

Why Decipherment?

- It's fun and cool
 - ancient languages
 - secret societies
- Breaking codes was the first application of NLP
- Intellectual root of NLP
 - language models, log-odds ratios, smoothing
 - ASR and MT use "decoders"
- View foreign language as a code for English

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.

Outline

- Classical military/diplomatic ciphers (15 mins)
- Foreign language as a code (10 mins)
- Automatic decipherment (55 mins)
- **Break** (30 mins)
- Unsolved ciphers (40 mins)
- Writing as a code for speech (20 mins)
- Undeciphered writing systems (15 mins)
- Conclusions (15 mins)

Classical military/diplomatic ciphers

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."

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

#

.

.

V

##

#

.

.

.

V

#####

.

.

V

###

.

.

V

###

V

.

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	### 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	### V
U	
V	
W	1 .
X	5
Y	7 ### V
Z	2 .

Pattern word dictionaries

KDCY **LQZKTLJKX** CY MDBCYJQL: "TR

abcdeafdg

HYD FKXO

abnegated
abnegates
advocator
airedales
alienages
alienated
alienates
amperages
cadencies
capricorn
cogencies
escapeway
healthily
imbeciles
imperiled
incurious
inherited
injurious
landslide
octagonal
oklahoman
overboard
repairman
sacristry
unrebuked
unsecured

MKX **RLQQIQ** HYDL

abccdc

MKL DXCT

CDLQ JQI KTM

basses
bassos
bosses
breeze
budded
...
cheese
cusses
dosses
finnan
fleece
fosses
freeze
...
terror
tosses
tweeze
wadded
wheeeze

PTBMYEQI KH CY LG 'L TC."

abcdefghijklm

consumptively
copyrightable
documentarily
lycanthropies
musicotherapy
semivoluntary
subordinately
unpredictably

OR, NORWEGIAN!

filmprodusent
kurspamelding
publikasjoner
stylemarginpx
upproblematisk

Fundamental Questions

- How much English does a system need to know to break a cipher?
- How long does the cipher need to be, to admit a unique solution?
- How much computational effort is required to decipher?

and...

How to Make Things Harder?

- Homophonic cipher
 - ciphertext values from 00 to 99
 - A → 02, 14, 16, 22, 49, 51, 58, 90
 - B → 04, 76
 - C → 15, 56, 71
 - etc
 - flattens out ciphertext distribution
 - “a cab...” becomes “22 56 14 04...”
 - still deterministic in the deciphering direction
- Polyalphabetic ciphers
 - the secret key changes at each plaintext letter token
 - e.g., rotate through 10 different keys
- Transposition ciphers

or perhaps:

A = 8 i l y r

B = u

C = o n

D = f

E = x f A k f t z 3

F = p

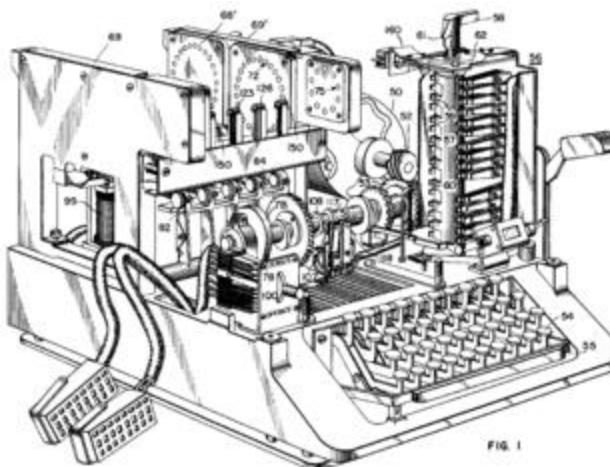
G = y ...

Cipher Types

- http://cryptogram.org/cipher_types.html
 - documents ~70 types
- E.g., RUNNING KEY cipher
 - key = agreed-upon standard English text
 - $\text{ciphertext}(i) = [\text{plaintext}(i) + \text{key}(i)] \bmod 26$
 - effectively uses 26 substitution keys
 - breakable!
 - we search for a key and (resulting) plaintext that are both natural language

How to Make Things Efficient?

- Mechanical encryption/decryption devices



German Enigma Machines (1926-45)

Substitution system

$N \rightarrow J$

Substitution table changes
with every keystroke:

$NNN \rightarrow JTE$

Rotates through 1000s of
substitution keys.



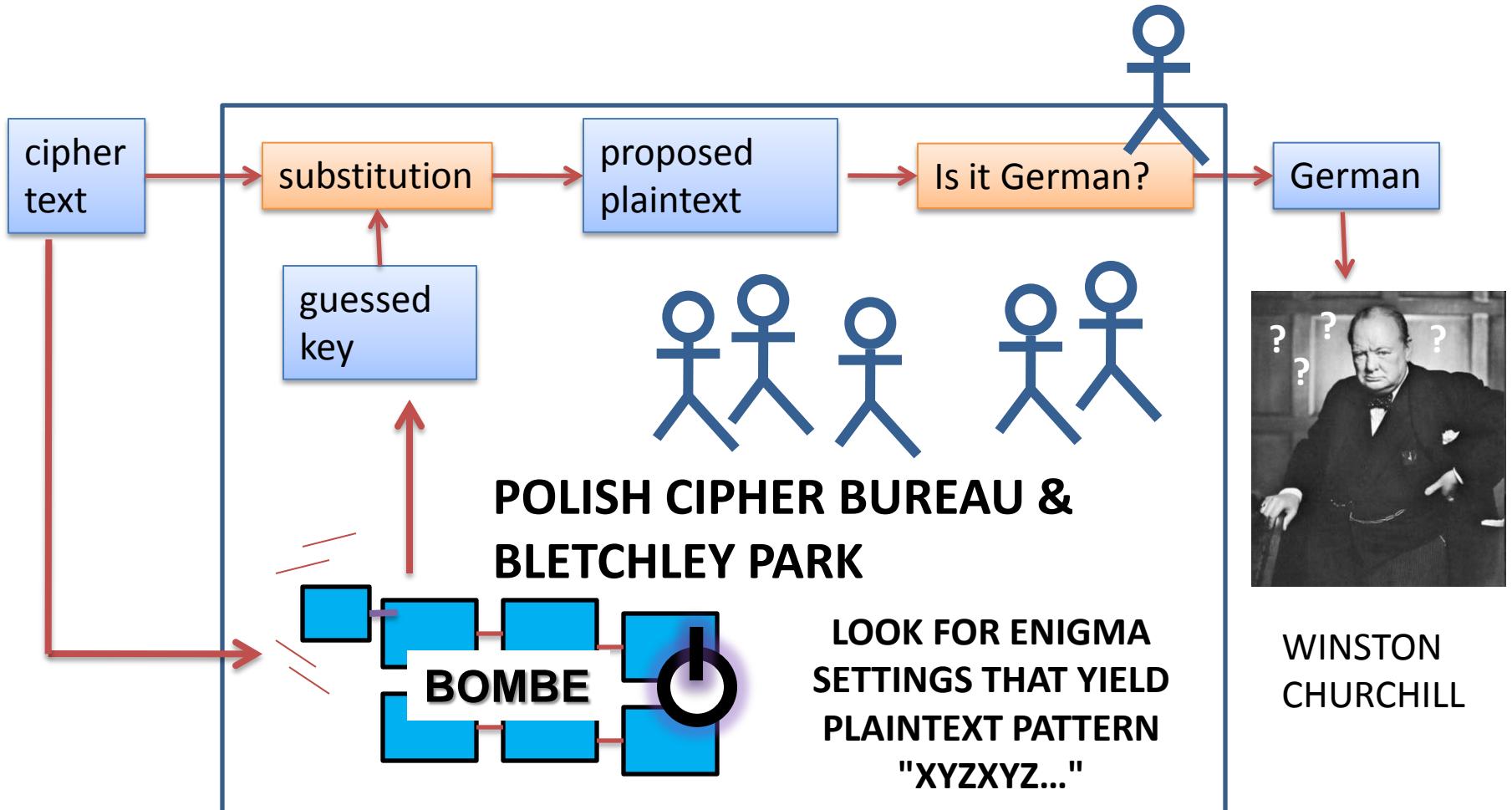
Secret key =
initial rotor
ordering and
settings

>Billions of initial
configurations.

Reversible behavior

$NNN \rightarrow JTE \rightarrow NNN$

Breaking Enigma



Enigma

- Mathematical breakthroughs:
 - Log-odds for weight of evidence [Good, Turing]
 - Smoothing with prior [Good, Turing]
 - Information theory [Shannon]
- 1945: War ends
- 1973: Wartime Enigma decipherment leaked
- 1975: Last surplus Enigma given to developing countries
- 1996: One Turing Enigma treatise declassified
- 2012: Another declassified (but have to go to England)

elegant,
powerful,
widely-applicable
mathematics

Turing Enigma Treatise

(aka NR 964, Box 201, RG 457, aka "The Prof's Book")

140pp (written sometime between 1939 and 1942)

One method is to try independently all the possible positions for the middle wheel. We shall want to know the middle wheel couplings which are consequences of these various assumptions. This can be done by using inverse rods for the middle wheel. The rods are paired off in pairs of R.H.W. couplings, i.e. M.W. output. This has been done for the case of fx, ep which arose in the DANZIGVON crib in Fig 55, assuming that the U.K.W. does not rotate. The pairs in each column of this set up give possible M.W. couplings. We have now to find out whether these couplings are good. Our procedure is rather different according as the U.K.W. does or does not rotate. In the case that the U.K.W. does not rotate it will be necessary to make a Foss sheet (the rows and columns lettered preferably with the diagonal alphabet) in which, in the RW square are entered the positions of the left hand wheel at which the RW is one of the pairs in the L.H.W. output alphabet Fig 51. This is known as the 'short catalogue' for this wheel.

elegant,
powerful, war-winning
widely applicable
mathematics

if we worked this
hard on machine
translation ...



Foreign language as a code

Alan Turing, on Thinking Machines

Instead we propose to try and see what can be done with a 'brain' which is more or less without a body, providing at most, organs of sight speech and hearing. We are then faced with the problem of finding suitable branches of thought for the machine to exercise its powers in. The following fields appear to me to have advantages:-

- (i) Various games e.g. chess, noughts and crosses, bridge, poker.
- (ii) The learning of languages.
- (iii) Translation of languages.
- (iv) Cryptography.
- (v) Mathematics.



of these (i), (iv), and to a lesser extent (iii) and (v) are good in that they require little contact with the outside world. For instance in order that the machine should be able to play chess its only organs need be 'eyes' capable of distinguishing the various positions on a specially made board, and means for announcing its own moves. Mathematics should preferably be restricted to branches where diagrams are not much used. Of the above possible fields the learning of languages would be the most impressive, since it is the most human of these activities. This field seems however to depend rather too much on sense organs and locomotion to be feasible.

The field of cryptography will perhaps be the most rewarding. There is a remarkably close parallel between the problems of the physicist and those of the cryptographer. The system on which a message is enciphered corresponds to the laws of the universe, the intercepted messages to the evidence available, the keys for a day or a message to important constants which have to be determined. The correspondence is very close, but the subject matter of cryptography is very easily dealt with by discrete machinery, physics not so easily.

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)

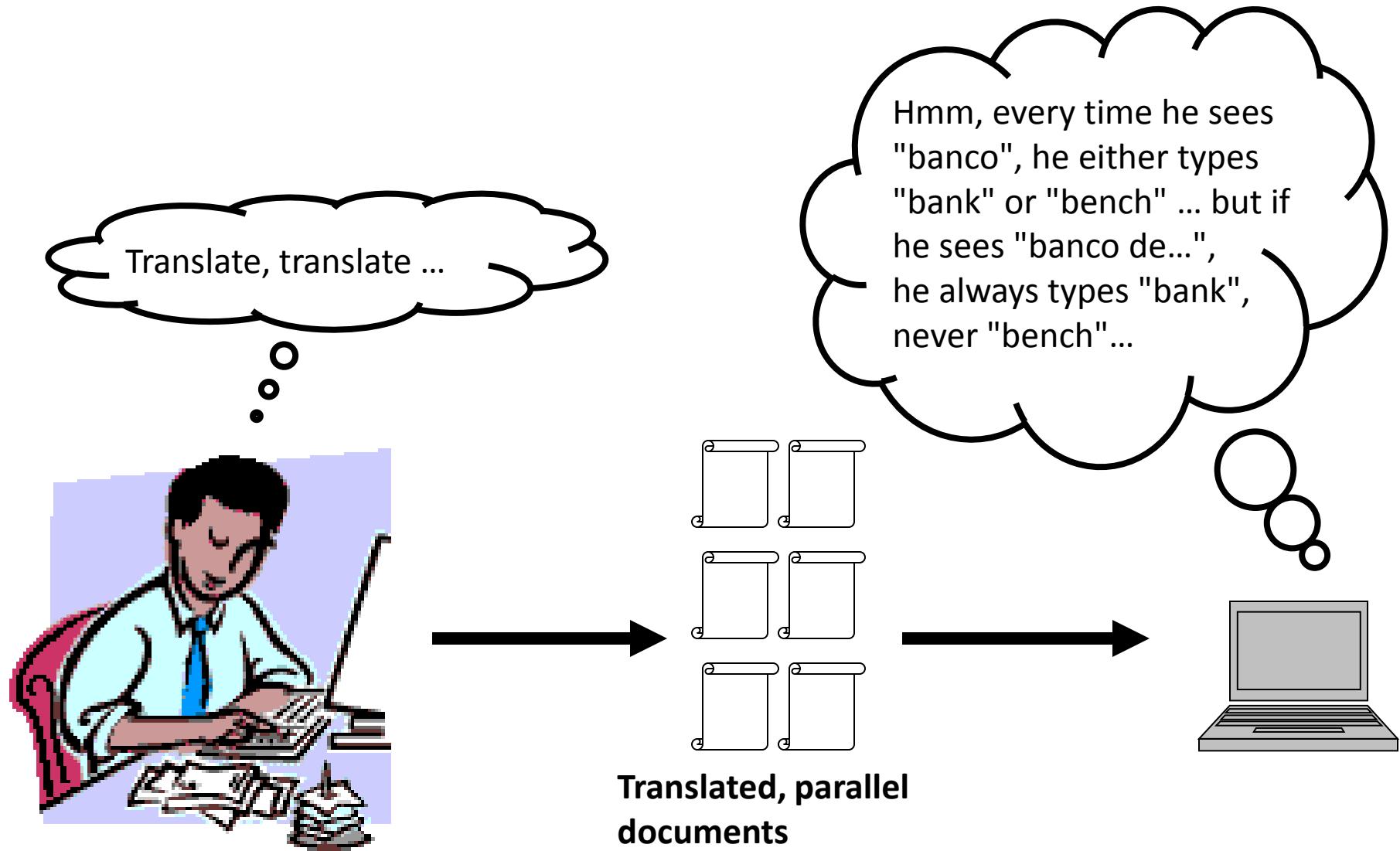


Weaver saw a colleague decoding intercepts into Turkish, without "knowing" Turkish.

