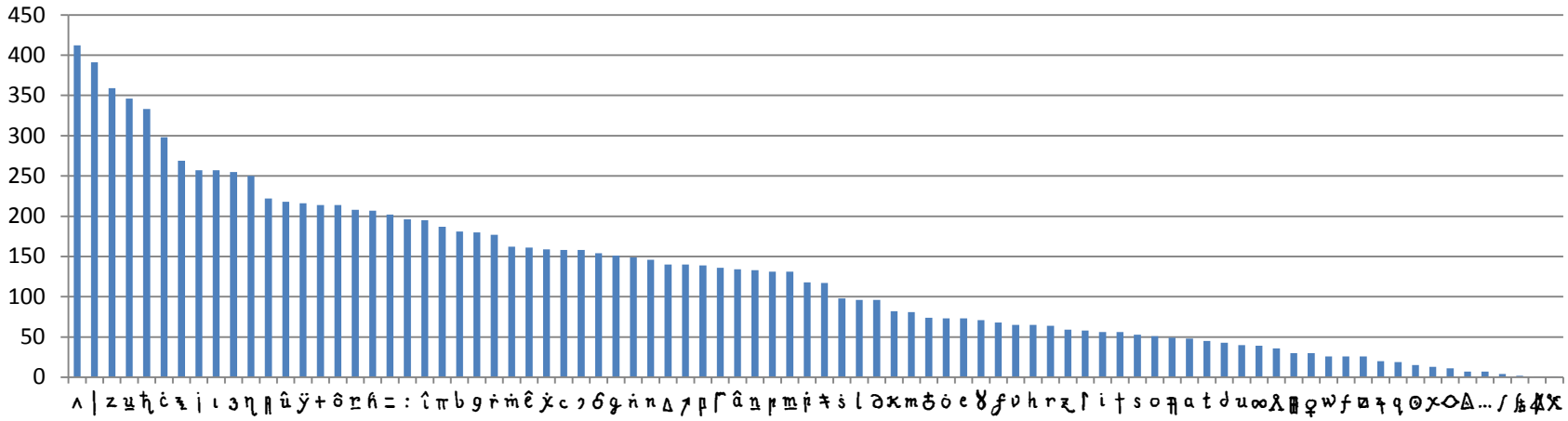
Paragraphs and section titles always begin with **capitalized Roman letters**.

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

Some scratch-outs, rare

Preview text fragments (“catchwords”)

Letter Frequencies



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

tendencies:

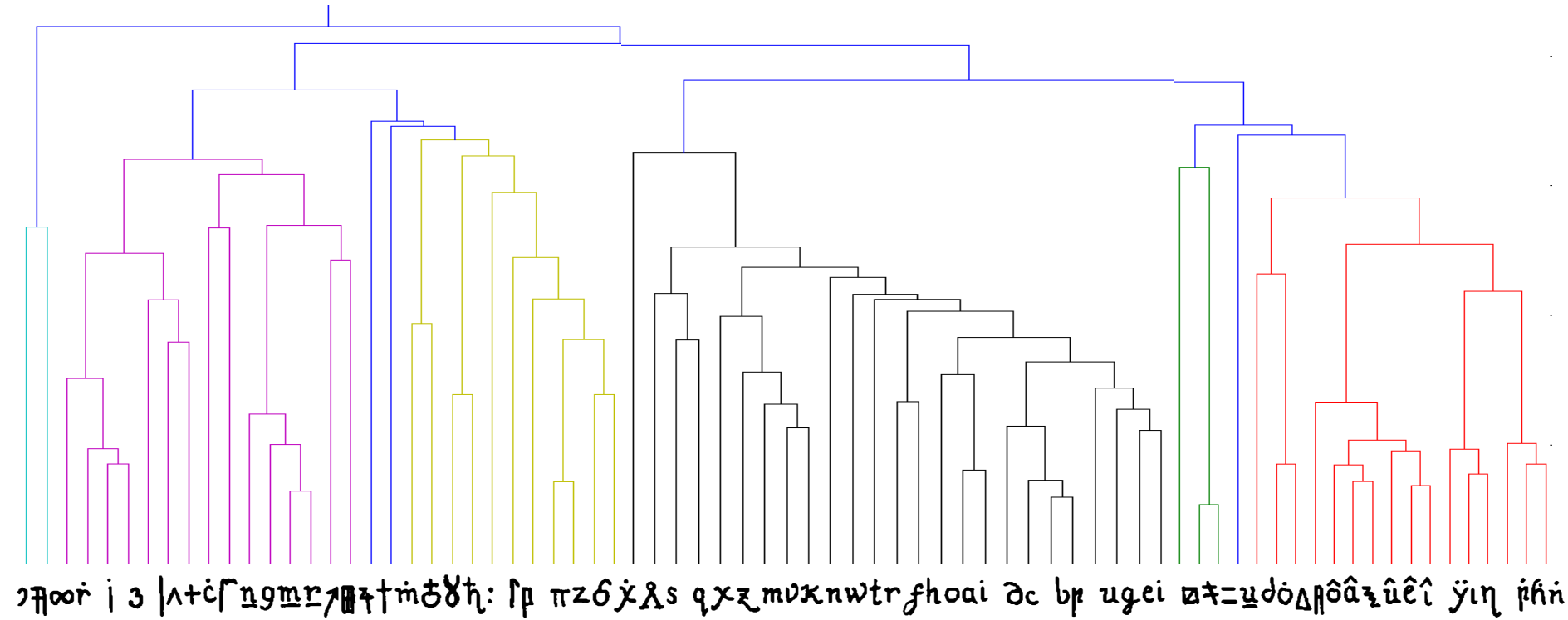
â, ê, î, ô, û followed by ʒ and ¡

â, ê, î, ô, û preceded by z and π

Clustering of Cipher Letters

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

Scipy software

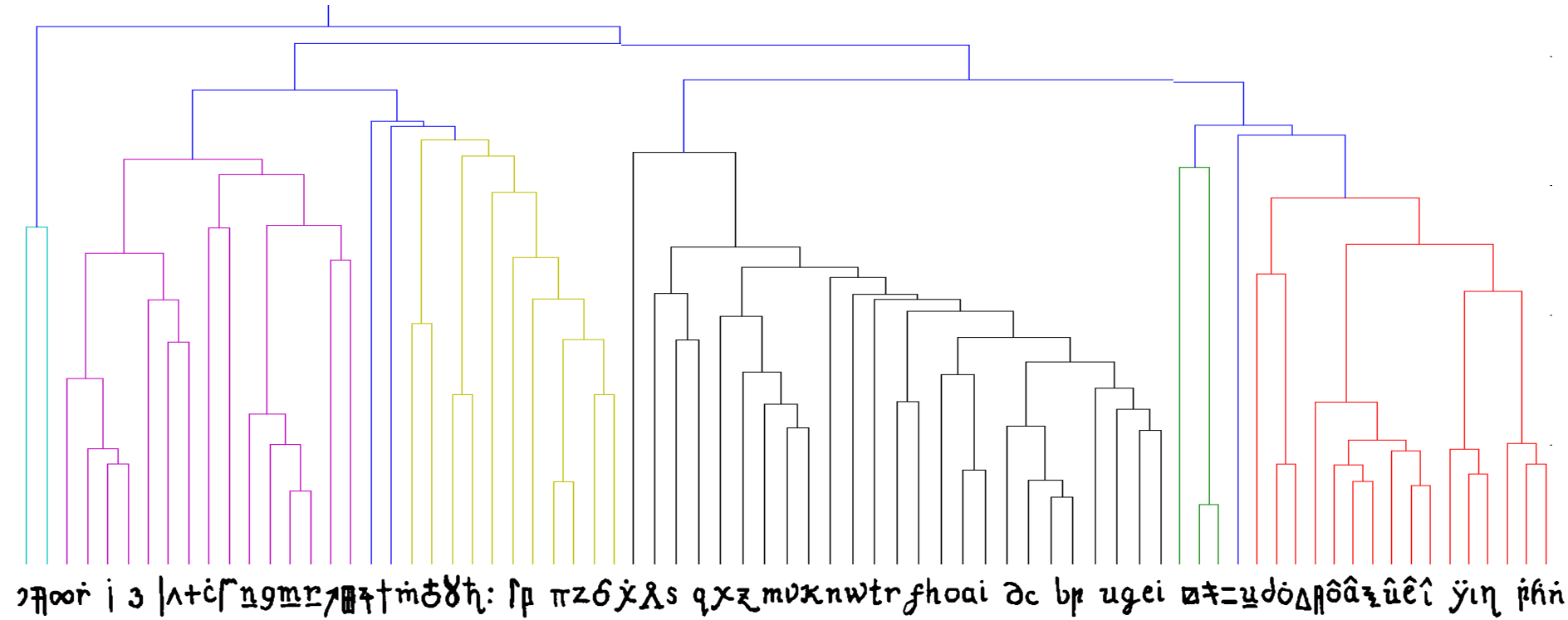


thanks Jon Graehl

Clustering of Cipher Letters

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

Scipy software



underlined letters

unaccented Roman letters

circumflexed vowels

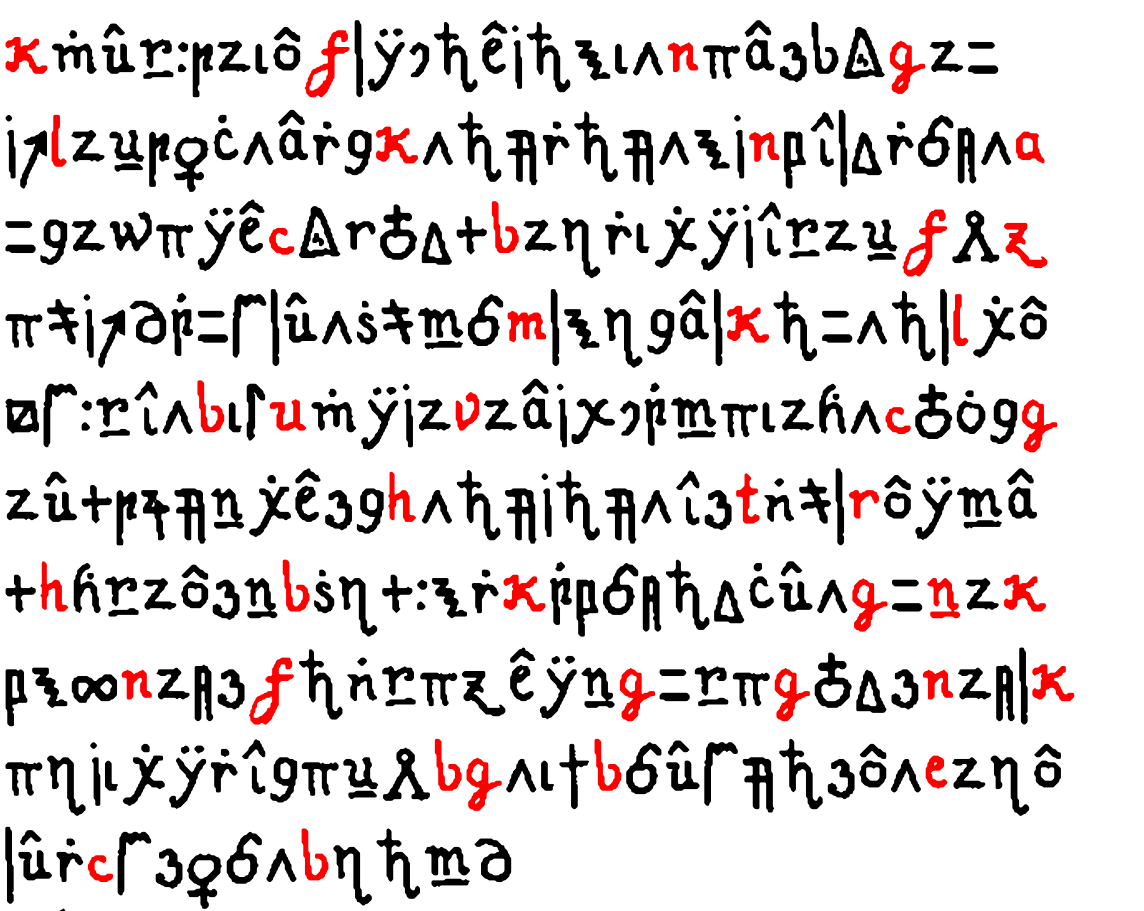
thanks Jon Graehl

First Decipherment Approach

unaccented Roman letters that cluster:

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

most common letter = 12%
least common = very small



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

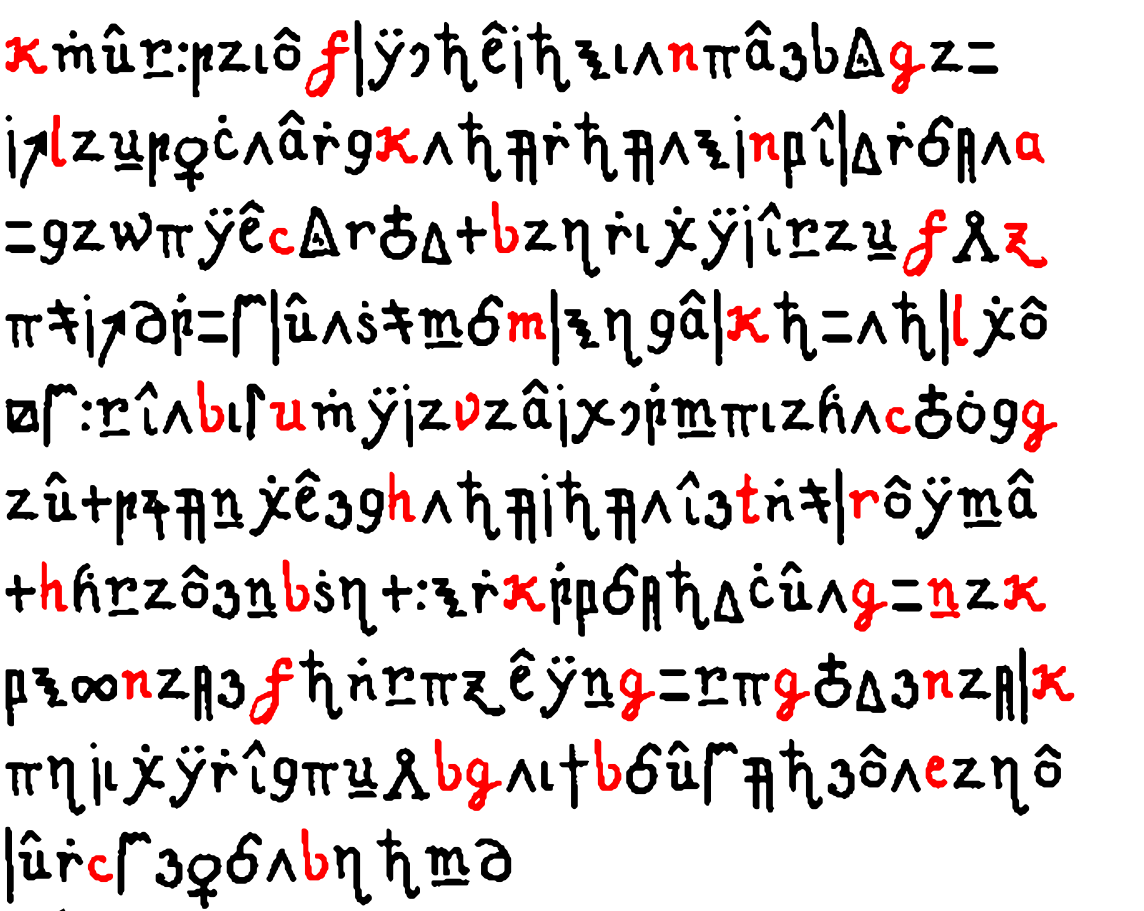
Decipher against
80 plaintext languages.

First Decipherment Approach

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

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 ...

D… …nst
80 … languages.

FAIL

Second Decipherment Approach

Homophonic cipher, e.g.:

A = 𝔄 𝔦 𝔩 𝔶 𝔯

B = 𝔲

C = 𝔬 𝔫

D = 𝔄

E = 𝔵 𝔣 𝔇 𝔨 𝔣 𝔦 𝔷 𝔷

F = 𝔭

G = 𝔶

etc.

𝔨𝔪𝔲𝔫:𝔭𝔷𝔦𝔬𝔣|𝔶𝔬𝔥𝔢|𝔥𝔷𝔦𝔞𝔫𝔭𝔞𝔷𝔟𝔇𝔤𝔷=
𝔦𝔩𝔩𝔷𝔲𝔭𝔮𝔠𝔞𝔞𝔯𝔤𝔨𝔞𝔥𝔄𝔯𝔥𝔄𝔞𝔷𝔦𝔫𝔭𝔦|𝔇𝔯𝔬𝔄𝔞𝔞
=𝔤𝔷𝔴𝔫𝔶𝔢𝔠𝔇𝔯𝔡𝔇+𝔟𝔷𝔥𝔯𝔦𝔵𝔵𝔶|𝔦𝔫𝔷𝔲𝔣𝔏𝔷
𝔫𝔨|𝔩𝔡𝔭=𝔣|𝔲𝔞𝔰𝔨𝔪𝔬𝔪|𝔷𝔥𝔤𝔞|𝔨𝔥=𝔞𝔥||𝔩𝔵𝔬
𝔴𝔣:𝔫𝔦𝔞𝔟𝔦𝔣𝔲𝔪𝔶|𝔷𝔳𝔷𝔞|𝔵𝔬𝔯𝔪𝔫𝔦𝔷𝔥𝔞𝔠𝔡𝔬𝔤𝔤
𝔷𝔲+𝔭𝔨𝔄𝔫𝔵𝔢𝔷𝔤𝔥𝔞𝔥𝔄𝔦𝔥𝔄𝔞𝔦𝔷𝔱𝔫𝔨|𝔯𝔬𝔶𝔪𝔞
+𝔥𝔥𝔫𝔷𝔬𝔷𝔫𝔟𝔰𝔥+:𝔷𝔯𝔨𝔯𝔭𝔬𝔄𝔥𝔇𝔠𝔲𝔞𝔤=𝔫𝔷𝔨
𝔭𝔷∞𝔫𝔷𝔄𝔷𝔣𝔥𝔫𝔫𝔫𝔷𝔢𝔶𝔫𝔤=𝔫𝔫𝔤𝔡𝔇𝔷𝔫𝔷𝔄|𝔨
𝔫𝔥𝔦𝔵𝔶𝔯𝔦𝔤𝔫𝔲𝔏𝔟𝔤𝔞𝔦+𝔟𝔬𝔲𝔣𝔄𝔥𝔷𝔬𝔞𝔷𝔥𝔬
|𝔲𝔯𝔠𝔣𝔷𝔵𝔬𝔞𝔟𝔥𝔥𝔪𝔡

Homophonic Cipher

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

FAIL

But, slight numerical preference for German

Cipher Characteristics

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

tendencies:

â, ê, î, ô, û followed by ɜ and ¡

â, ê, î, ô, û preceded by z and π

should appear adjacent in German text

Make full digraph table for cipher and for German

Key Observation #1

In Copiale, ʔ almost always followed by ħ

In German, C almost always followed by H

(German CH is like English QU)

So guess: ʔ = C, ħ = H

One Thing Leads to Another

ɔħ = CH → ɔħʌ = CHT → ʌ = T ?

Each step is guesswork.

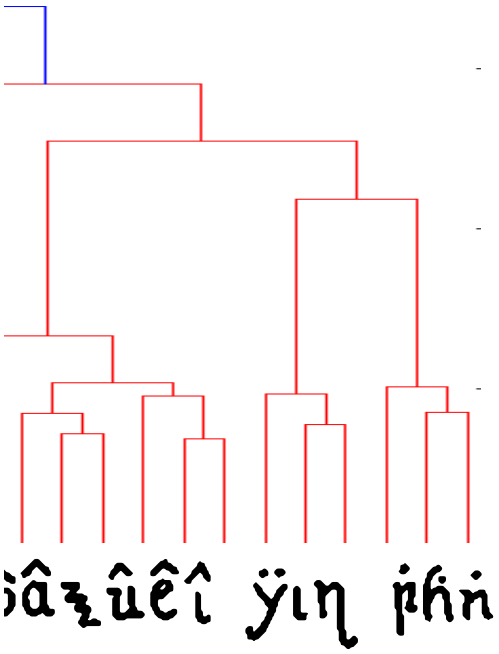
Must be willing to retract.

Weird task, not knowing German.

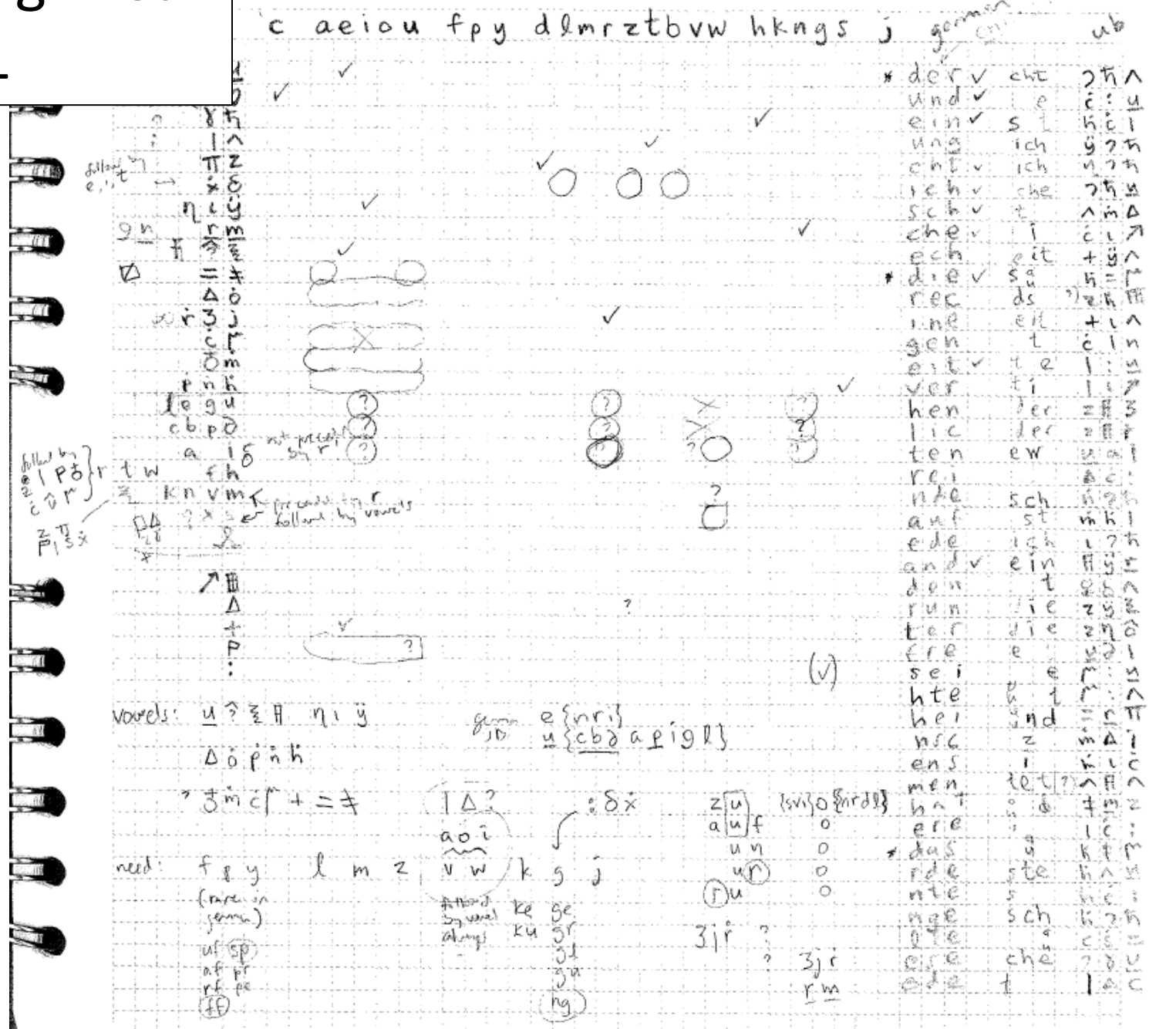
No longer care what the book says.