... maybe a computer could translate into English without "knowing" English?

OUR HERO

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

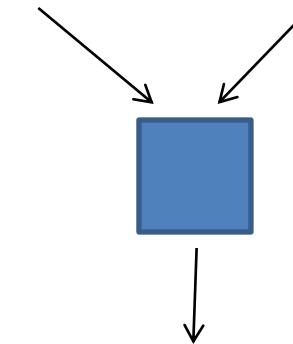
Learn Translation Knowledge from Non-Parallel Text?

English/Albanian
Parallel text



Translation model

English Albanian
text text



Translation model

Is this what Weaver had in mind?
We'll come back to this idea.

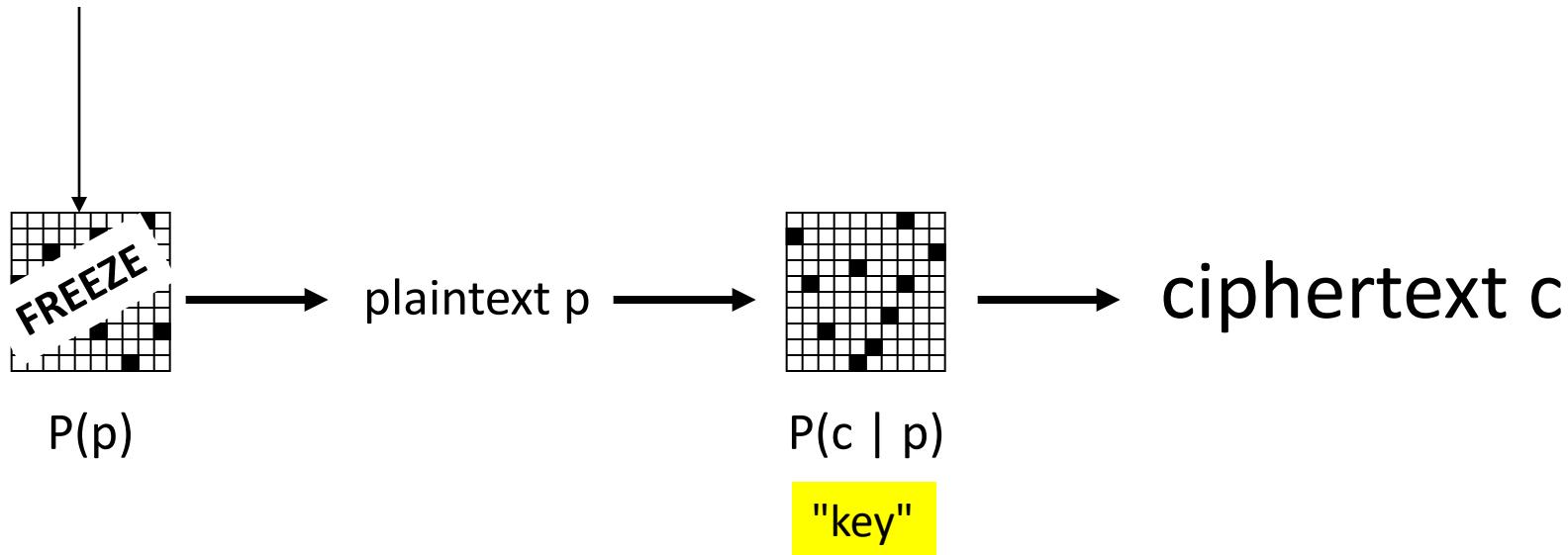
Automatic decipherment

Letter Substitution Cipher

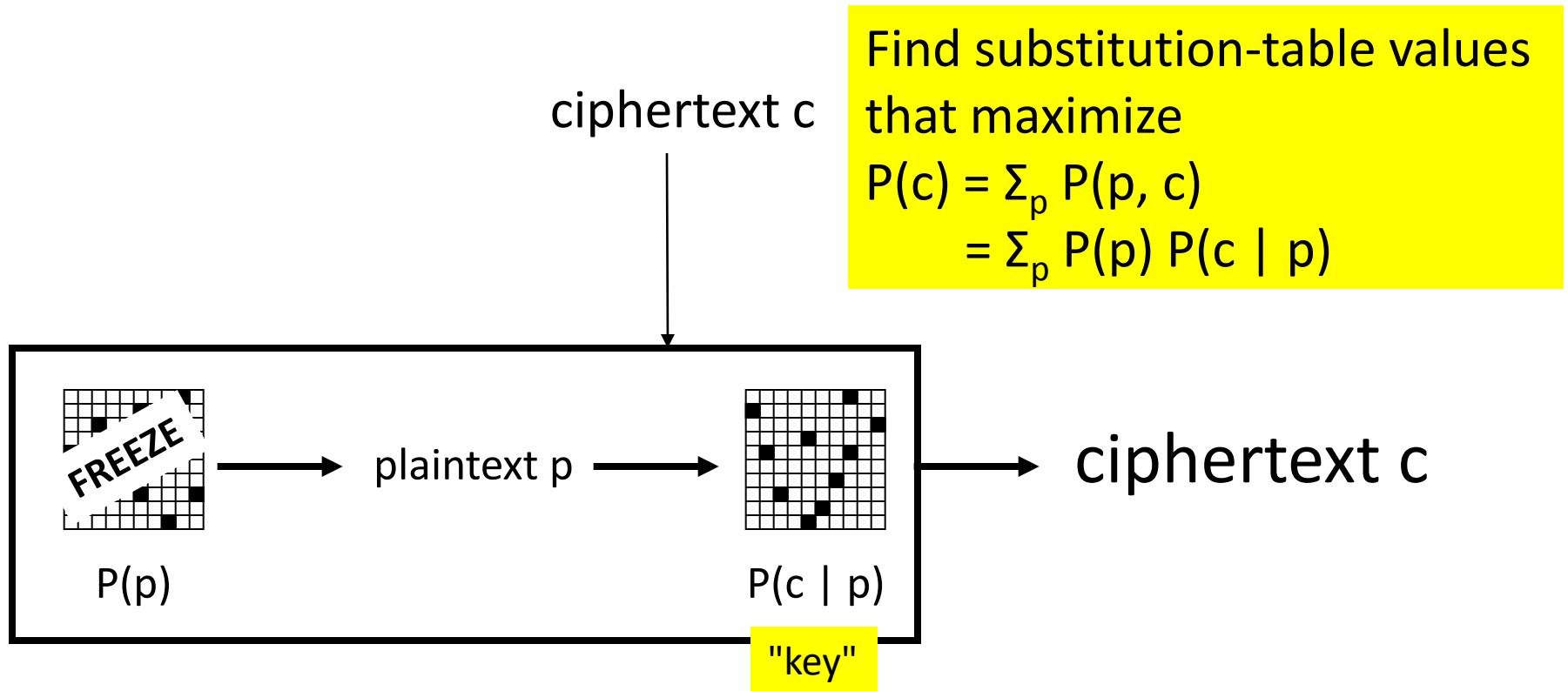
ciphertext c

Letter Substitution Cipher

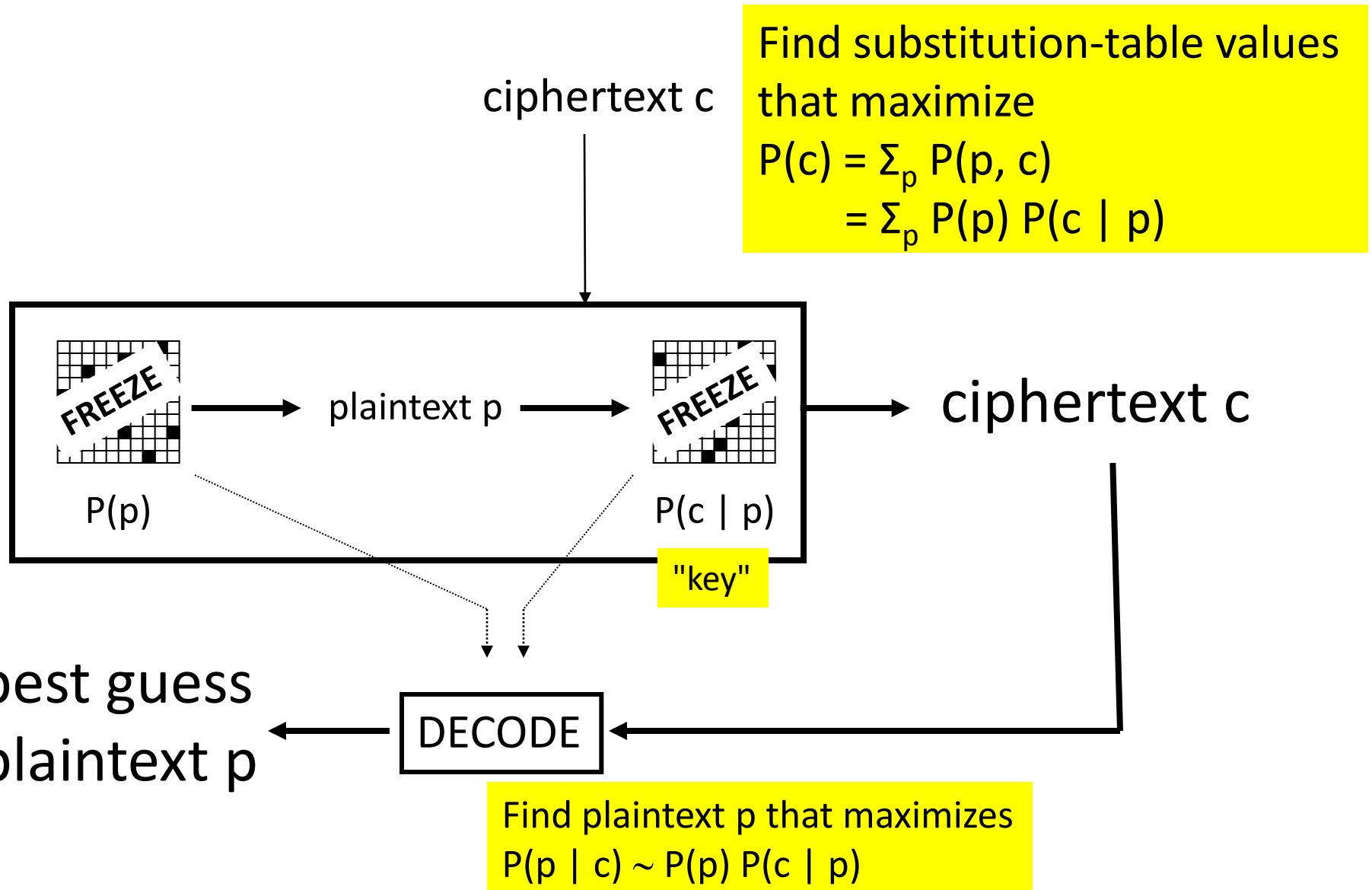
plaintext samples,
unrelated to ciphertext