Cluster diagram crucial:

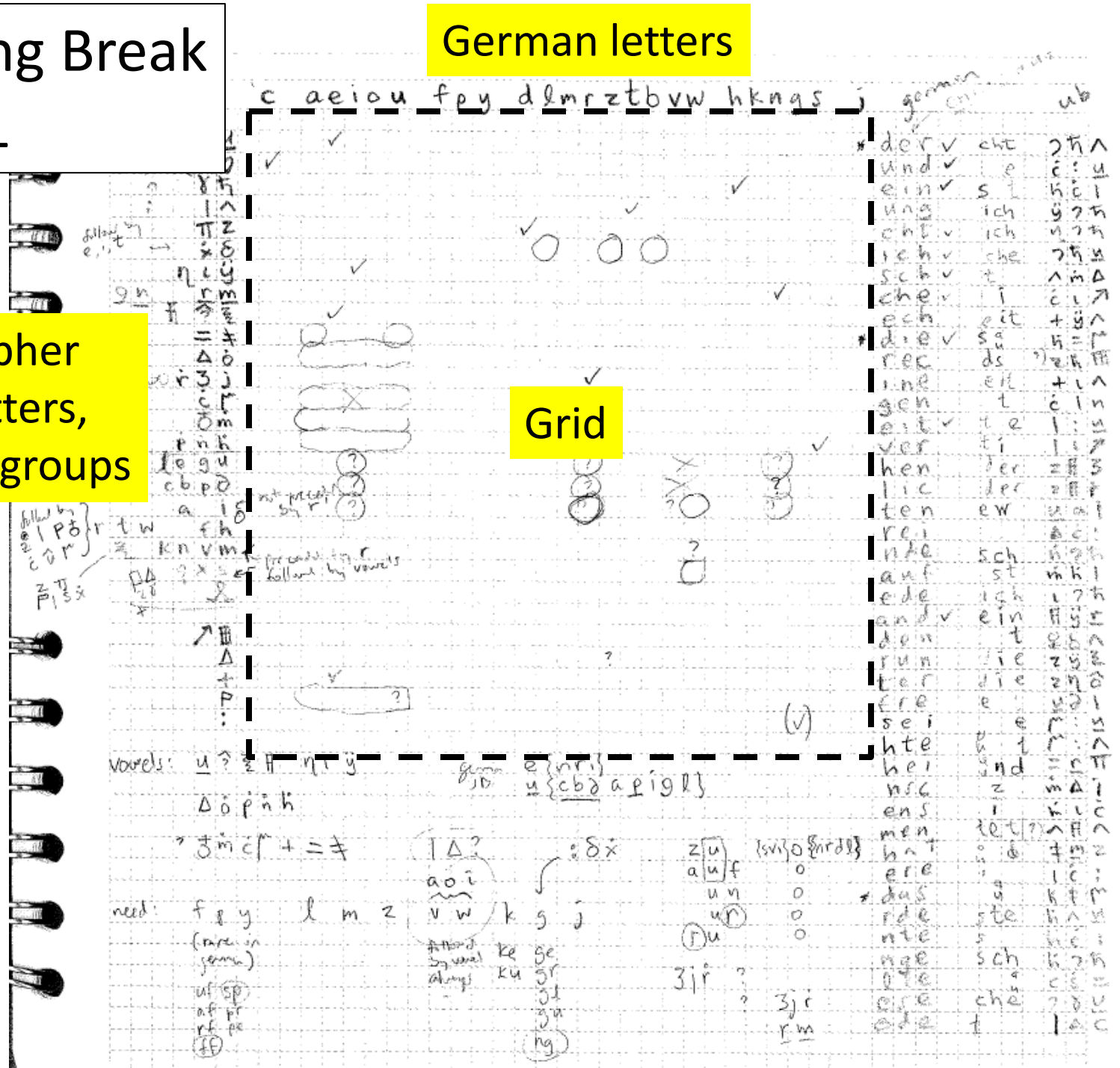
ÿ = I → ɩ = I , ɳ = I



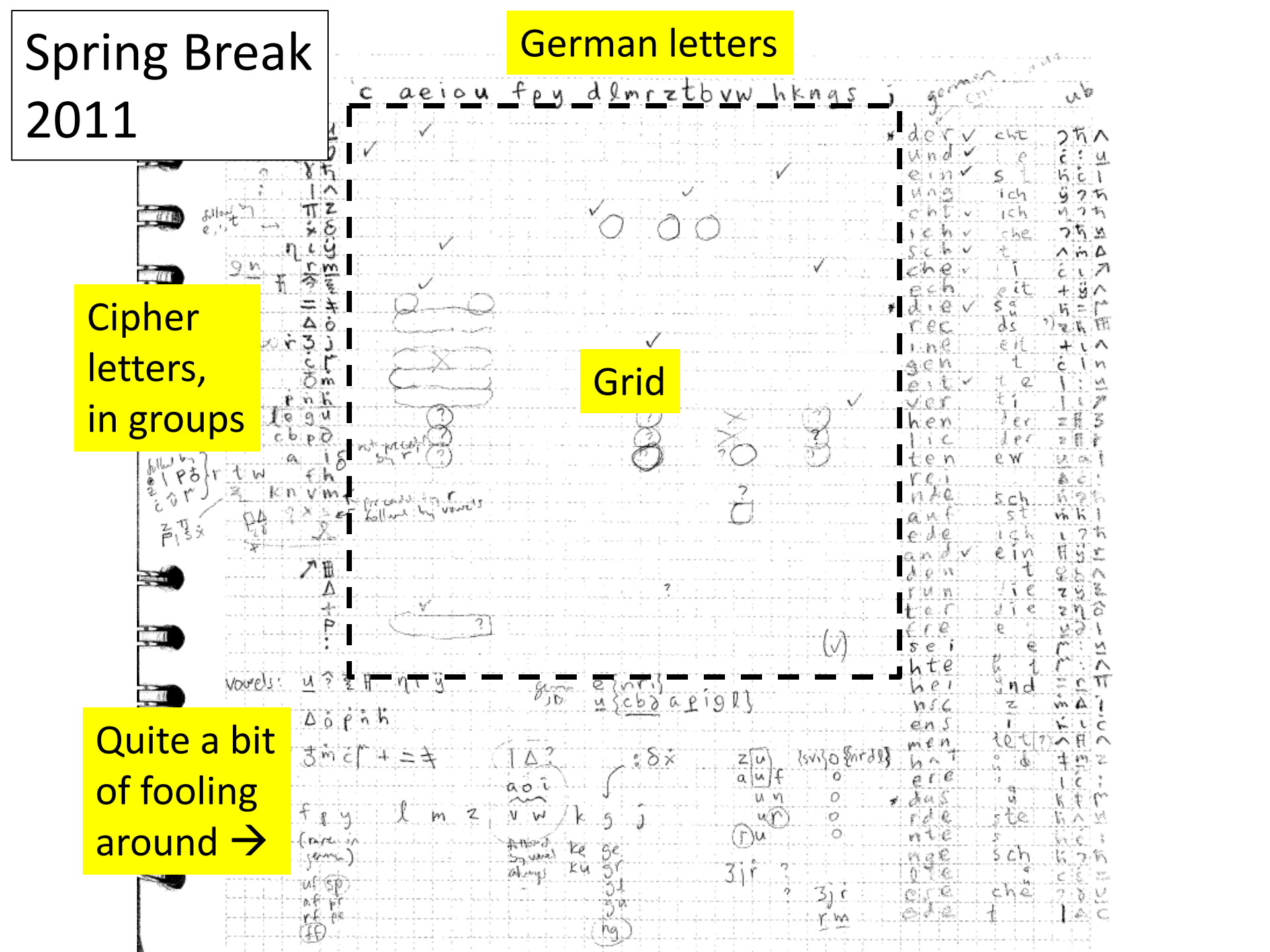
Spring Break 2011



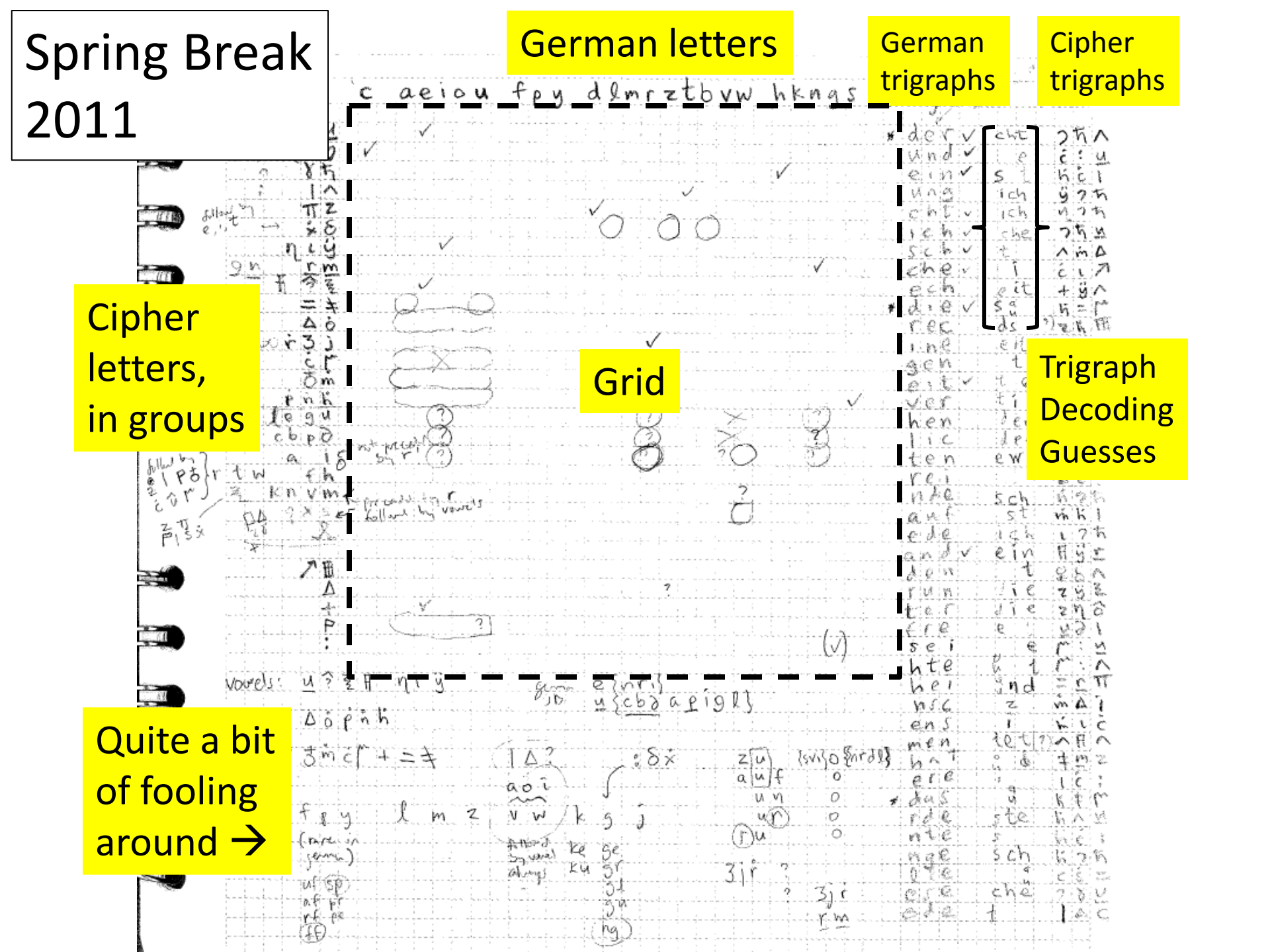
Spring Break 2011



Spring Break 2011



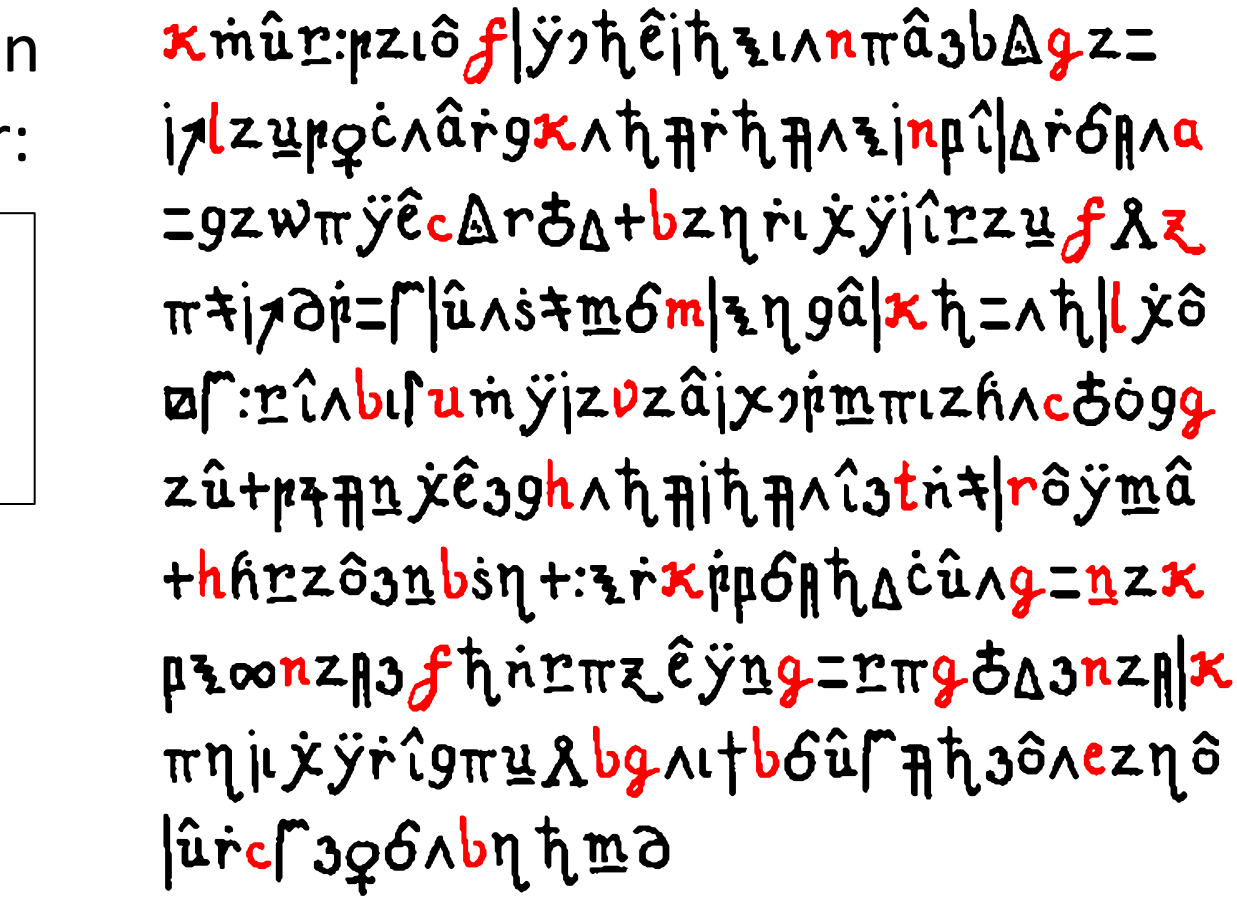
Spring Break 2011



Key Observation #2

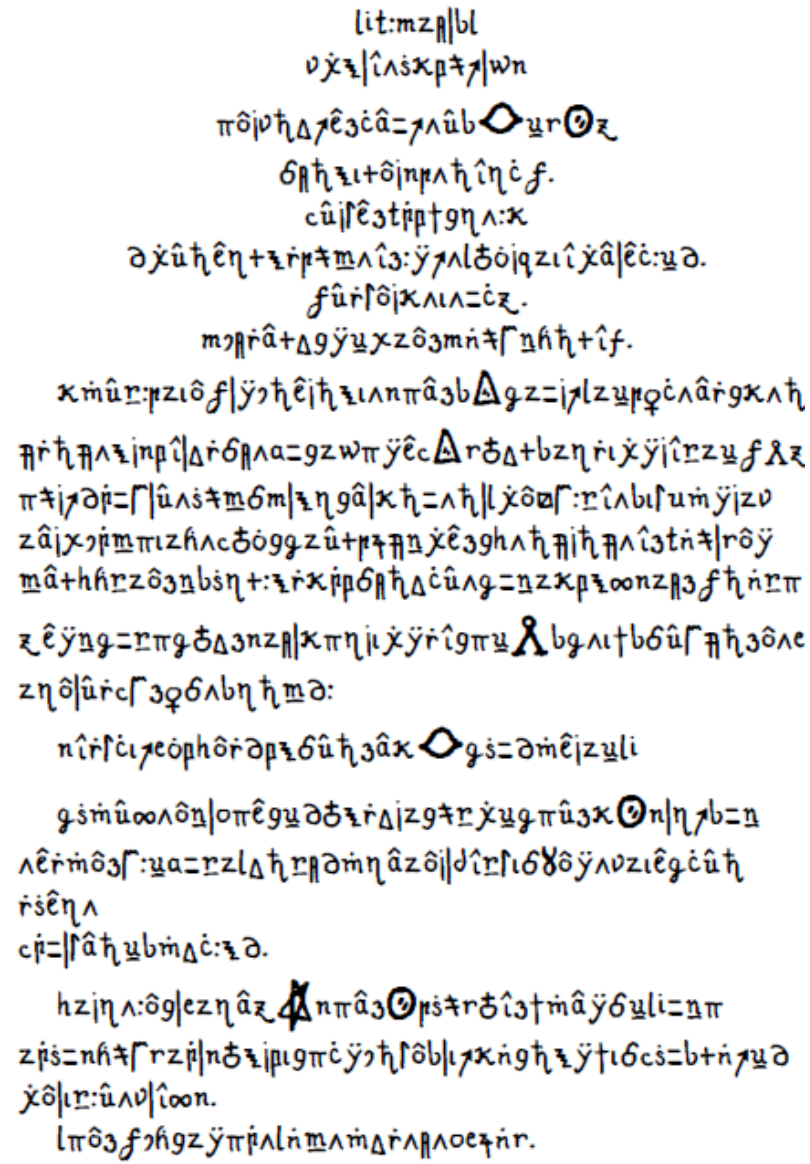
unaccented Roman letters that cluster:

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



Actually, those are space bars

Copiale Decipherment



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

wenn die sicherheit der △ durch den ältern thürheter besorget und die △ vom dirigirenden ⅄ 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:

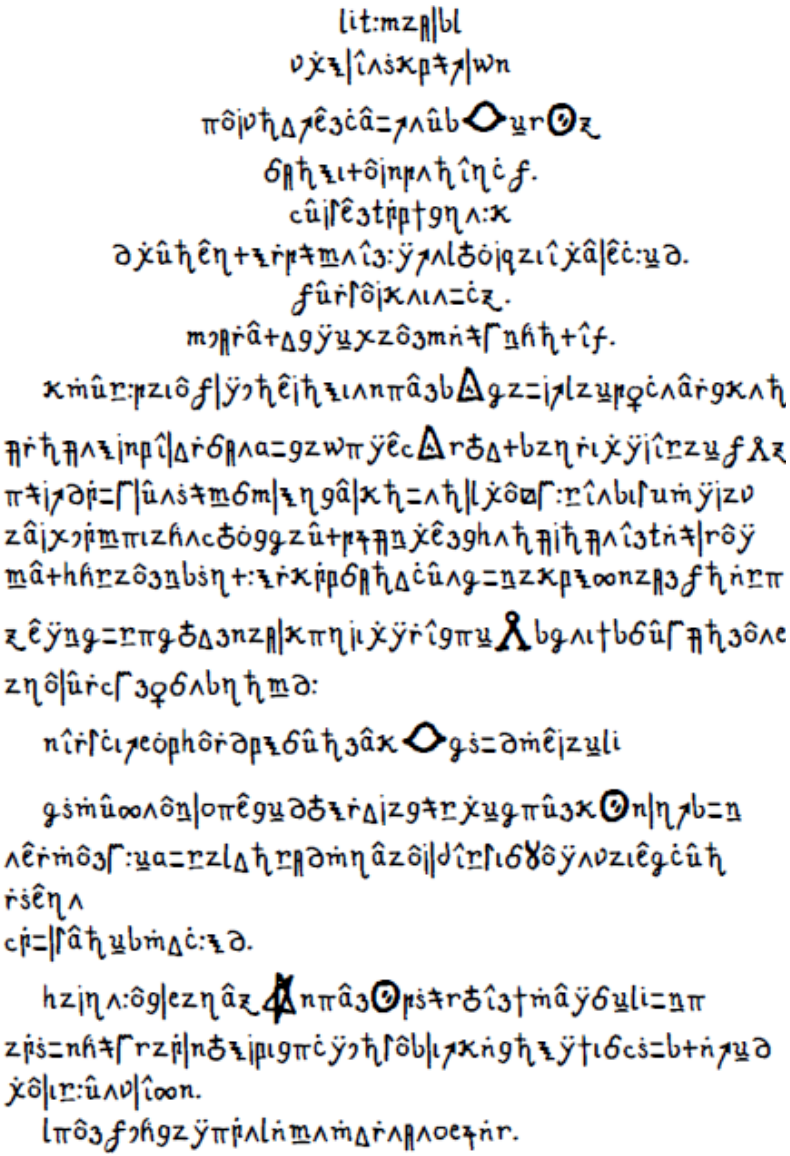
erstlich ob er begehre ◇ zu werden

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 ⌔.

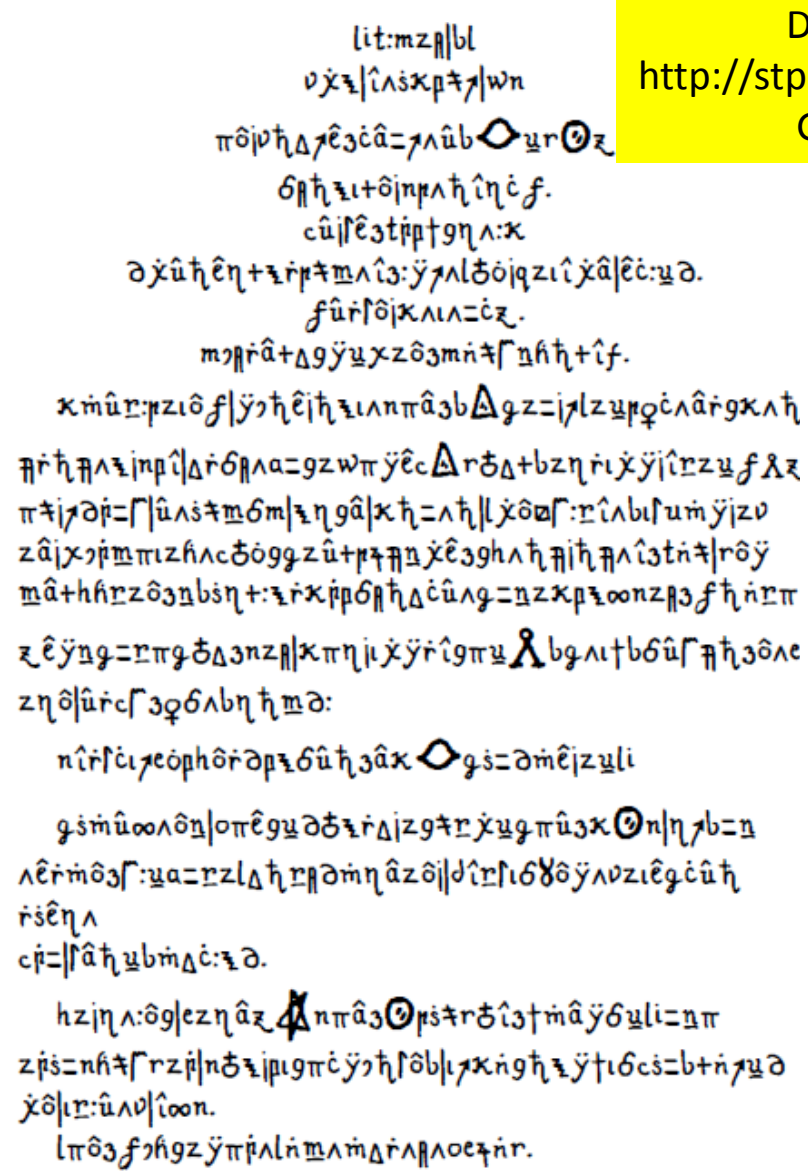
Secondly, if he submits to the rules of the ⊙ and without rebelliousness suffer through the time of apprenticeship.

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.

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:

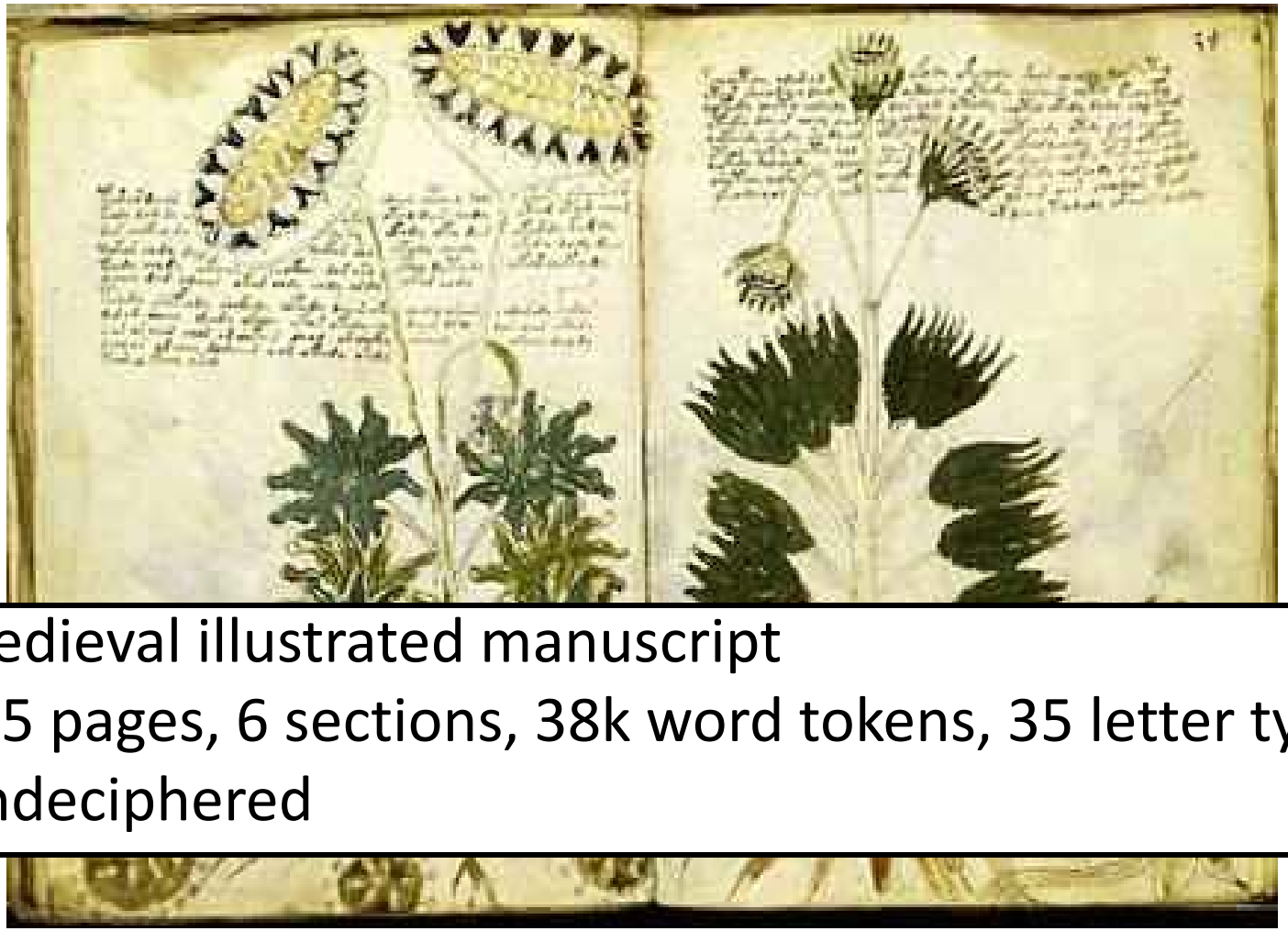
First, if he desires to become ◇.