Letter Substitution Cipher



Letter Substitution Cipher

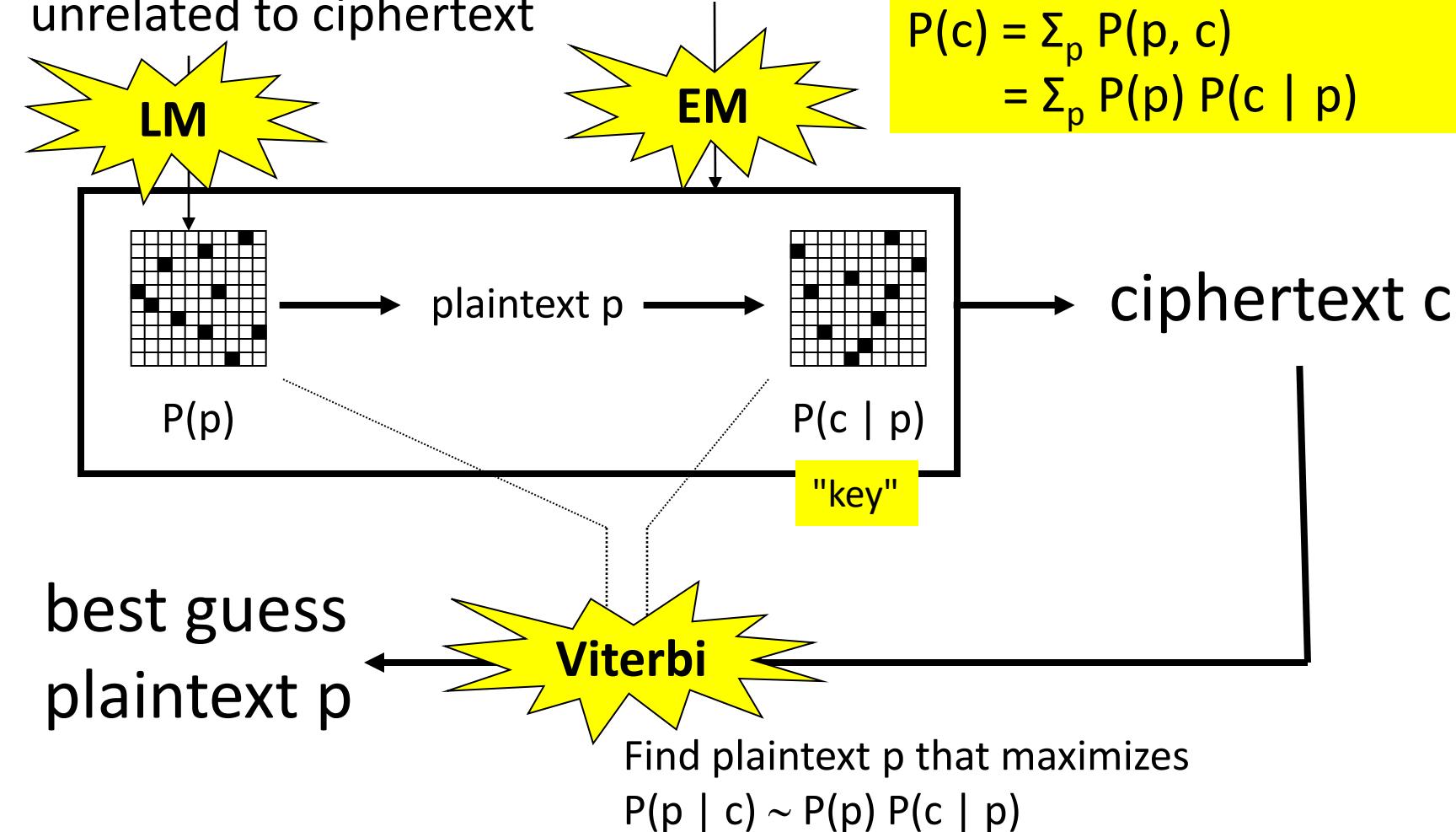


Letter Substitution Cipher

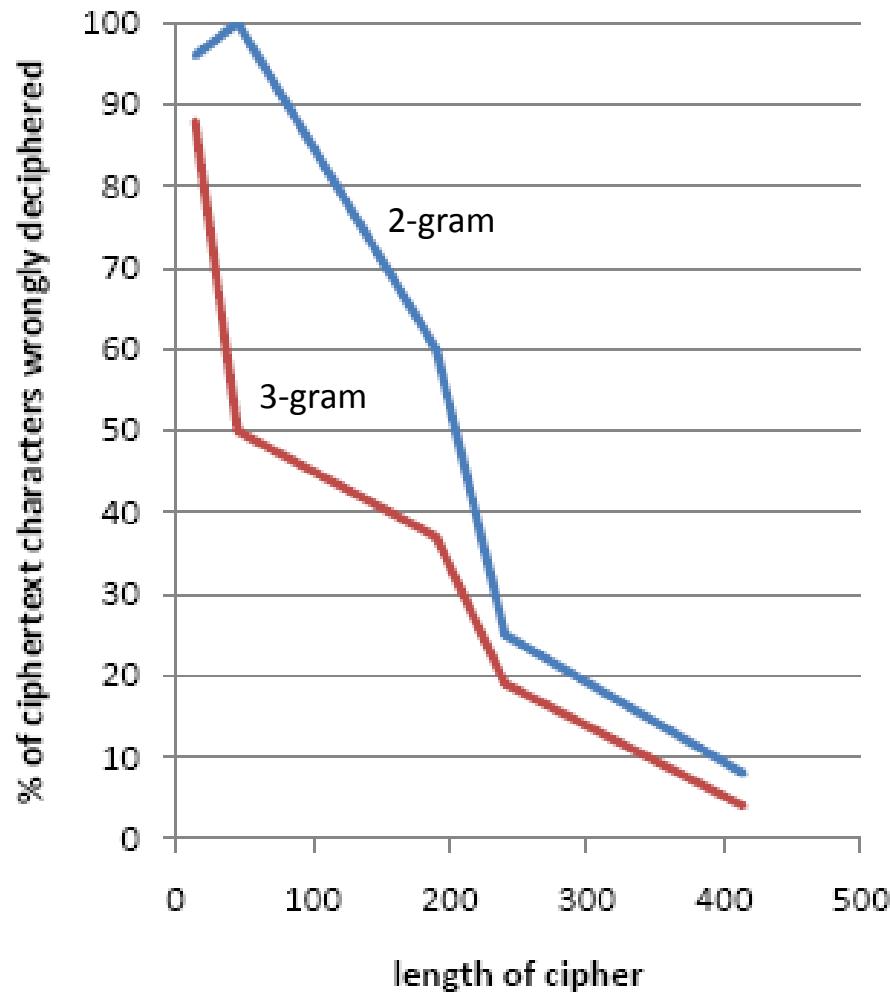
plaintext samples,
unrelated to ciphertext

ciphertext c

Find substitution-table values
that maximize
 $P(c) = \sum_p P(p, c)$
 $= \sum_p P(p) P(c | p)$



Decipherment Accuracy vs. Cipher Length



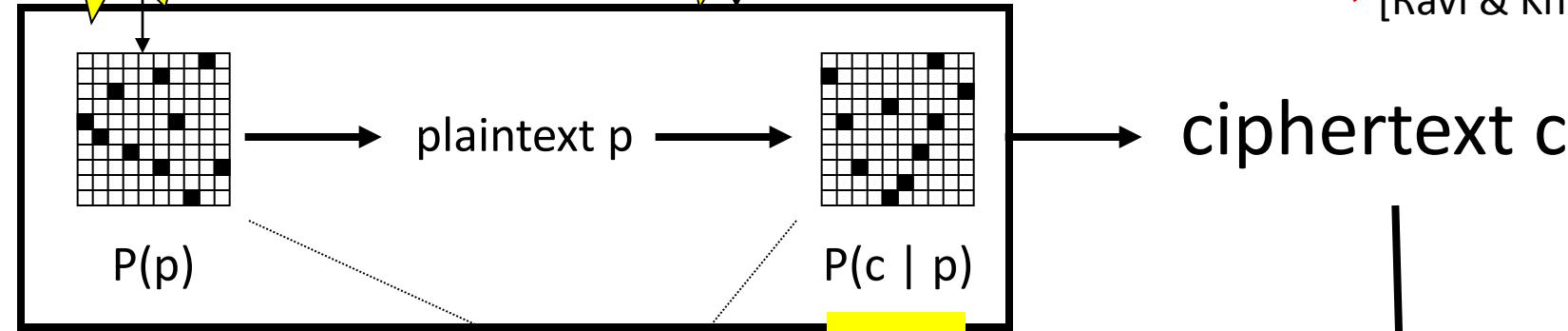
Letter Substitution Cipher

plaintext samples,
unrelated to ciphertext

ciphertext c

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]

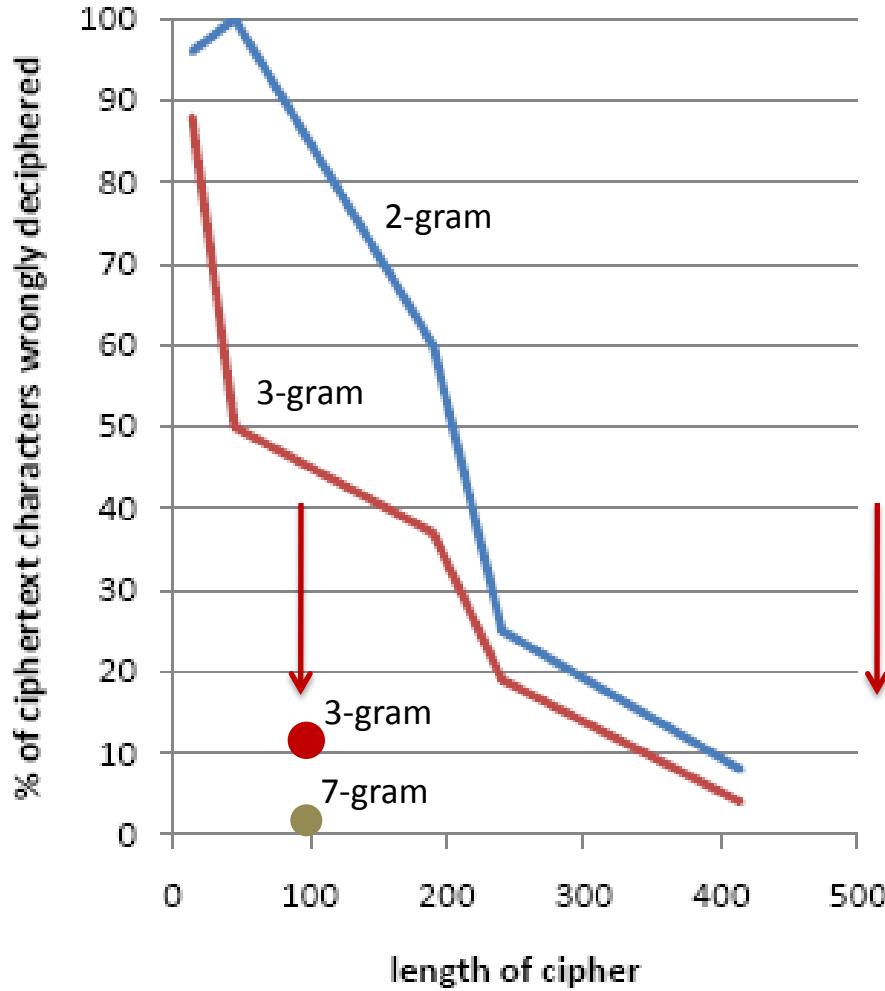


best guess
plaintext p

Find plaintext p that maximizes
 $P(p | c) \sim P(p) P(c | p)^3$

[Knight/Yamada 99]

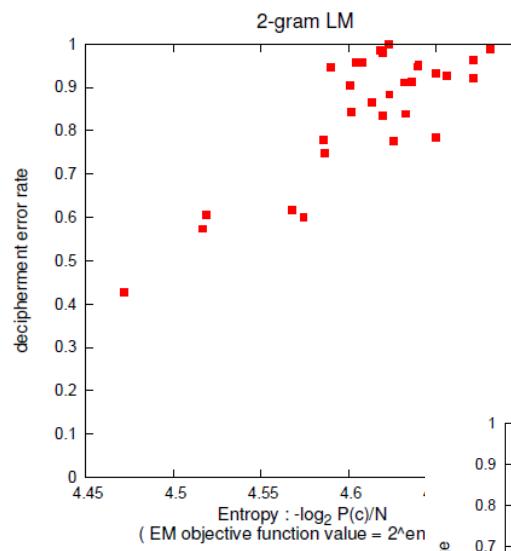
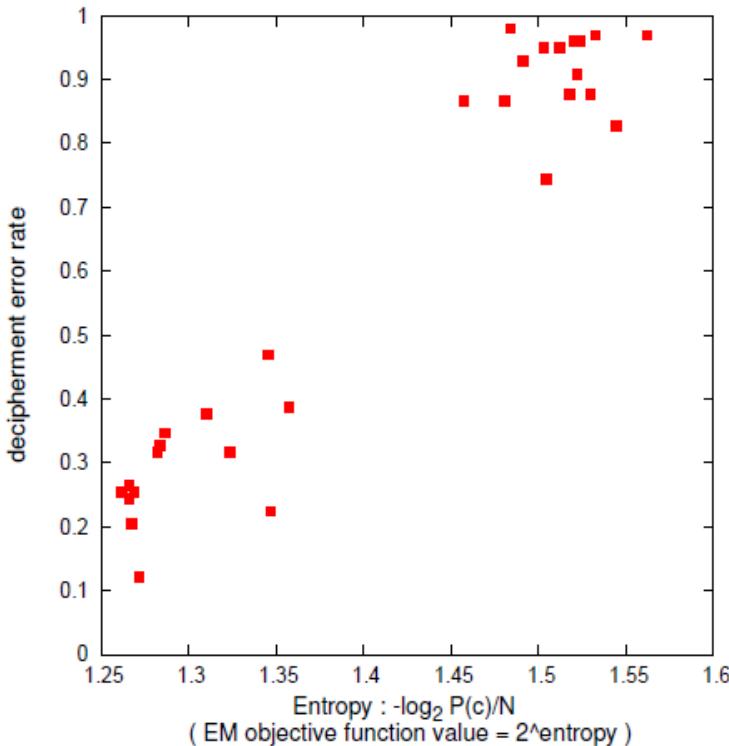
Reducing LM Weight During EM



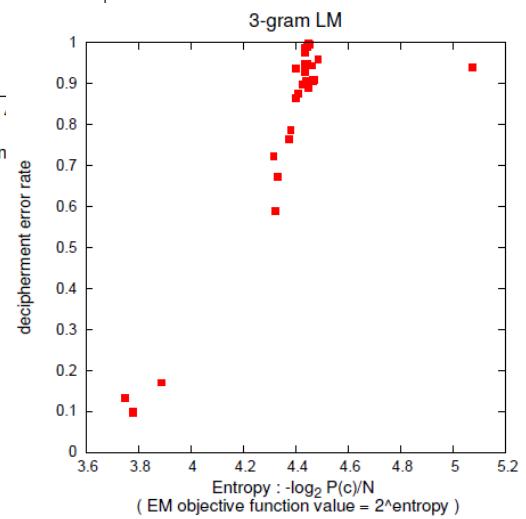
Set EM to maximize
 $P(c) \approx \sum_p P(p)^{0.5} P(c | p)$
instead of
 $P(c) \approx \sum_p P(p) P(c | p)$

Random Restarts are Critical

English 98-letter cipher, 3-gram LM



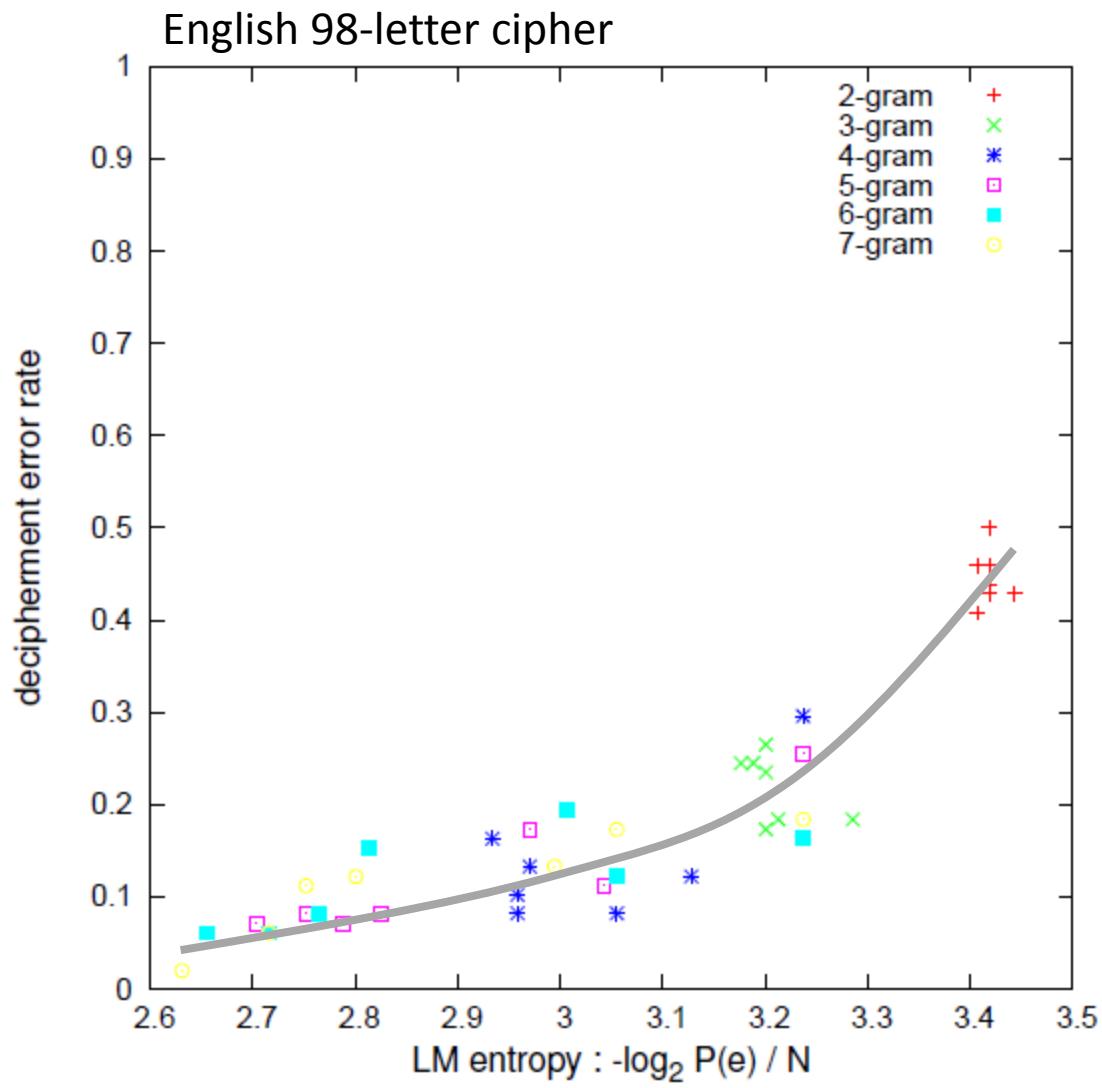
Japanese syllable cipher



even people do restarts!

[Ravi & Knight 09b]

Good Language Models are Critical



[Ravi & Knight 09b]

Searching for Deterministic Keys

- Peleg & Rosenfeld, 1979
 - relaxation search
- ...
- Ravi & Knight, 2008
 - ILP, exact search
- Corlett & Penn, 2010
 - A* exact search
- Nuhn, Schamper, and Ney, 2013
 - beam search

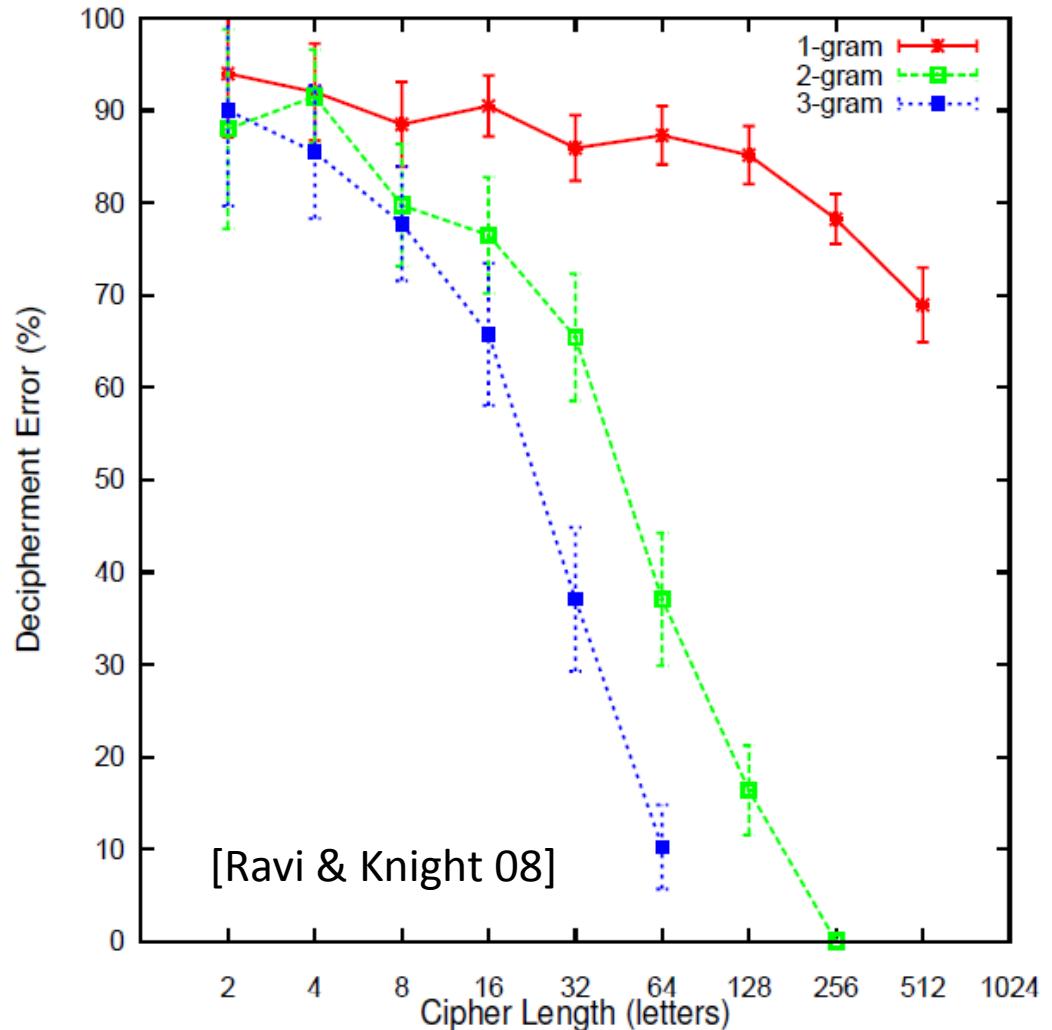
Deterministic Keys

- * Use ILP to search only deterministic keys.
- * Exact, no restarts.



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

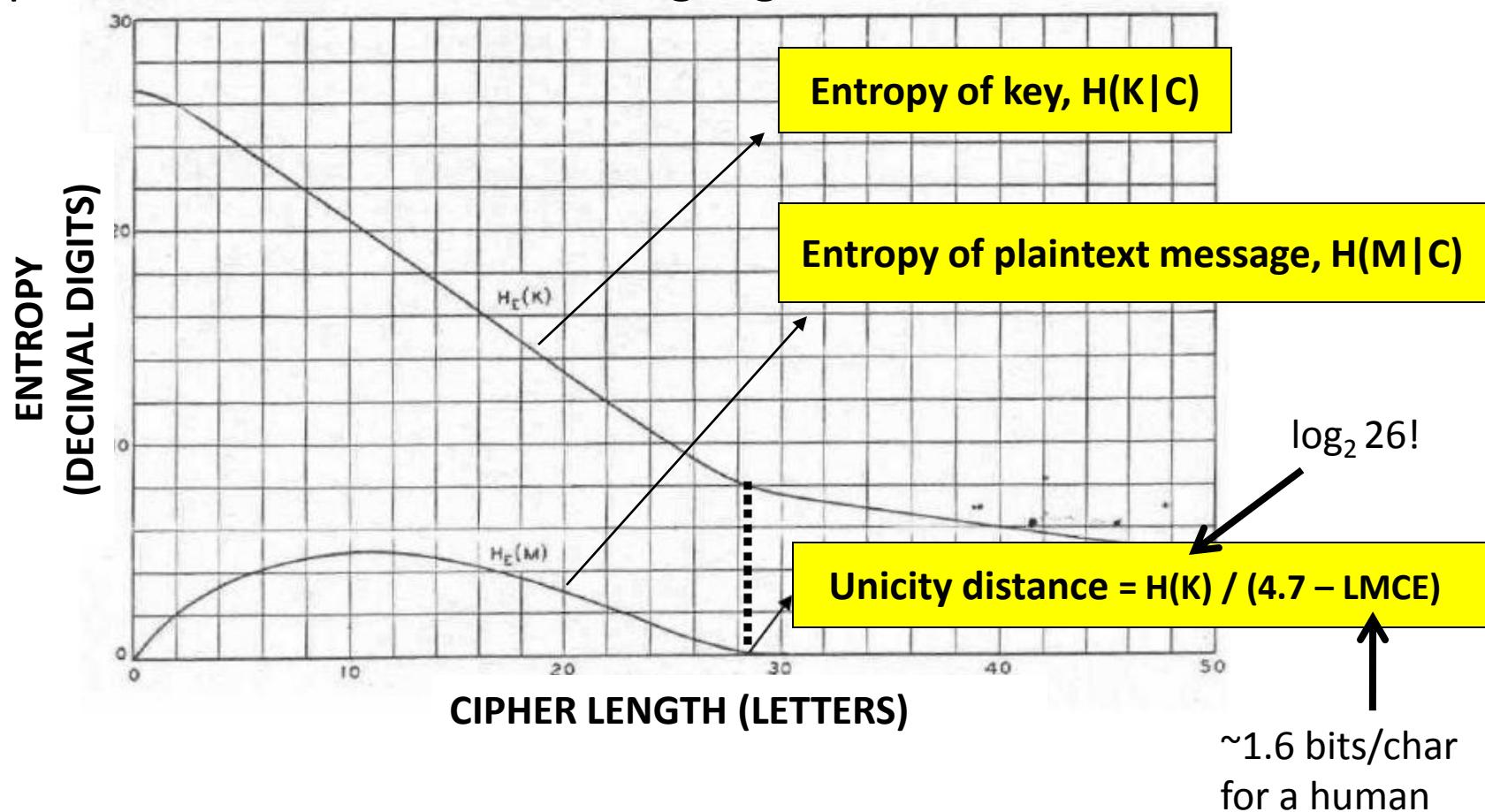
Using 2-gram letter-based LM



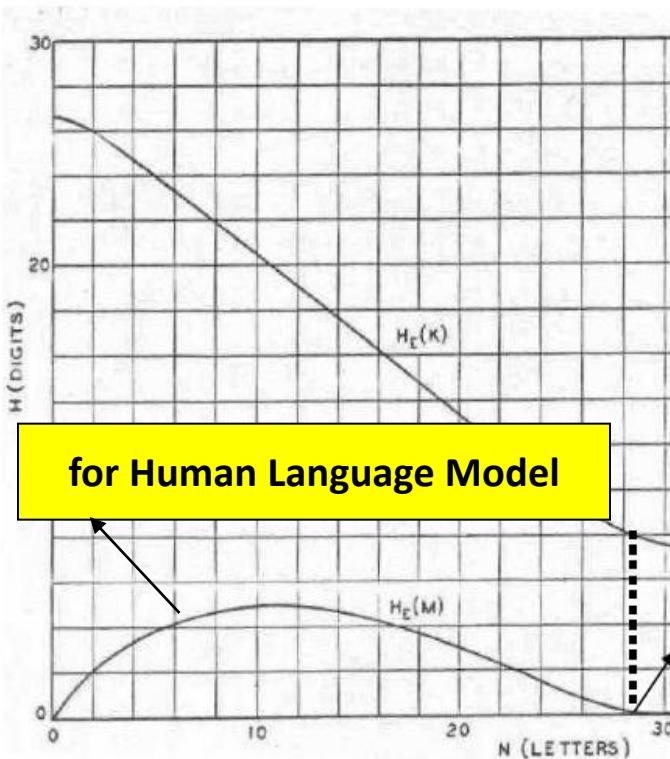
[Shannon 46, 49]

"Communication Theory of Secrecy Systems"

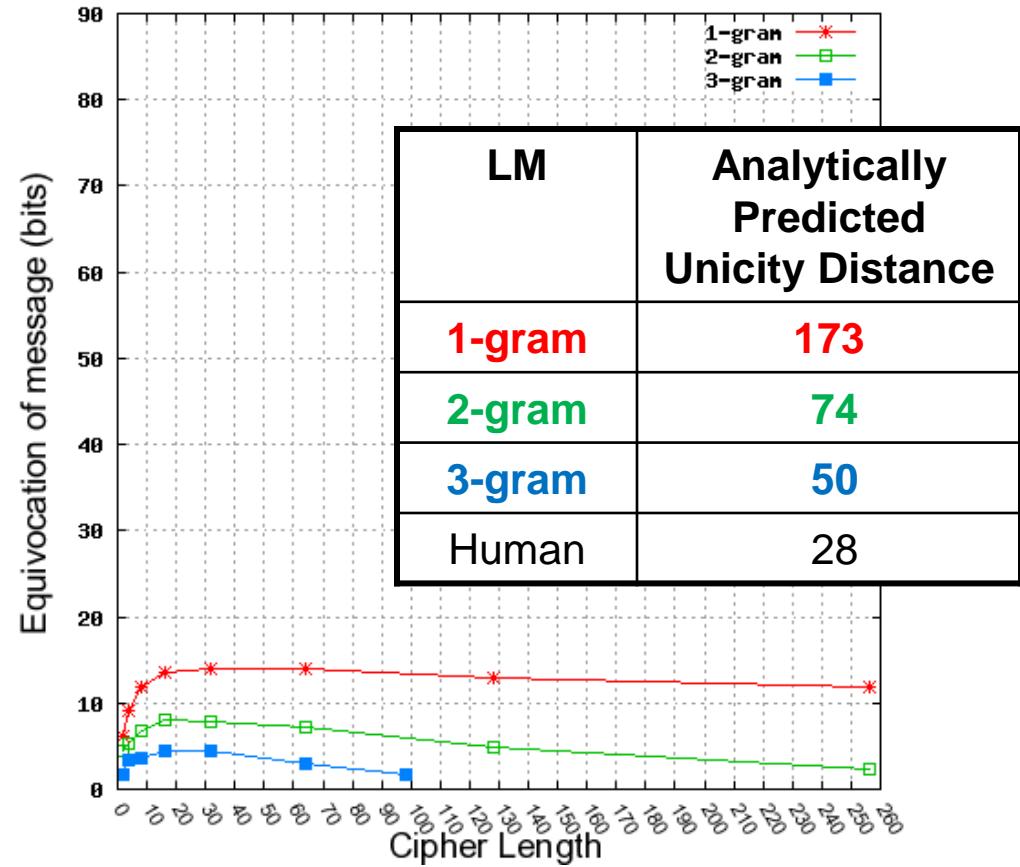
- Shannon analytically predicted uncertainty about key and message
- Graphed it for a human-level language model



Verifying Shannon's Prediction of Plaintext Message Uncertainty



ANALYTIC CURVES (Shannon)



ACTUAL CURVES

Some Recent Historical Decipherments

- Jefferson cipher (L. Smithline)
 - <http://online.wsj.com/article/SB124648494429082661.html>
 - For more than 200 years, buried deep within Thomas Jefferson's correspondence and papers, there lay a mysterious cipher -- a coded message that appears to have remained unsolved. Until now.
- Civil War ciphers (K. Boklan)
 - Cryptologia, 30:340–345
 - We study a previously undeciphered Civil War cryptogram, limiting ourselves to pencil and paper, and discover not only a missive of military importance, but in the process identify a new Confederate codeword. Our methods rely not only upon cryptanalysis of the encryption method but also on the exploitation of an elementary mistake.
- German Naval Enigma
 - <http://www.enigma.hoerenberg.com>
 - The "Breaking German Navy Ciphers" Project was founded in 2012. The goal is to break original radio messages, which were encoded with the famous German ENIGMA cipher machine. Up to now, we've succeeded in deciphering 53 original World War II Enigma M4 messages. Many of these messages had never been broken before, so you can read them for the first time in history.