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.

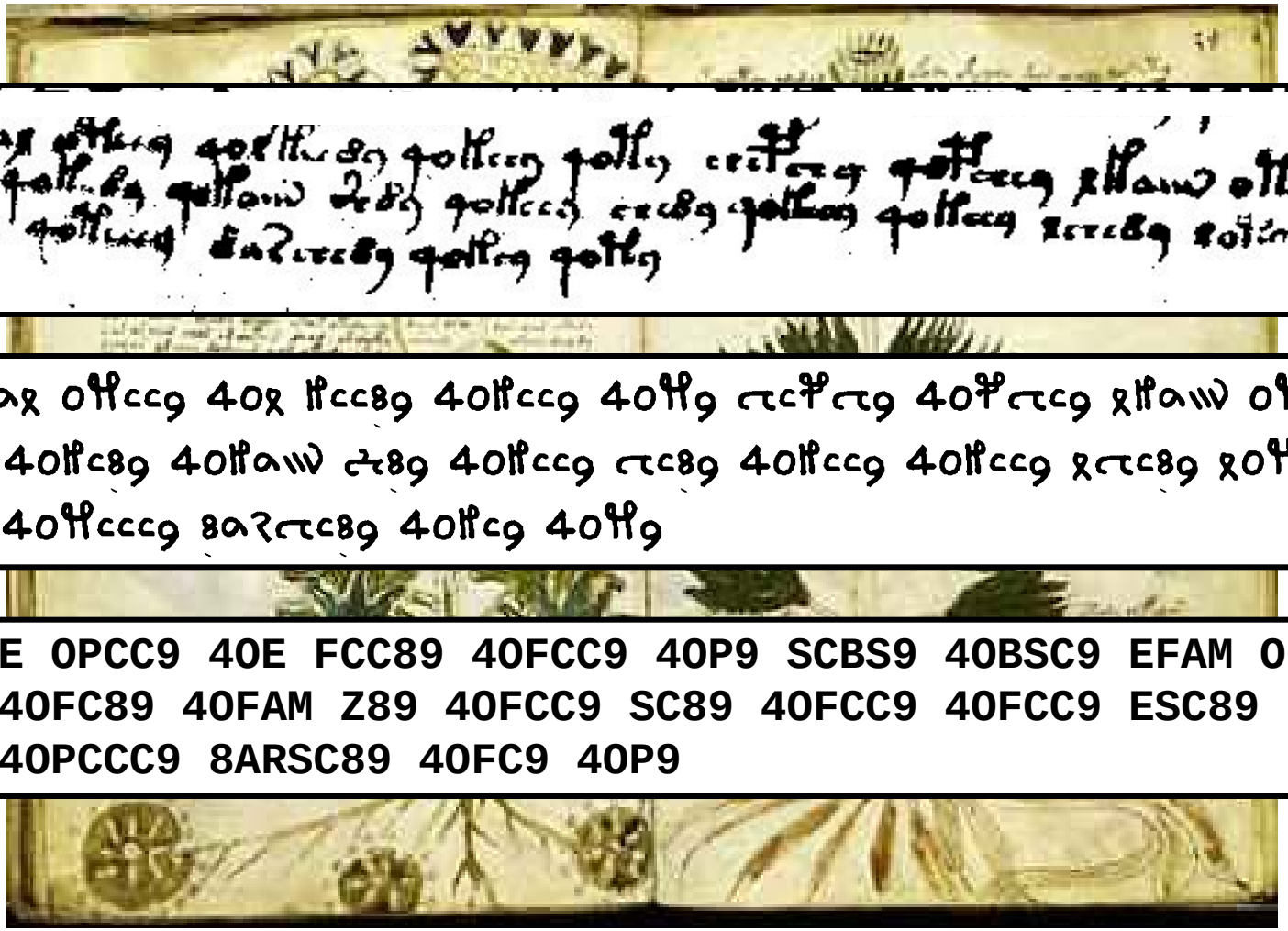
The candidate answers yes.

Voynich Manuscript



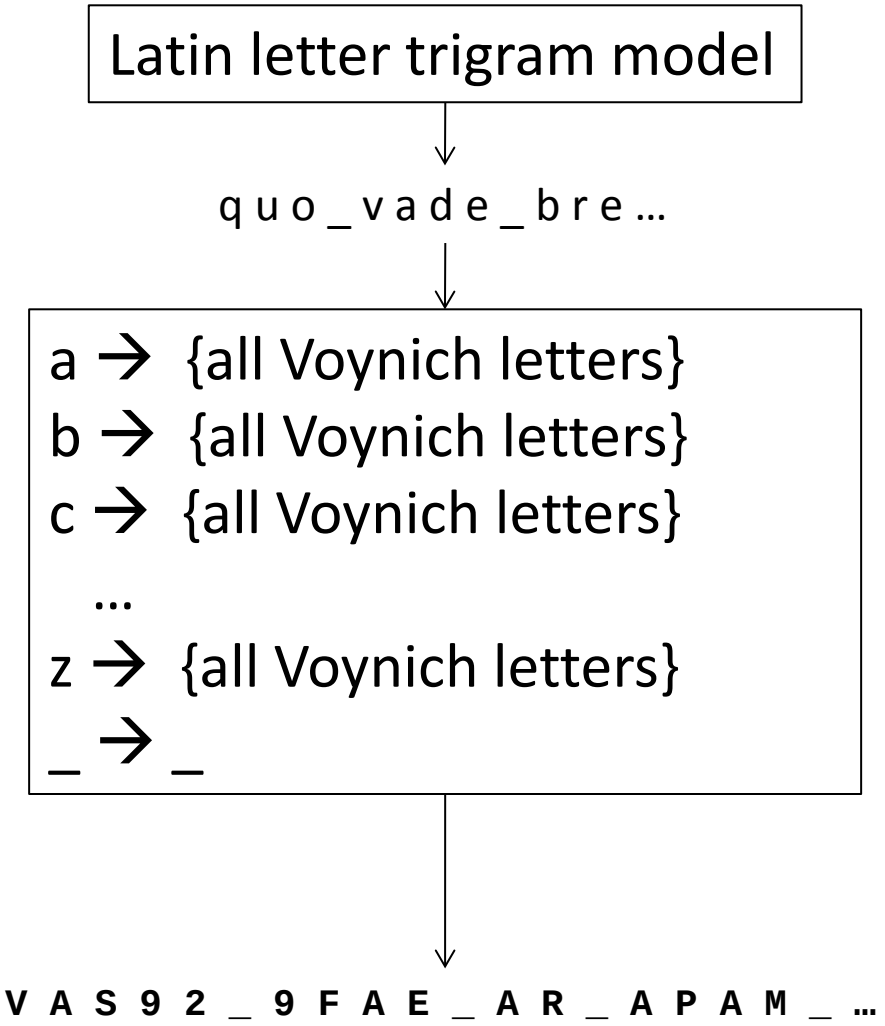
- Medieval illustrated manuscript
- 235 pages, 6 sections, 38k word tokens, 35 letter types
- Undeciphered

Voynich Manuscript



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

Voynich Letter Substitution

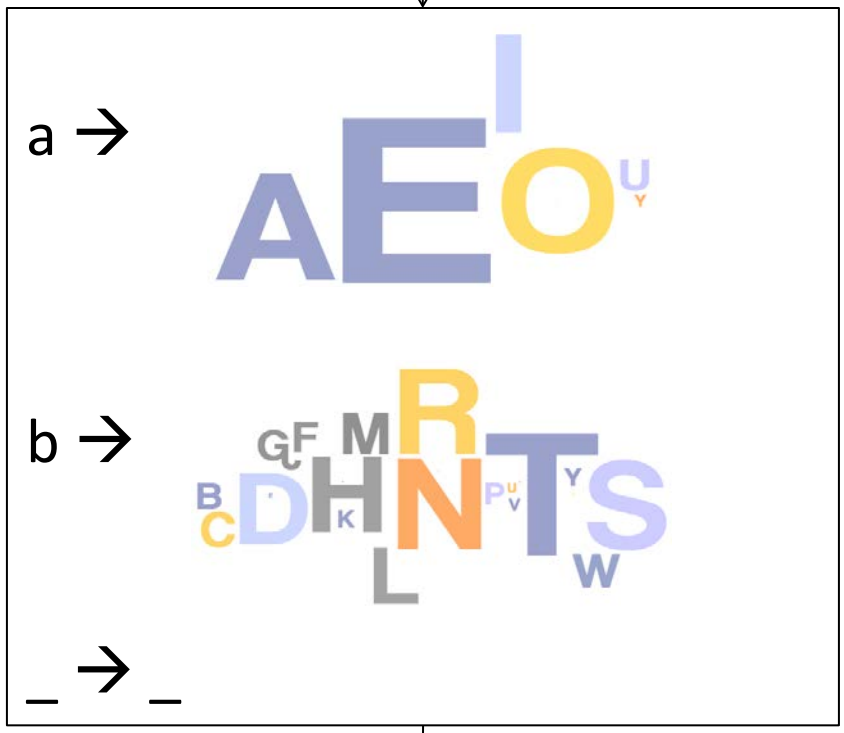


Input	Decipherment
VAS92 9FAE AR APAM ZOE ZOR9 QOR92 9 FOR ZOE89 …	quiss squm is onum pom quss hates s qum hatis …

Letter Clustering

Trigram model over {a, b, _ }

a a _ b a b _ a b a a _ …



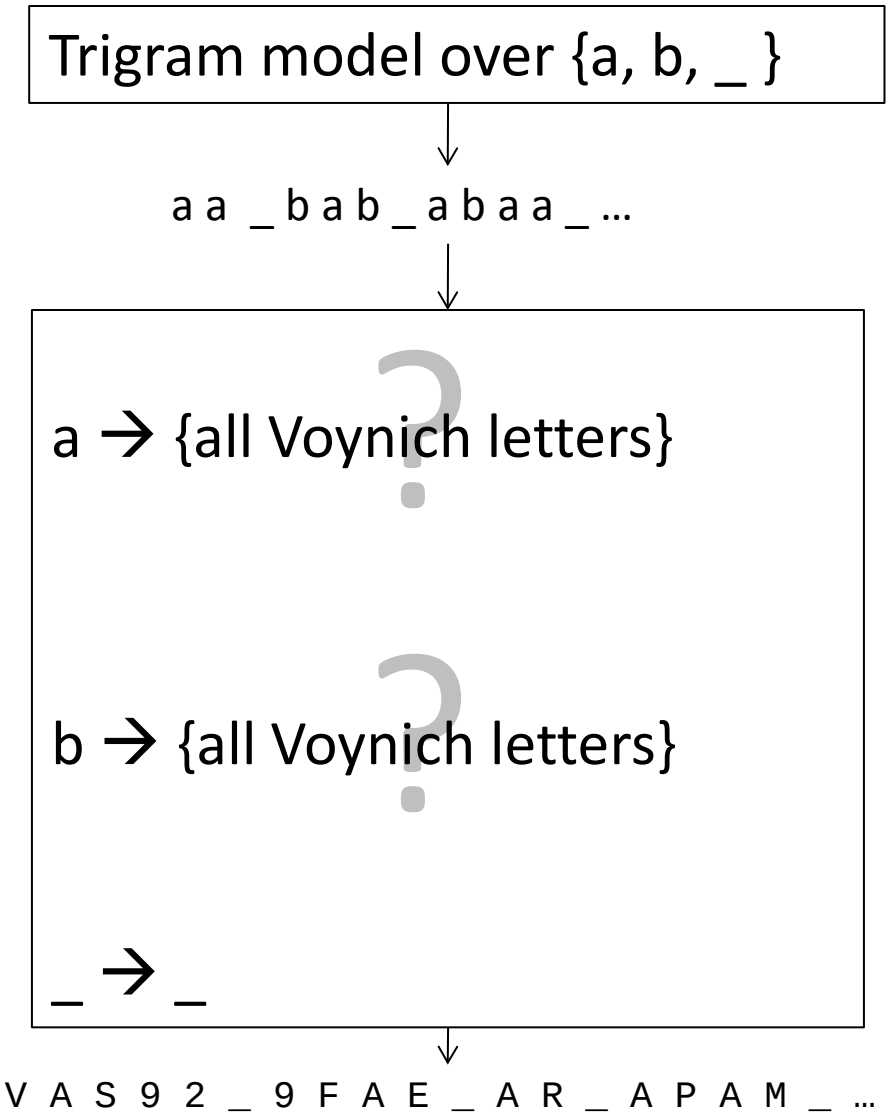
i n _ t h e _ t o w n _ w h e r e _ i _ was …