Copiale Cipher

zmzfszhijxjézprémfcjizirbqjrlxjzöylzalbtralnpar
hjzéjyazut+qalpizgadpzyxháylagdraf+lamzuzcuhirppjizéjy
ej. hizpnurðcfozplgobijnjh+bzilossen=romz||blud:jj|o
cq||larihdgmpusjhrakzéluod||:zalbqzlpizb||:olbýtjmlgip
mptóruod||:sénpusituj+nhcug=rzgmaazlalayg||w+qz
j|m=mptaxcicjcz||izgznccof||lbnjh+vzjy||ébgnpzljqzuihál
hizwzé||w||bcjlm||xudhlinptb||zr+pmprj+hpusijolzn||n||
gj+rvuzfizizm||zrétdriéccj:||rinq||mpáaoelptá+nvzjbfzju
pu||hs=l||oirájnrémpubfjumrcé||rleznnff||cýé||aw
zjih||bzjá||umlijpámlozgq+ligxu||hljlr||bðhj||lðaxst:||z
az+qzjumof||xmtphéylvzál||bcjlm||fj||usot||wurðcý||u
ðjzö||azá||vptod=alravlyxja||vlgz||+bdrlpnatá
mriwo||þrimamznh||Anisut||v||ða+bj=hc||upirgzn
hizfrph+abdrj:||ralölf||qnezbda||fúthzozhr:||lpo||ggz||uw
pjzárge=gptvñuzcñrzkic:||urqkipm||uzuwpz||párlzg
gptójuv||:||injxjha3ndusnpr+||a||ðdt+himc||m||izn
ðnúm:||élosz=la||jyr||vneim||:||ðhj||azqg||mpjic:||vldzalnac:||n

341

τάλαντον μείζι την πράξη καθέστως ή μηδὲν Δοδέκατον
δύεται πλην λιγοστού πιθανού πάντας οὐδεὶς αὐτόν.

'Néostlin.

Caprodēi πνηράν & Δανάϊχυλέηρbg=3ghit|

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

Copiale Cipher

Section headers

h̄īc̄īj̄ȳp̄ūt̄h̄l̄p̄īr̄ḡd̄j̄x̄b̄d̄ȳl̄ād̄f̄t̄ūm̄z̄ēc̄ūh̄īr̄p̄īr̄d̄ȳ
ē.k̄ī=p̄n̄ūr̄d̄c̄q̄p̄l̄ḡd̄j̄n̄h̄+b̄z̄īōs̄ēl̄p̄=r̄ōm̄z̄p̄l̄n̄ūd̄:h̄p̄/ō
c̄q̄l̄ār̄īh̄l̄d̄ḡp̄ūc̄l̄āl̄z̄ēl̄ūd̄p̄:z̄āl̄b̄x̄ūp̄īb̄z̄p̄:ōl̄b̄ȳh̄m̄l̄ḡp̄
m̄p̄ōr̄ūd̄p̄p̄:s̄ēp̄ūīūj̄l̄t̄īh̄c̄ūḡ=r̄z̄ḡm̄āz̄ūl̄āȳḡ/w̄q̄x̄
j̄m̄=p̄t̄x̄c̄īh̄c̄īr̄ī:īḡz̄n̄īō/l̄b̄h̄t̄+v̄z̄/l̄b̄ḡōp̄m̄z̄īj̄z̄ūh̄â̄l̄
h̄=f̄w̄z̄ē/w̄/b̄c̄n̄m̄h̄ūd̄h̄l̄īn̄p̄b̄h̄īr̄+p̄m̄p̄ī+l̄p̄ūīf̄l̄z̄n̄/n̄/j̄
ḡī+v̄ūp̄h̄z̄īz̄m̄/r̄+ēr̄īc̄p̄c̄:p̄n̄q̄p̄m̄p̄āōl̄p̄ū+v̄z̄b̄r̄īū
p̄ū/h̄s̄=l̄ōīr̄āj̄n̄r̄ēm̄ūb̄s̄h̄ūr̄c̄ēp̄l̄ēz̄n̄:f̄c̄ȳē/w̄
z̄īh̄/b̄z̄ȳ/ūm̄h̄p̄āl̄ō=ḡq̄+l̄ḡt̄x̄ūh̄āl̄p̄/b̄d̄h̄īj̄l̄ōs̄īt̄:z̄
ā+q̄z̄n̄īm̄p̄/x̄m̄p̄h̄ēȳl̄n̄z̄â̄/b̄c̄īūp̄â̄/f̄z̄ūs̄p̄ī/w̄ūr̄c̄ȳ/ū
d̄x̄ō/l̄āz̄ū/n̄p̄ōd̄=āl̄ēūl̄ūx̄j̄ā/l̄v̄l̄ḡr̄+/b̄d̄īn̄ūp̄â̄/z̄
m̄r̄īw̄ō/p̄m̄ām̄z̄h̄â̄/Δn̄īs̄ūp̄/ū/l̄d̄+b̄=h̄c̄ūp̄īr̄ḡz̄h̄
h̄=l̄r̄īh̄t̄â̄b̄d̄p̄:īr̄l̄ōūp̄īr̄ēs̄b̄d̄j̄s̄ūh̄īz̄ōz̄h̄:l̄p̄ō/ḡḡ=ūw̄
p̄īz̄â̄r̄ȳā=ḡp̄v̄īn̄z̄c̄n̄p̄z̄h̄c̄:ūr̄ḡh̄p̄m̄ô̄/ūz̄ūw̄r̄z̄p̄ūr̄=ḡz̄
ḡp̄ōj̄ūn̄/l̄īn̄j̄h̄āz̄b̄ūn̄p̄r̄+z̄/ā/d̄+h̄m̄c̄f̄m̄/f̄īz̄p̄
d̄n̄ūm̄:é̄l̄ōs̄=l̄īj̄ȳr̄/d̄ūēm̄p̄r̄:d̄h̄īāz̄ḡf̄p̄r̄īc̄:/l̄d̄ēl̄ōd̄āc̄:h̄

Some scratch-outs, rare

Preview text fragments
("catchwords")

z̄āl̄ōm̄īr̄īm̄f̄j̄t̄c̄p̄k̄x̄r̄z̄k̄d̄c̄:īx̄m̄z̄h̄ô̄Δd̄b̄é̄f̄c̄:ū
d̄ȳēc̄:â̄īr̄n̄l̄ī=p̄r̄ōȳp̄b̄+t̄h̄c̄s̄â̄n̄l̄ūx̄z̄h̄īc̄l̄ūb̄.
Δp̄á̄ō+d̄x̄/īc̄:ūb̄.
Δh̄īz̄j̄k̄p̄c̄â̄p̄l̄m̄p̄ȳé̄r̄ḡō+c̄p̄p̄h̄īr̄j̄ȳ/ū=m̄b̄z̄īz̄t̄
ūp̄x̄:īz̄k̄z̄h̄b̄d̄ēȳr̄é̄ḡc̄â̄h̄īr̄x̄d̄l̄s̄z̄x̄h̄ūs̄h̄īr̄p̄āx̄-m̄p̄ūp̄
w̄p̄ē/ūc̄p̄h̄s̄ēm̄l̄ī:p̄r̄ēp̄j̄f̄h̄ḡz̄īp̄r̄l̄b̄n̄ȳz̄z̄n̄ūp̄d̄s̄ūp̄
h̄n̄ȳḡd̄/ūp̄m̄â̄x̄k̄p̄ōc̄īâ̄p̄ȳb̄s̄ūh̄c̄f̄ȳz̄=ūx̄ūp̄ūn̄f̄=z̄l̄s̄b̄x̄
r̄p̄ȳé̄b̄m̄ī:ūd̄h̄īm̄z̄t̄x̄h̄m̄b̄ḡh̄j̄p̄ūz̄āḡp̄w̄+p̄l̄k̄s̄:z̄r̄ūl̄ī
x̄ūc̄ūn̄x̄ī/f̄h̄r̄āl̄ōh̄:p̄b̄=+v̄p̄ūl̄īd̄:ȳp̄m̄s̄īm̄p̄ō+h̄c̄ḡn̄r̄
z̄īr̄b̄z̄x̄ūd̄h̄s̄ūm̄h̄ēx̄+ēz̄ȳūh̄z̄m̄h̄h̄t̄ōr̄d̄ūf̄h̄h̄īc̄ū
f̄l̄b̄m̄â̄r̄p̄ēōf̄d̄c̄īh̄c̄ȳr̄j̄īāḡx̄w̄z̄p̄/n̄īz̄h̄īd̄p̄â̄īn̄
p̄ēz̄x̄d̄l̄īr̄b̄ēn̄r̄p̄āḡȳf̄p̄â̄z̄h̄īh̄c̄ēl̄ūf̄m̄z̄
c̄s̄p̄f̄ȳīūz̄ū/l̄p̄ī=l̄m̄p̄ȳâ̄b̄d̄z̄ȳ/f̄p̄m̄d̄
x̄l̄c̄r̄â̄j̄m̄c̄:z̄+h̄h̄c̄r̄ēȳr̄j̄īp̄īn̄d̄p̄s̄x̄n̄/c̄
w̄īn̄r̄f̄+ḡp̄ē=l̄c̄ȳp̄d̄/l̄z̄d̄p̄/d̄r̄ȳ/m̄z̄.
Δp̄é̄ō+b̄/l̄n̄.

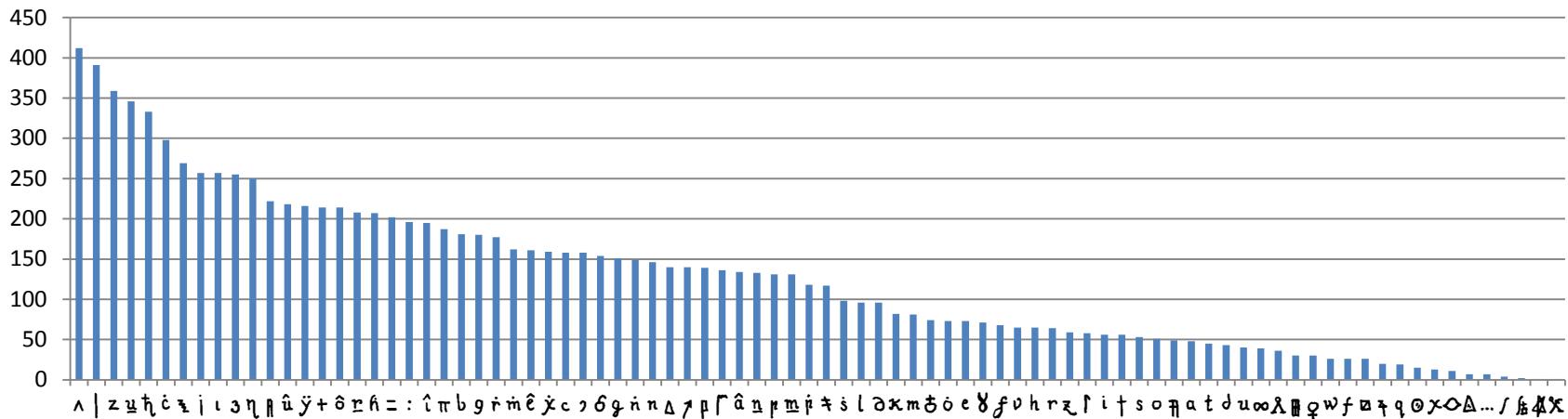
Caproðc̄é̄īp̄v̄ōp̄r̄ā/Δn̄â̄ȳx̄ūl̄c̄p̄b̄ḡ=z̄ḡh̄z̄/

Non-enciphered inscriptions:
Copiales 3 and Philipp 1866

Lines ≈
equal length

Paragraphs and section titles
always begin with
capitalized Roman letters.

Letter Frequencies



digraphs:

՚ Ւ 99

ծ :

Ւ ՞ 49

։ Ա 48

Զ Ր 44

trigraphs:

՚ Ւ ՞ 47

ծ : Ա 23

՞ Հ Ւ 22

Յ Ռ Ւ 18

Ւ Ծ | 17

tendencies:

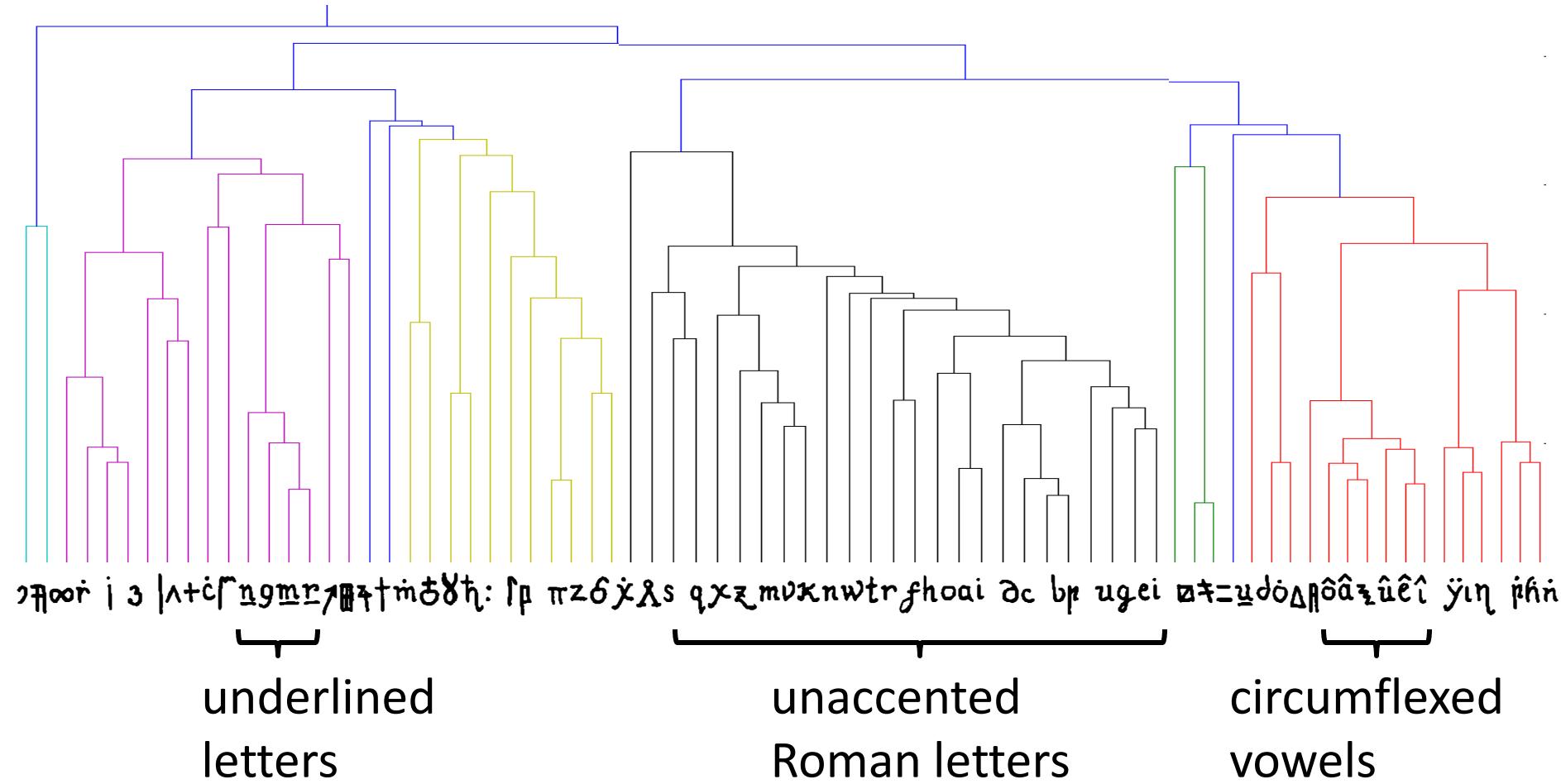
â, ê, î, ô, û followed by Յ and Ի

â, ê, î, ô, û preceded by Զ and Ր

Clustering of Cipher Letters

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

Scipy software



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

**zfnqlknacbfxmk
lbuvcghtrhbkgnkn
fggnkbgbeeb ...**

Decipher against
80 plaintext languages.

Second Decipherment Approach

Homophonic cipher, e.g.:

A = ፩ i l y የ
B = ኃ
C = ዕ n
D = ቅ
E = ጂ f ል ቁ f ካ እ 3
F = p
G = ዕ



etc.

κτημάτων προστατεύεται το ιερό της Αγίας Ειρήνης
που βρίσκεται στην πόλη της Καστοριάς.
Η εκκλησία έχει κατασκευασθεί με πολύ χαροκόπια
τέχνη, αποτελούμενη από δύο νάρθηκες, έναν κεντρικό
θόλο και δύο πλατύτερες βόρεια και νότια περιστοιχίες.
Το εσωτερικό της εκκλησίας είναι διακοσμημένο με
πολύτιμα ορείχαλκα, γαλαζοπούρια, λαζαρίτη
και άλλα πολύτιμα λεπτομερή.
Οι τοιχογραφίες στην εκκλησία είναι θεματικές
και αφορούν στην ζωή της Αγίας Ειρήνης, την ιστορία
της πόλης και την θρησκευτική παράδοση της περιοχής.
Η εκκλησία έχει θεωρηθεί ένα από τα σημαντικότερα
τοποθετήσεις της ιερής Αγίας Ειρήνης στην Ελλάδα,
και έχει αποδοθεί σημαντική ιστορική και πολιτιστική
αξία.

Homophonic Cipher

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

FAIL

But, slight numerical preference for
German

Cipher Characteristics

digraphs:

, ī	99
č :	66
ī ^	49
: ü	48
z R	44

trigraphs:

, ī ^	47
č : ü	23
ī , ī	22
ÿ , ī	18
ī č	17

tendencies:

â, ê, î, ô, û followed by ɔ and ï

â, ê, î, ô, û preceded by z and ñ



should appear
adjacent in German text

Make full digraph table for cipher and for German

Key Observation #1

In Copiale, \mathfrak{C} almost always followed by \mathfrak{H}

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

So guess: $\mathfrak{C} = C, \mathfrak{H} = H$

One Thing Leads to Another

ſt̄ = CH → ſt̄Λ = CHT → Λ = T ?

Each step is guesswork.

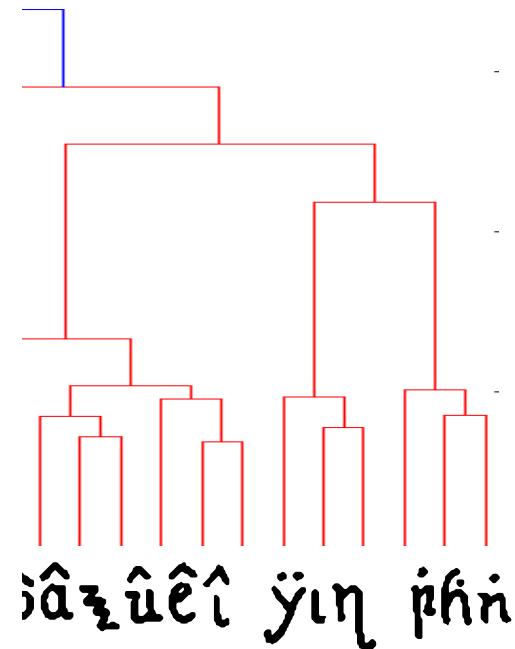
Must be willing to retract.

Weird task, not knowing German.

No longer care what the book says.

Cluster diagram crucial:

ÿ = I → ȳ = I , ȳ = I



Spring Break 2011

Cipher
letters,
in groups

Quite a bit
of fooling
around →

German letters

c aeiou fpy dlmrztbvw hknqsg

German trigraphs

Cipher trigraphs

Grid

*	der	v	[cht]	cht
	W	nd	v	e: u
	e	in	v	h̄: i
	u	n	a	ḡ: ɔ
	c	h	t	n̄: ə
	r	e	h	χ̄: ʌ
	s	c	h	ʌ̄: ɒ
*	e	h	e	i: ə
	e	ch	en	ɛ: ʌ
	d	ie	e	+ y: ɔ
	r	ec		h̄: ɪ
	in	g		ɪ̄: ʊ
	ge	n		ə̄: ə
	e	t		ə̄̄: ə̄
	v	er		ə̄̄̄: ə̄̄
	h	ien		ə̄̄̄̄: ə̄̄̄
	l	ic		ə̄̄̄̄̄: ə̄̄̄̄
	t	ten		ə̄̄̄̄̄̄: ə̄̄̄̄̄

Trigraph Decoding Guesses

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

Kmûr:rzlôf|y, hêi hziln pâzba g z= iplz u kôc lârg k l h pâz i npûl d r n h la = g z w p y ê c A r t ð a + b z q r i x y j i r z u f l z p t i p d r i = f | u l s t p n m | z n g â | k h = l h | l x ô o f : r i l b i f u m y j z v z â j x , r p i z h i l c t ð o g g z û t p q m x e z g h l h p i h p l i z t n f | r ô y m â + h h r z ô z n b s n t : z r k p d h h d c n l g = n z k p z o o n z p z f h i n r p z e y n g = r p g t ð a z n z p k p n j i x y r i g p u l b g l i t b d n f p h h z ô l e z n ô f i n r c f z d n l b n h h m

Actually, those are space bars

Copiale Decipherment

lit:mz||bl
v̄x̄|̄l̄s̄k̄p̄t̄|w̄n

ποιηταιεωνα=γλυπ̄θυρ̄θυ

δηκ̄ι+δημητ̄ηιη̄c̄f̄.

cūl̄ēt̄p̄t̄ḡl̄ā:

δχ̄η̄ēn̄+z̄r̄k̄p̄l̄īz̄:ȳγ̄l̄d̄ōq̄z̄īx̄āj̄ēc̄:ūd̄.
f̄ūr̄f̄īk̄l̄āl̄ēc̄.

μ̄p̄r̄ā+Δ̄ḡȳūx̄z̄d̄ēm̄n̄f̄ūh̄īh̄+f̄.

κ̄m̄ūr̄p̄z̄īōf̄j̄ȳt̄h̄ēīh̄īl̄āl̄p̄āz̄b̄Δ̄ḡz̄īp̄l̄z̄ūp̄q̄c̄l̄ār̄ḡk̄l̄h̄

π̄r̄h̄ēl̄īn̄p̄l̄īd̄r̄f̄l̄āz̄ḡw̄p̄ȳēc̄Δ̄r̄d̄+b̄z̄h̄r̄īx̄ȳīr̄z̄ūf̄λ̄z̄
π̄x̄īd̄ī=ḡl̄n̄l̄s̄k̄p̄l̄īz̄n̄|̄ḡā|̄x̄h̄=l̄h̄|̄l̄x̄d̄w̄r̄:r̄īl̄b̄īl̄ūm̄ȳj̄z̄
z̄ā̄īx̄īp̄l̄īz̄h̄l̄āc̄t̄ōḡz̄ū+r̄t̄p̄p̄x̄ēz̄ḡh̄l̄h̄īh̄l̄īz̄t̄n̄f̄r̄ȳ
m̄ā̄t̄h̄h̄r̄z̄d̄z̄n̄b̄s̄h̄t̄:z̄r̄k̄īp̄l̄h̄d̄c̄ūl̄ḡ=ūz̄k̄p̄z̄ōn̄z̄f̄h̄r̄p̄
z̄ē̄ȳn̄=r̄p̄ḡk̄d̄ēn̄z̄p̄|̄k̄p̄h̄īx̄ȳr̄īḡp̄ūλ̄b̄ḡl̄īt̄b̄d̄n̄f̄h̄ēn̄ē
z̄n̄ō̄īr̄c̄z̄d̄l̄b̄h̄t̄d̄:

n̄īr̄c̄īḡēōp̄h̄ēr̄d̄p̄z̄d̄n̄h̄z̄ā̄k̄Δ̄ḡs̄=d̄m̄ēj̄z̄ūl̄ī

ḡs̄m̄n̄ōl̄d̄ūp̄ēḡūd̄t̄z̄r̄d̄īz̄ḡz̄r̄j̄ūx̄ūp̄n̄ēk̄Θ̄n̄p̄h̄b̄=n̄
λ̄ēr̄m̄ōḡ:ūā=r̄z̄l̄d̄h̄r̄p̄m̄h̄āz̄d̄j̄l̄īr̄n̄ī6̄8̄ȳl̄ūz̄īēḡc̄ūh̄
r̄s̄ēn̄λ̄
c̄p̄=f̄h̄ūb̄m̄d̄c̄:z̄d̄.

hz̄īl̄:ō̄ḡēz̄n̄h̄ā̄z̄Δ̄n̄p̄ā̄z̄Θ̄p̄s̄r̄t̄īz̄t̄m̄ȳb̄ūl̄ī=n̄p̄
z̄p̄s̄=n̄h̄f̄r̄z̄p̄l̄īt̄z̄p̄ūp̄c̄ȳh̄f̄b̄b̄l̄īx̄n̄ḡh̄īȳt̄īb̄c̄s̄=b̄t̄n̄p̄ūd̄
j̄d̄l̄īr̄:ūl̄v̄īōn̄.
l̄p̄ōz̄f̄īh̄ḡz̄ȳp̄r̄l̄īp̄l̄īt̄d̄r̄l̄l̄ūōz̄īr̄.

gesetz buchs

der hoherleuchte Δ e Θ

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 geführet dieser frägt ihn:

erstlich ob er begehre Δ zu werden

zweytens denen verordnungen der Θ sich

unterwerffen und ohne wiederspenstigkeit die lehrzeit
ausstehen wolle.

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

der candidat antwortet ja.

Copiale Decipherment

lit:mzplbl
vixz̄l̄as̄kp̄t̄l̄wn

ποιητή Δημήτριος ο Έρθρος

69. Եթէ չու տօյորակ է ի՞նչ է
շայի թէ տըսդ ցող և ք

Ճշնհեղ+քրիտլնէց: յշլէթօլզւիյալիք: Ա
füröffjeklassecz.

mɔʃʃɪrā+Δgjūxəzəʒmən̩tʃɪf.

κατηγορίας της οποίας είναι η πατέρας Δασιάρχης πρόεδρος της Επιτροπής της Αρχής για την ανάπτυξη της Κύπρου. Ο πρόεδρος της Επιτροπής είναι ο Αρχηγός της Επιτροπής για την ανάπτυξη της Κύπρου.

የኢትዮጵያውያንኩንቃቄዎንግሥት የሚመለከት ማስተካከል

hzηλ:ôg|ezηâz **Α**πά3Θρστι3τμάýδυli=ηπ
zρσηήγτρηπτή|ριցπέýσήροβηγκηή|ýτιδεσ=б+н|ρу
жô|р:ùл|юон.
Лпôзғиңзýпіалéшләтділәоеғнір.

First lawbook of the e

Secret part.

First section

Secret teachings for apprentices.

First title.

Initiation rite.

If the safety of the **A** is guaranteed, and the **A** is opened by the chief **A**, 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 **A**, who asks him:

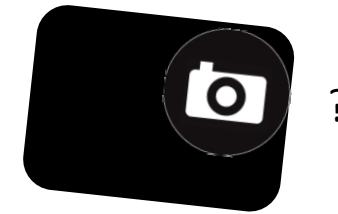
First, if he desires to become 

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

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

The candidate answers yes.

Historical Archives



French plaintext

a Tönningen le 20 de Mars l'an 1713.

Pierre, que j'ay le honneur de vous écrire plus au longue hier par le canal de relay qui m'a porté la Vôtre du 21 de Fevrier; celle du 20 n'étant pas venue, je vous adresse ces mots en une doublette. Par un avis affidé de Husum j'ay été averti, que le Roi avoit pris part et fier la fidélité de nouveaux au Roy de Baviere, & qu'il risqueroit le sacrifice de ses combattants pour abîmer Tönningen & la reduire dans l'état que nous avons mis Altona; disant qu'il avoit assez du monde pour faire teste aux Turques, & que ceux qu'il laissait ici seraient plus que sacrifiés que de mordre de son daffin. Toutes les préparatifs sont faits pour une charge d'une bombardement formelle. Ce n'est pas que je craigne, mais bien la famine. Si cher Comte Vous ne faites en sorte que la France & l'Angleterre nous assiste par mere je craignez que le monde se fonderait en peu. Je fais auj, si plaisir à Dieu, mon devoir en soldat tant que sangue dure. Bien sait que cette situation n'a pas été éviter & que sans la résource de Tönningen les ennemis nous avoient déjà par leur supériorité à leur désertion, car l'infanterie diminuoit par maladie de jour en jour, la cavallerie force de fatigues meconante, faute de fourrage & vivres abattue guère d'avoir de morts, n'avoit pu produire qu'une triste fin! travailler pour nous pour l'amour de Dieu.