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



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 _ …

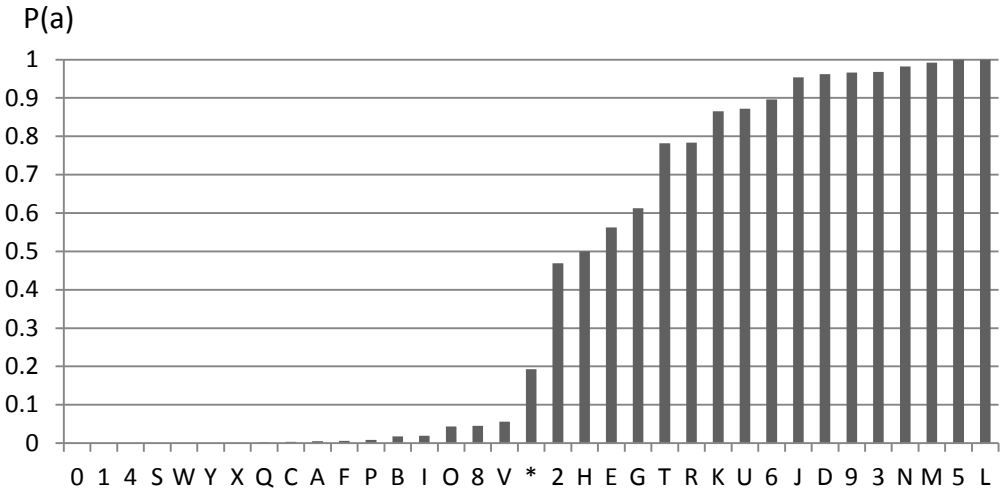
Letter Clustering

P(letter | tag)

P(tag | letter)

English

Voynich

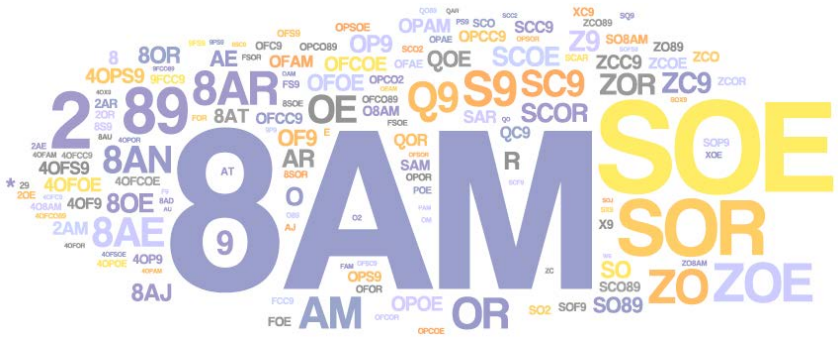


Word Clustering

Bigram model over {a, b}

a a b a b a b a a …

a →



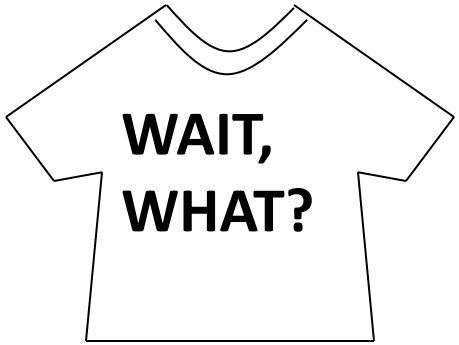
b →

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

Sample tagging with learned model:

a a a a a
VAS92 9FAE AR APAM ZOE

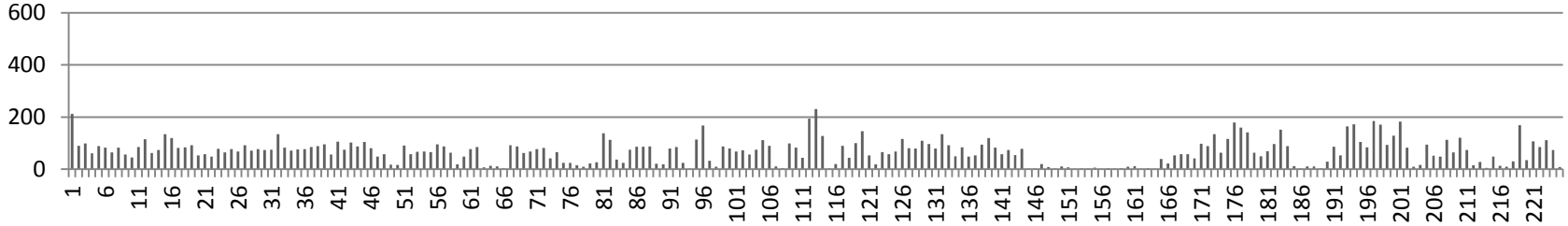
a a a a a …
ZOR9 QRC2 9 FOR ZOE89 …



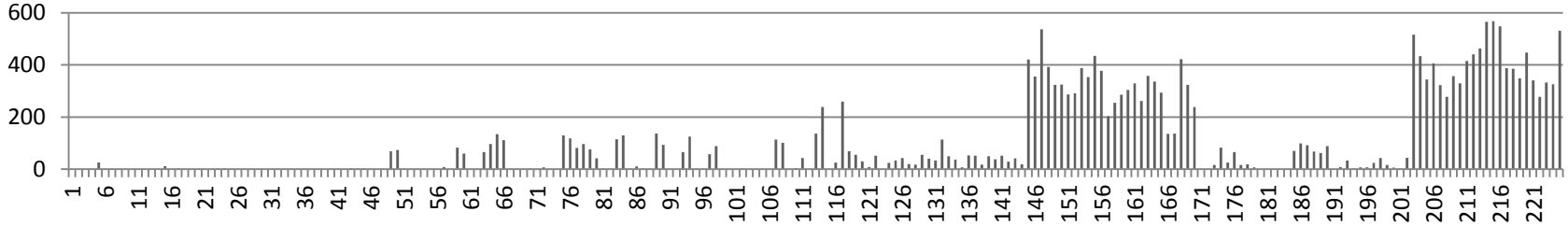
Word Clustering

Voynich words tagged as “a”

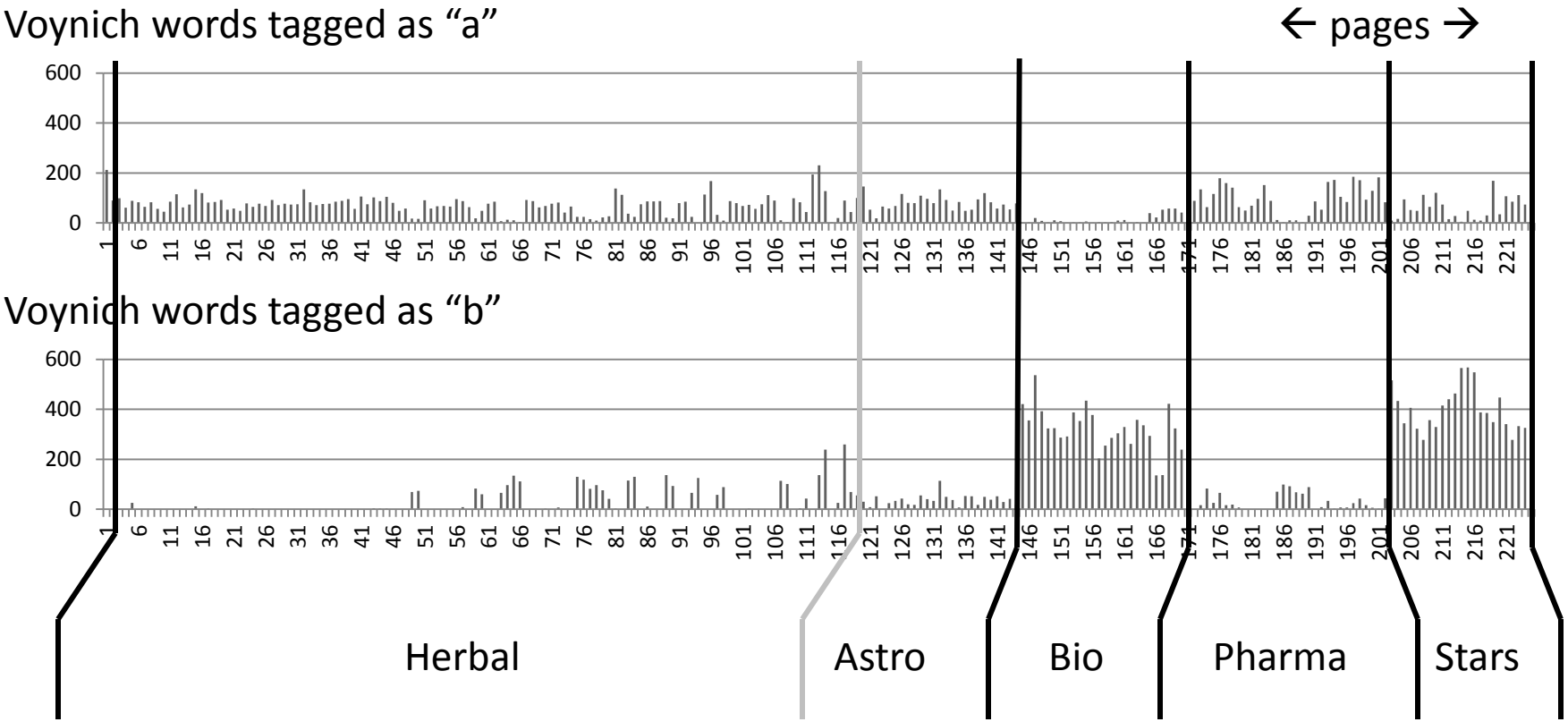
← pages →



Voynich words tagged as “b”



Word Clustering



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

1970s NSA report
recently declassified!

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.

s a general
Analyses),
aper in the
mmer, 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]

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.

??

1970s paper on HMM Voynich

1950s

1960s

1970s

1980s

1990s

2000s

2010s

2020s

Association for Computational Linguistics

1993 paper on Statistical Machine Translation

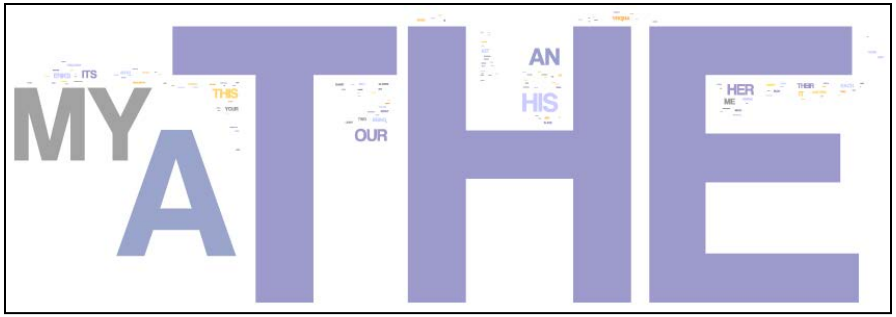
2011 paper on HMM Voynich

ACL applies statistics to language processing problems

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

"Slacker"

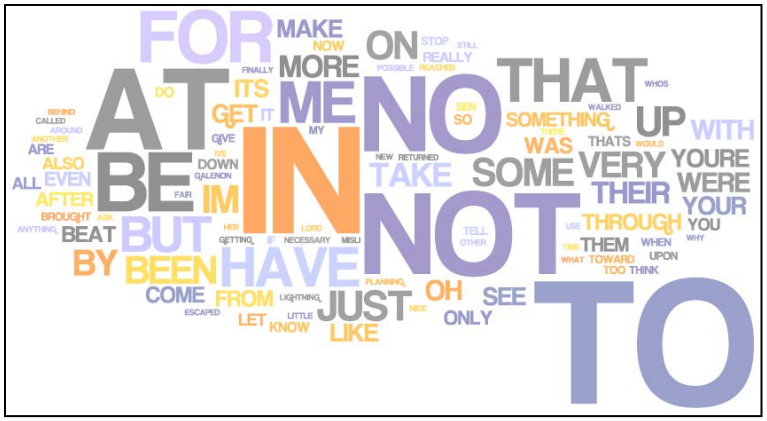
10-Class Word Clustering: English



etc

etc

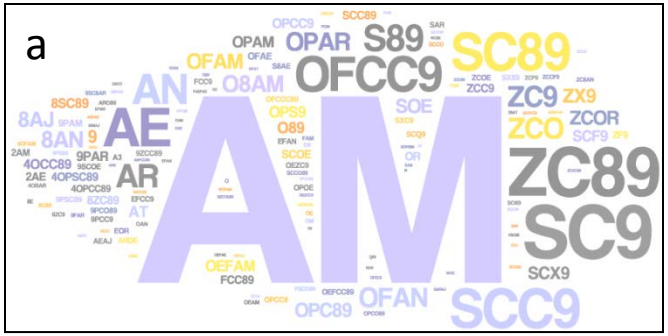
etc

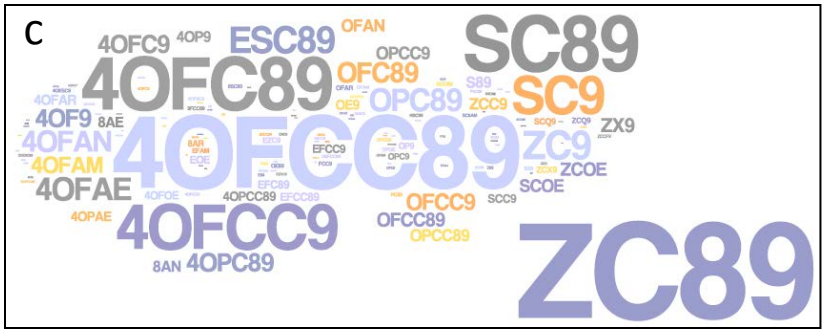


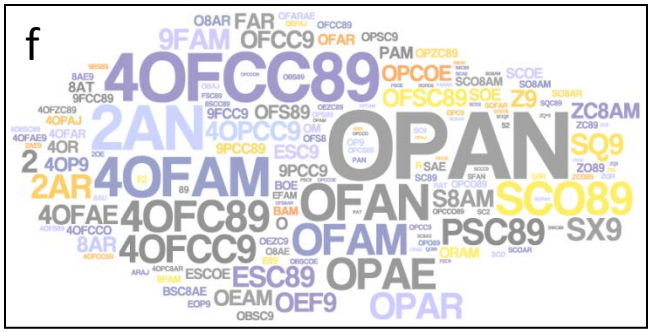
etc

etc

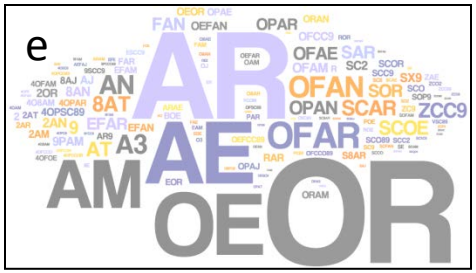
etc

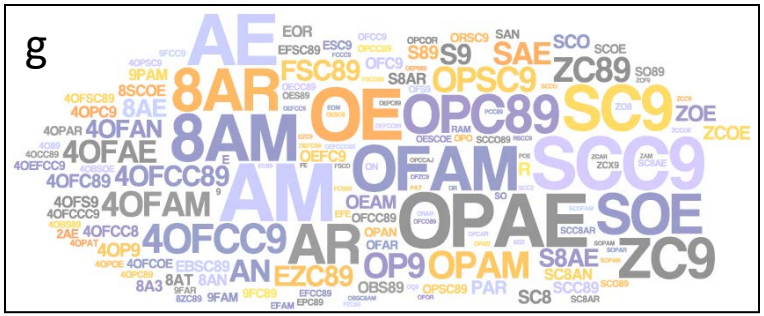


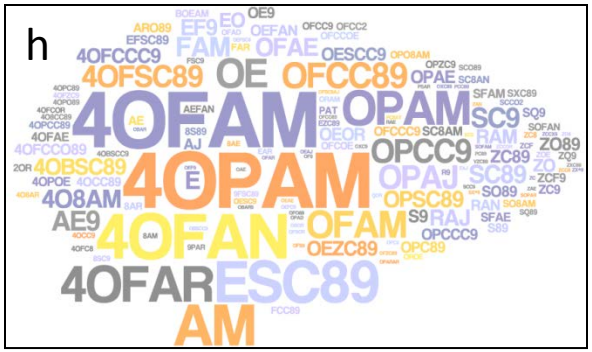


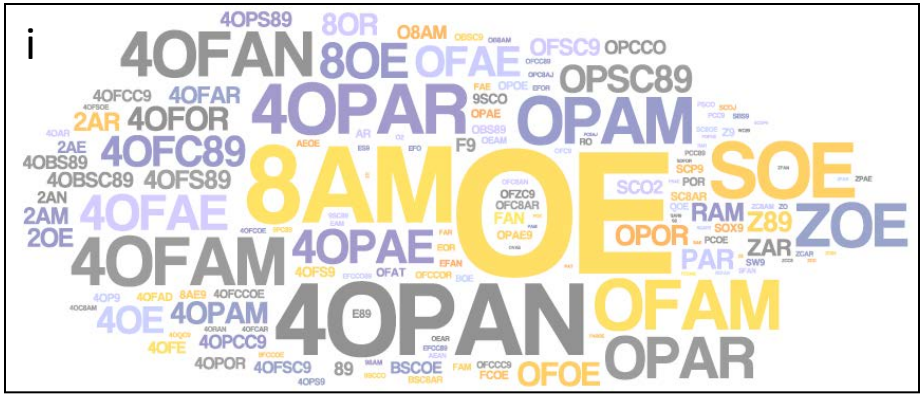


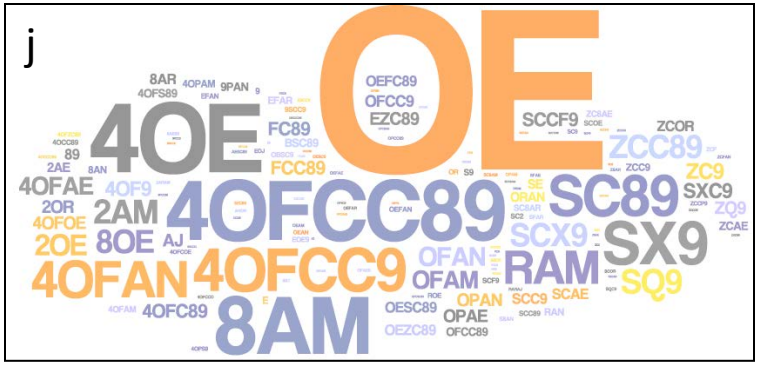
10 Classes:
Voynich-B







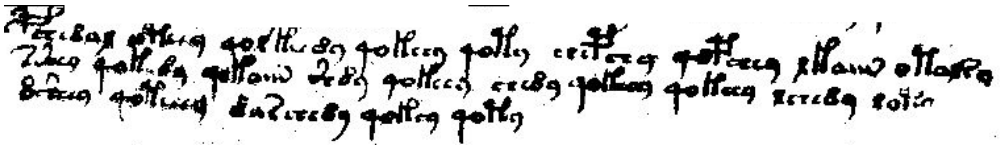




Other Unsolved Ciphers

Voynich Manuscript (1400)

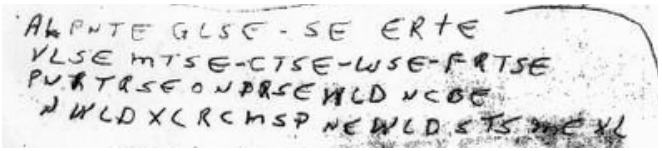
Reddy & Knight 11



Zodiac Killer (1967)

Ravi & Knight 11

FBI cipher (1999)



Kryptos (1990)

```
                                OBKR
UOXOGHULBSOLIFBBWFLRVQQPRNGKSSO
TWTQSJQSSEKZZWATJKLUDIAWINFBNYP
VTTMZFPKWGDKZXTJCDIGKUHUAUEKCAR
```

Lost Language Decipherment

Snyder, Barzilay, Knight 10

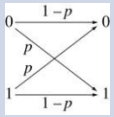
Unsupervised Training for NLP

Machine Translation …

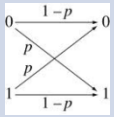
Plan for This Talk

- 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	

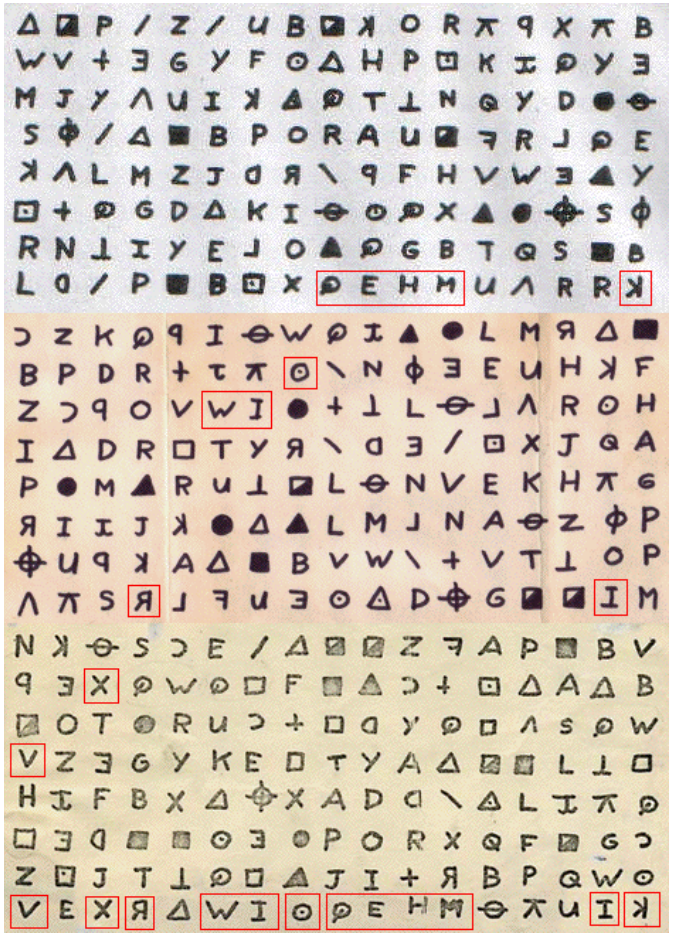
Future Prospects for Translation

	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 ??? ??? ???

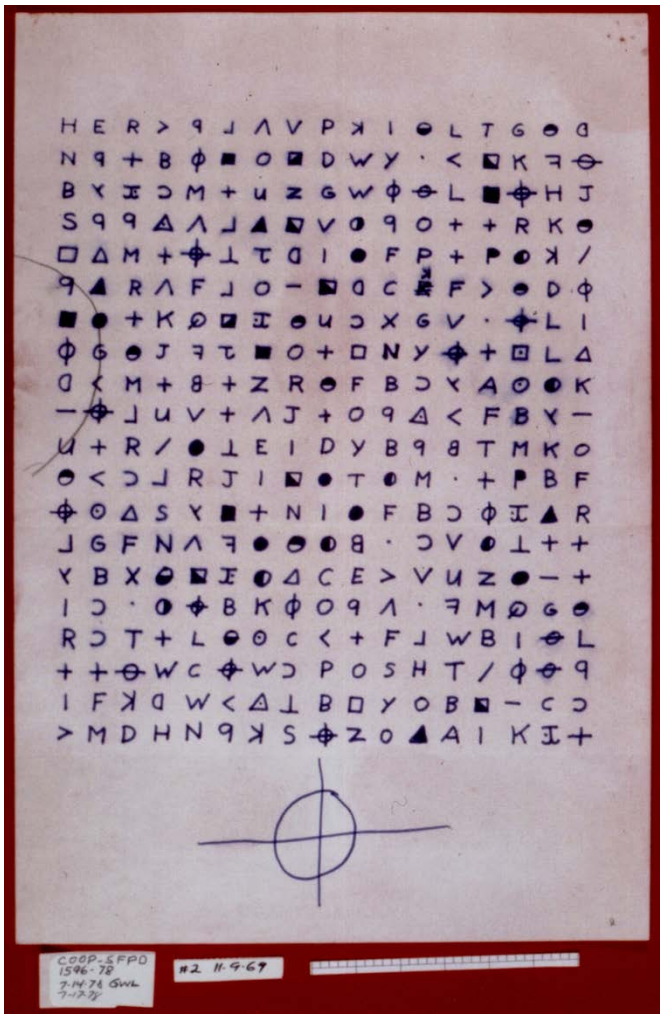
end

Zodiac Killer Ciphers

Zodiac 408 (solved, 1969)