46. 131. 42. 79. 175. 79. 93. 128. 46. 63. 130. 409. 32. 26. 302. 41. 66. 63. 81. 70. 138. 88.
53. 146. 86. 34. 363. 81. 1051. 176. 376. 84. 48. 86. 84. 131. 52. 302. 764. 58. 60. 80. 154. 96.
132. 707. 132. 99. 59. 93. 164. 59. 63. 151. 234. 373. 967. 164. 32. 81. 148. 266. 1008. 86.
580. 31. 45. 79. 999. 34. 163. 61. 149. 997. 34. 82. 63. 41. 75. 764. 373. 82. 125. 39. 79. 151. 111. 120.
706. 43. 104. 144. 81. 302. 118. 138. 87. 764. 373. 82. 125. 39. 79. 151. 111. 120.

Ciphertext

| 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 |
|----|----|---|---|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|---|---|
| 31 | 34 | 3 | 2 | 39 | 40 | 41 | 42 | 43 | 44 | 45 | 46 | 47 | 48 | 49 | 50 | 51 | 53 | 58 | 60 | 61 | 62 | 63 | 64 | | |
| 32 | 68 | | | 75 | 76 | 77 | 78 | 79 | 80 | 81 | 82 | 52 | 84 | 59 | | | | | | | | | | | |
| | | | | | | | | | | | | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | | | | | | | | | | | |

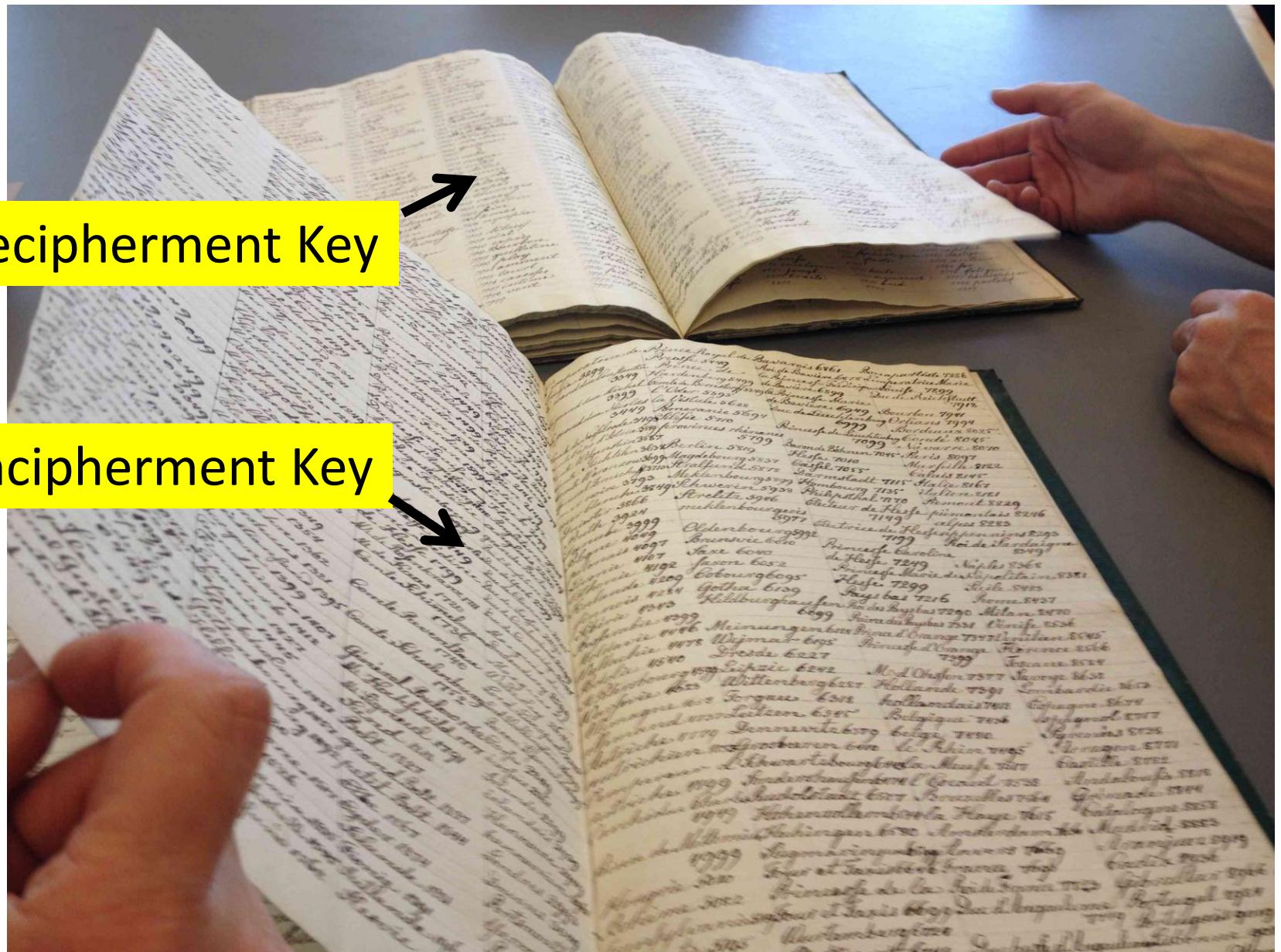
Stockholm Aug 1934
M. G. Hallé

Solution (1934)

Word Substitution Encipherment Key

| | | | | | | | |
|--------------|-------------------|------------------|--------------------|-------------------|-------------------|-------------------|------------|
| ment 3495 | assemblée 3960 | alive 2914 | autre 0574 | autres 0569 | nous avions 1523 | tache 0177 | barbier 2 |
| ristice 5327 | asfert 3977 | ativement 2440 | aucunement 1483 | autrui 0790 | vous aviez 1563 | badaud 0404 | barbouil |
| noire 5559 | asfes 4017 | atives 2968 | aucunes 1529 | aux 0025 | avaient 2478 | badi 0477 | barbe 23 |
| nuire 5382 | asfement 4041 | atlas 2495 | aucuns 1568 | auxquels 0045 | je eus 2215 | bardinage 0595 | barde 23 |
| mate 5019 | asfertion 4061 | atmosphère 2521 | audace 1612 | auquelles 0090 | je eus 2108 | bafton 0221 | baril 23 |
| nt 5032 | asfere 4076 | atoise 2530 | audacieux 1626 | auxiliaire 0122 | eut 1824 | bagage 0259 | barque |
| ubuf 5056 | asfesfuer 4105 | atoires 2556 | au deça 1669 | aval 0113 | nous eumes 1365 | bagatelle 0317 | bar 239 |
| ack 5077 | asfex 4142 | atome 2534 | au de là 1691 | avare 0193 | vous eutes 1108 | bague 0258 | bar 292 |
| ang 5109 | asfidu 4157 | à tort 2803 | au devant 1765 | avance 0427 | eurent 2421 | bagquette 0374 | barrea |
| nt 5152 | asfig 4180 | atrobilaire 2825 | audience 1792 | avarie 0443 | j'eus 1606 | baign 0209 | barrie |
| rog 5171 | asfith 3443 | atrice 2860 | auditeur 1703 | avant 0164 | tu auras 1565 | baïl 0243 | barrie |
| station 5192 | asfiez 3445 | atrices 2878 | auditoire 1729 | augment 1024 | avant de 0457 | aura 2225 | baill 0864 |
| l 5214 | asfign 3464 | atrice 2907 | augment 1024 | augur 1033 | avant que nous | aurons 1010 | bain 0290 |
| es 5243 | asfimil 3486 | atrocite 2932 | auguste 1123 | avantage 0571 | vous aurez 1464 | bais 0960 | bambo |
| r 5267 | asfimul 3518 | attack 2962 | aujourd'hui 1081 | avantageux 0542 | auront 2074 | baisf 0972 | bastar |
| re 5283 | asfions 3541 | attaque 2984 | aulique 1236 | avant posé 0554 | j'aurais 2269 | bal 0912 | baston |
| regards 5300 | asfies 3555 | attaque 2604 | aulo 1267 | avant garde | tu aurais 2167 | balai 0644 | baisque |
| 5329 | asfies 3577 | attein 2626 | aumone 1312 | avant hier 003 | aurait 1245 | balance 0657 | baistes |
| é 5365 | asfist 3616 | atteinte 2657 | aumonier 1385 | avant midi 004 | nous aurions 1863 | balance 0740 | bastin |
| 5384 | asfoci 3641 | attel 2681 | aune obre | avant propos 0271 | vous auriez 2007 | balbuti 0762 | bastor |
| rd 4616 | asfomm 3654 | attelage 2709 | auparavant 03 | aware 0296 | auraient 2196 | balaine 1031 | bastta |
| 4637 | asfot 3679 | attenant 2736 | auprès 0652 | avarice 0305 | j'aie 1920 | baliste 1050 | basti |
| el 4669 | asfoup 3715 | attend 2762 | auquel 0693 | avec 0328 | tu aies 1664 | balverres 1133 | baston |
| 93 | asfouir 3744 | attendu 2776 | auvéole 0730 | avenant 0385 | ait 1534 | balle 1209 | bat 3 |
| 706 | asfujett 3757 | attendu que 2209 | auréole 0730 | avenir 0815 | nous ayions 1805 | balourd 1295 | batte |
| 4739 | assur 3786 | attendr 2242 | aurore 0742 | aventure 0877 | vous ayiez 1753 | balsamique 1302 | bat |
| 4759 | assurance 3002 | attent 2264 | auroréboriale 0765 | avenue 0922 | ayent 1491 | balustr 1364 | bat |
| 4782 | assurement 307 | attentat 2296 | auspices 0786 | avér 0940 | j'euge 2452 | balustrade 1643 | batte |
| el 4214 | asthmétique 3057 | attenu 2327 | ausfi 0805 | aversion 0996 | tu euges 1072 | bamboche 1652 | bat |
| rie 4238 | asthme 3078 | atter 2342 | ausfitôt 0831 | avert 0601 | éut 1933 | bane 1741 | bat |
| er 4270 | astre 3119 | attast 2368 | ausfitôt que 0861 | aveu 0625 | nous eusions 206 | banal 1771 | bat |
| 4294 | astring 3134 | attied 2393 | ausstère 0885 | aveugl 0671 | vous eufiez 1954 | band 1414 | bat |
| 4306 | astrolabe 3155 | attier 1817 | ausstérité 0916 | aveugle 0687 | eufsent 1980 | bandeau 1502 | bat |
| | astrologie 3176 | attrail 1841 | austral 0933 | avide 0713 | ayant 1282 | bandoulière 1581 | bat |
| | astronomie 3209 | attouch 1866 | autant 0951 | avidité 0725 | ayer 1594 | banlieue 2042 | bat |
| nt 4382 | astronomie 3223 | attraction 1880 | autant que 0939 | avil 0752 | ayons 1621 | bann 2054 | bat |
| | astronomique 3235 | battrait 1904 | autel 0713 | aviron 0782 | dvoisin 1786 | banni 2113 | bat |
| 442 | astuce 3279 | attrayant 1926 | auteur 0228 | avis 1011 | avort 2101 | bannissement 1804 | bat |
| 476 | astucieux 3308 | attribu 1969 | authenticité 0265 | avitaill 1065 | avou 2278 | banquieroute 1809 | bat |
| 4491 | at 3326 | attribut 1989 | authentique 0285 | avocat 1148 | avril 1851 | banque 1875 | bat |
| 07 | atelier 3361 | attrist 2004 | autocrate 0322 | avoine 1343 | aae 1678 | banquier 1903 | bat |

Word Substitution Keys



Word Substitution Keys

Numbers/Words Both
in Order!

| Lat | 901. | Conference | 952. Commiratio |
|--------|------|----------------|--------------------|
| do | | | |
| Dora | 9 | | |
| Dant | 9 | | |
| rius | 9 | | |
| Sarv | 931. | Confusione | 957. Constantingel |
| tion | 932. | Congratulation | 958. constituer |
| teria | 933. | Congratulera | 959. constitution |
| dilat | 934. | Congress | 960. conterance |
| ration | 935. | Conjunction | 961. Content |
| ceria | 936. | Conjangera | 962. Contentement |
| gnie | 937. | Conianctim | 963. Contentera |
| retra | 938. | Comivera | 964. Contester |
| lation | 939. | Coniventz | 965. Continent |
| lera | 940. | Consens | 966. Continuation |
| nt | 941. | Consentora | 967. Continuera |
| bea | 942. | Consequenter | 968. Continue |
| mont | 943. | Consequenter | 969. Contract |
| ura | 944. | Conservation | 970. Contrahera |
| lera | 945. | Conververa | 971. Contribution |
| scio | 946. | Consideration | 972. Contribuera |
| on | 947. | Considerable | 973. Controvorsia |
| era | 948. | Considerera | 974. Convent |
| te | 949. | Consilia | 975. Conversation |
| era | 950. | consilium | 976. Conversera |
| ora | 951. | Consiliarius | 977. Convoy |

| | Art | | |
|-----|----------------|------|-----------------|
| 175 | Sverige | 950 | Sincomalee |
| 901 | Stockholm | 521 | Point de Galle |
| 328 | Christiania | 789 | Colombo |
| 109 | Carlekerona | 987 | Ceylon |
| 393 | Lissabon | 997 | Bombay |
| 569 | Madeira | 195 | Suezkanalen |
| 596 | Equator | 591 | Alden |
| 747 | Rio Janeiro | 554 | Port Said |
| 783 | Montevideo | 973 | Alexandria |
| 558 | Buenos Ayres | 718 | Kairo |
| 323 | Magellansund | 506 | Malta |
| 768 | Eddelandet | 383 | Tunis |
| 706 | Salparaiso | 757 | Gibraltar |
| 174 | Callao | 347 | Norge |
| 936 | Paskow | 986 | Danmark |
| 762 | Marquesasøerne | 737 | England |
| 535 | Tahiti | 1011 | at. h. b. i. o. |
| 301 | Honolulu | | |
| 187 | + Caroliner | | |
| 314 | Ladronerne | 559 | Italien |
| 365 | Japan | 749 | United States |
| 504 | Yokohama | 168 | Europa |
| 561 | Nangasaki | 153 | Ostindien |
| 983 | Inland sea | 580 | Holland |
| 729 | China | 177 | Turkiest |
| 592 | Shanghai | 994 | Egypten |
| 649 | 10 | 158 | o. B. |