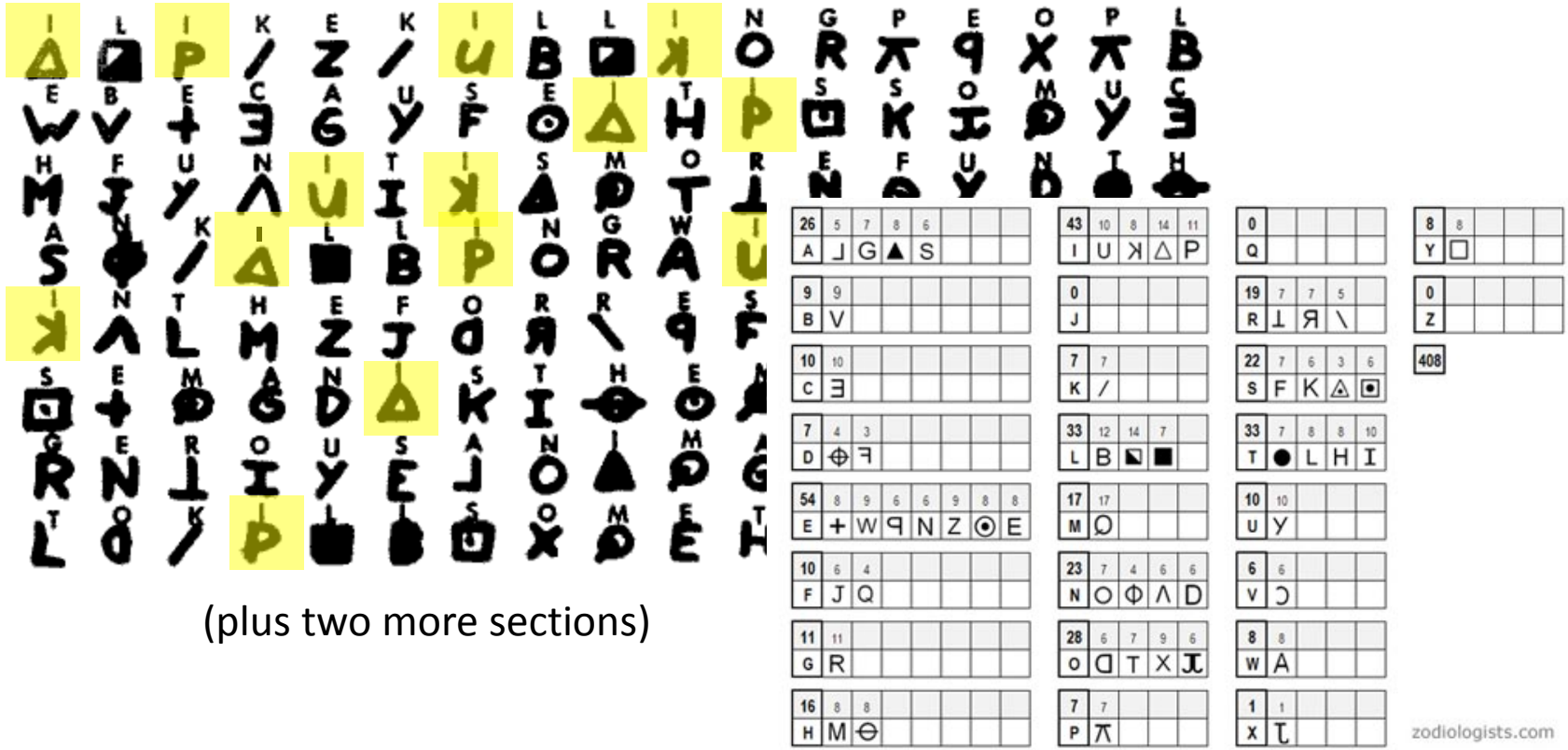


Zodiac 340 (still unsolved)



Zodiac Serial Killer

408-letter cipher (solved):



(plus two more sections)

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 EBEORIETEMETHHPITI”

Plaintext has many misspellings

Final 18 plaintext characters of 408 are “junk”

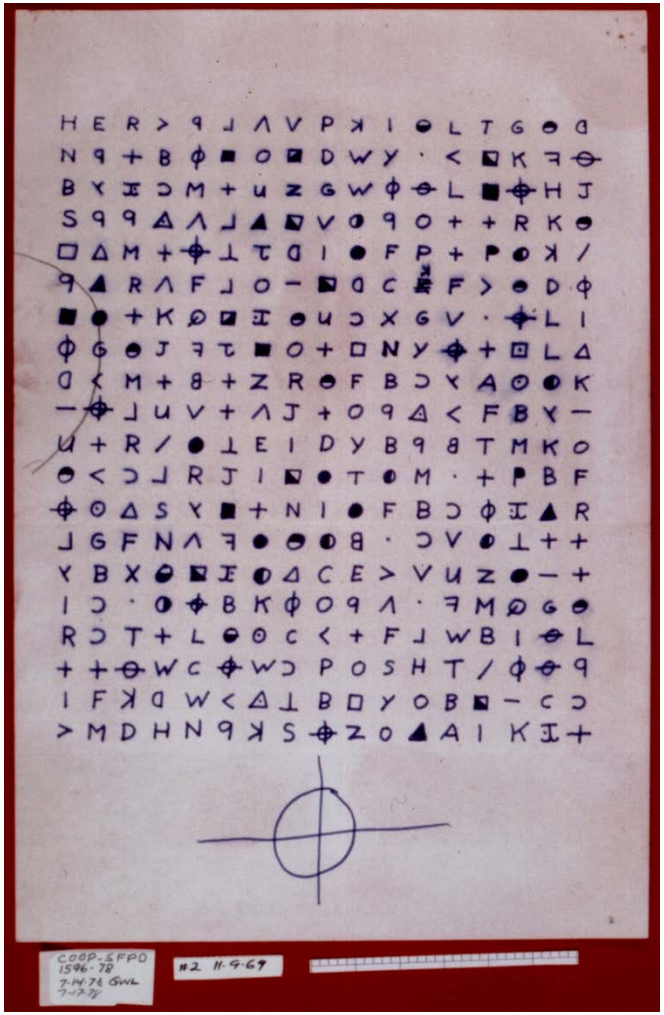
Deciphering Zodiac 408 Bayesian models

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!
Interpolated 5-gram and word 1-gram	Random	79.2
"	5-gram solution	**3.3 / 2.6**

[Ravi & Knight 11]

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.

Decipherment Papers by ACL-ers

- "Unsupervised Analysis for Decipherment Problems," (K. Knight, A. Nair, N. Rathod, and K. Yamada), Proc. ACL-COLING, 2006. (Rejected four times previously, but OK!)

- "Attacking Decipherment Problems Optimally with Low-Order N-gram Models," (S. Ravi and K. Knight), *Cryptologia*, 2009.
- "Probabilistic Methods for a Japanese Syllable Cipher," (S. Ravi and K. Knight), Proc. ICCPOL, 2009.

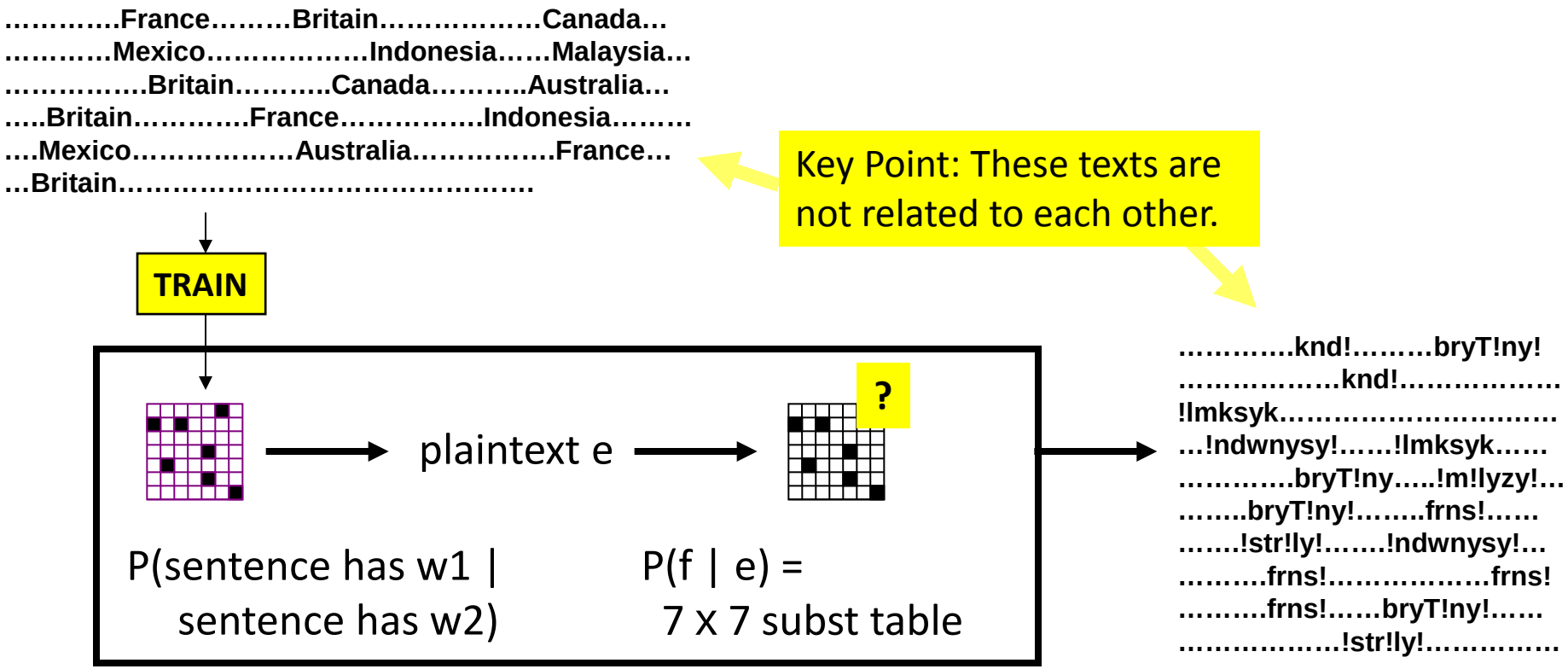
- "A Statistical Model for Lost Language Decipherment," (B. Snyder, R. Barzilay, and K. Knight), Proc. ACL, 2010.
- "An Exact A* Method for Deciphering Letter-Substitution Ciphers," (E. Corlett and G. Penn), Proc. ACL, 2010.

- "Deciphering Foreign Language," (S. Ravi and K. Knight), Proc. ACL, 2011.
- "The Copiale Cipher," (K. Knight, B. Megyesi, and C. Schaefer), Proc. ACL BUCC, 2011.
- "Bayesian Inference for Zodiac and Other Homophonic Ciphers," (S. Ravi and K. Knight), Proc. ACL, 2011.
- "What We Know About the Voynich Manuscript," (S. Reddy and K. Knight), Proc. ACL LaTECH, 2011.
- "Simple Effective Decipherment via Combinatorial Optimization," (T. Berg-Kirkpatrick and D. Klein), Proc. EMNLP, 2011.

- "Decoding Running Key Ciphers," (S. Reddy and K. Knight), Proc. ACL, 2012.
- "Large Scale Decipherment for Out-of-Domain Machine Translation," (Q. Dou and K. Knight), Proc. EMNLP, 2012.
- "Deciphering Foreign Language by Combining Language Models and Context Vectors," (M. Nuhn, A. Mauser, and H. Ney), Proc. ACL, 2012.

- "Decipherment Complexity in 1:1 Substitution Ciphers," (M. Nuhn, and H. Ney), Proc. ACL, 2013.
- "Beam Search for Solving Substitution Ciphers," (M. Nuhn, J. Schamper, and H. Ney), Proc. ACL, 2013.
- "Scalable decipherment for machine translation via hash sampling," (S. Ravi), Proc. ACL, 2013.
- "Unsupervised Consonant-Vowel Prediction over Hundreds of Languages," (Y. Kim and B. Snyder), Proc. ACL, 2013.

Word Substitution Cipher

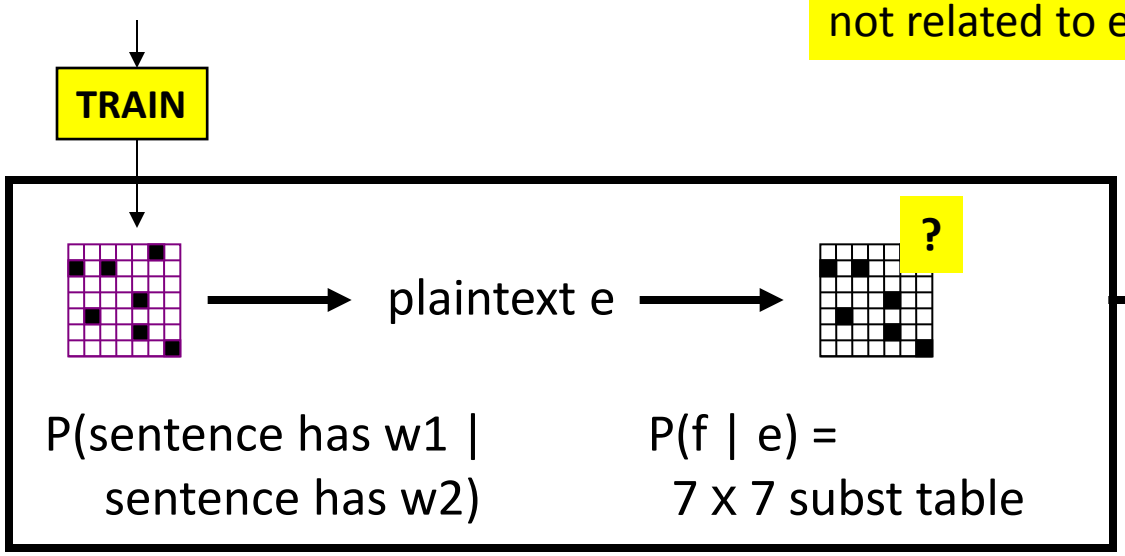


[Knight et al 06]

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.



…….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]