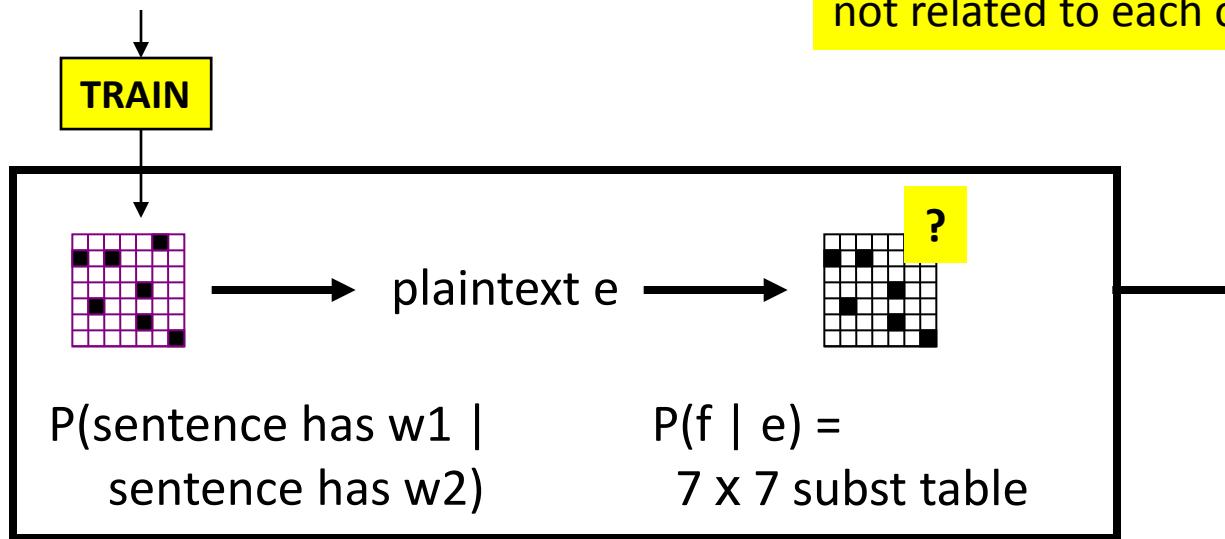
Neither in Order!

Word Substitution

- Interesting for NLP
- Language translation can be viewed as word substitution (and transposition)
- Certainly, that is how IBM models 1-5 view it

Word Substitution (Small-scale)

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

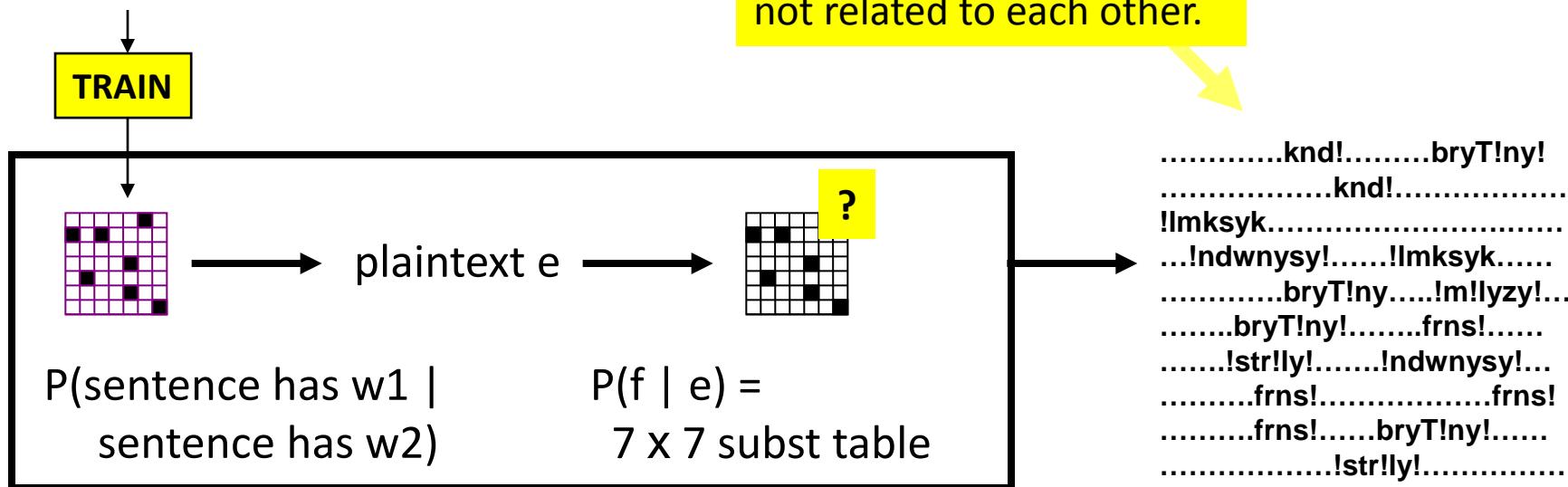


.....knd!.....bryT!ny!
.....knd!.....
!lmksyk.....
...!ndwnysy!.....!lmksyk.....
.....bryT!ny.....!m!lyzy!...
.....bryT!ny!.....frns!.....
....!str!ly!.....!ndwnysy!...
.....frns!.....frns!.....
.....frns!.....bryT!ny!.....
....!str!ly!.....

Word Substitution (Small-scale)

.....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.



| | | | | | | | |
|-----------|---|-----------|--------|-----------|--------|----------|--------|
| 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 (Giga-scale)

- Suppose I replace each English word on your hard drive with some integer.
- Can you recover your texts?
- In principle, apply the same techniques we used for letter substitution.
 - English word-bigram LM drives decipherment
 - But for EM, initially-uniform substitution table is too big!
 - $100,000 \times 100,000$

Word Substitution (Giga-scale)

- Gibbs sampling fixes memory problem

Cipher: 24234 1899 39902 5716 29948 ...

Plain: the man is car are ...

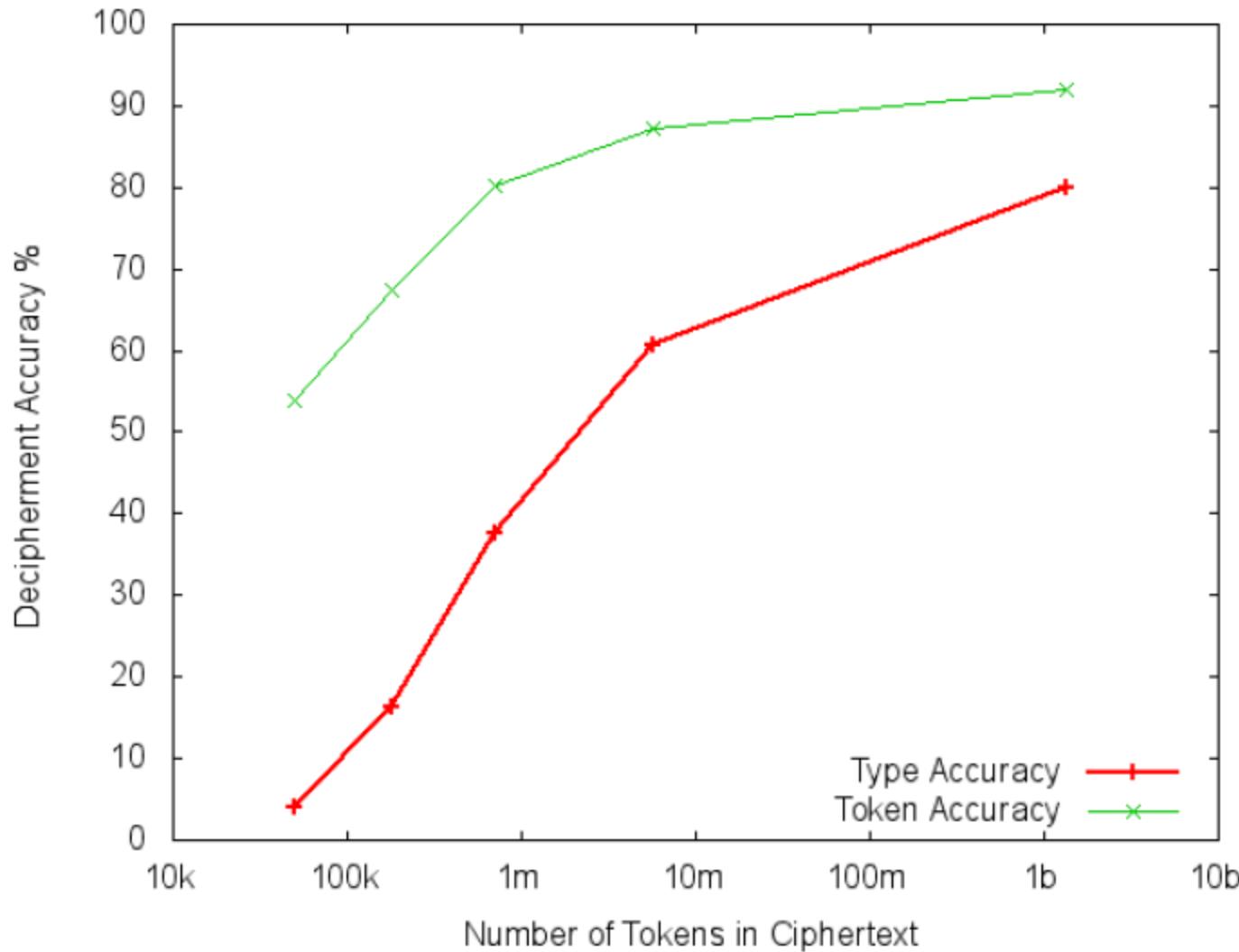
Resample:

a
an
apple
...
man
zoo

Still need to sample 100,000 alternatives at each cipher token, for each epoch.

- Slice sampling (Dou & Knight 12) fixes speed problem

Word Substitution (Giga-scale)



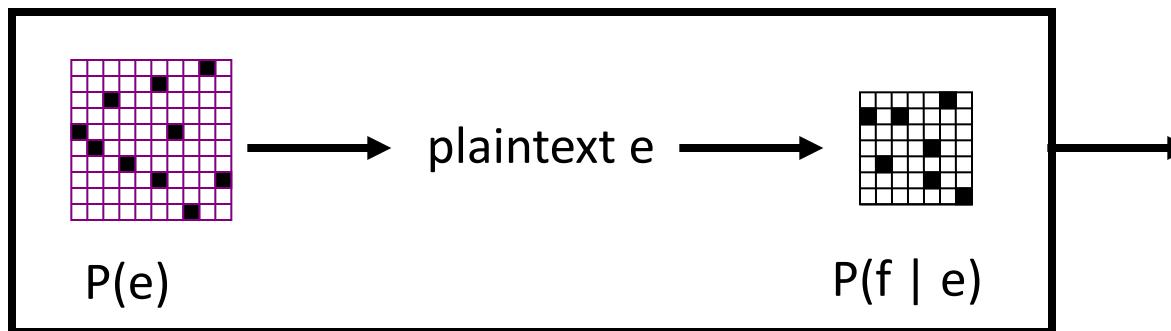
(Dou & Knight 2011)

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 يتمنى 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 يتمنى 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".

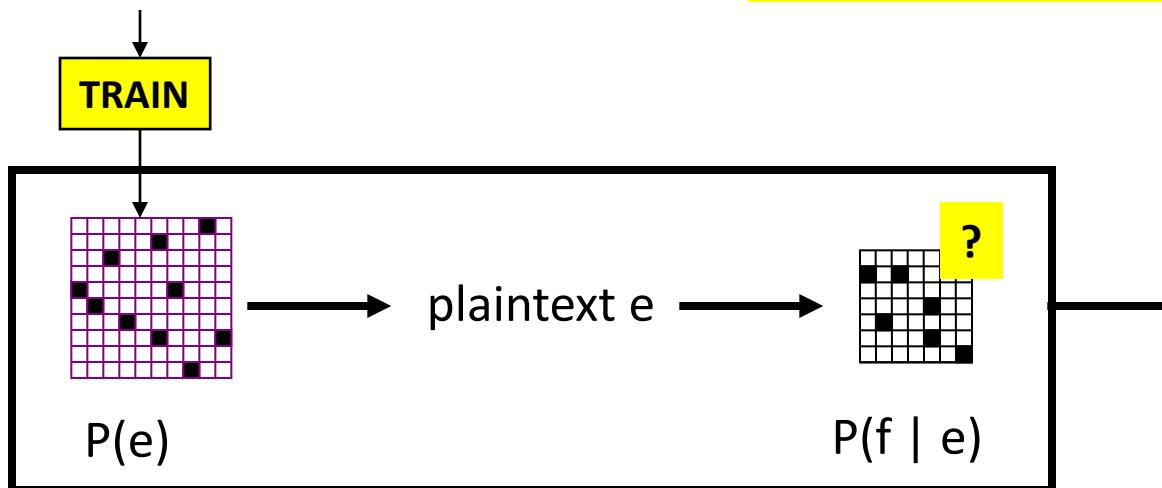
من جانبه شجب رئيس الحكومة الفلسطينية أحمد قريع الطلب الأحادي الجانبي لانسحاب الإسرائيلى من غزة، وأكد أن إسرائيل تريد مغادرة هذه الأرض لتعزيز سيطرتها على الضفة الغربية.

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 الجانب".

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 شارون التي قال فيها إنه يتمنى 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 يتمنى 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 الطلب الأحادي الذي ينادي لانسحاب الإسرائيلى من غزة، وأكد أن إسرائيل تريد مغادرة هذه الأرض 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 أحدى الجان

!!@!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

Time Expressions

@!m 1990
w!lth!ny&
fy !lywm
mn !lsh=hr !lj!ry
!lqrn
!y!m
@!m!aN
!!s!@&
17 shb!T 1994
th!lth snw!t
dqyq&
=hdh=h !lsn&
ywmyn
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& @sh!&
th!lth& ly!m
qbl !sbw@yn
fy !lywm !lt!ly
sh@b!n
tmwz
3 dhw !!Hj& 1414
fy shb!T !lm!Dy
qbl ywmyn

Time Expressions

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

| | | |
|---------------------|----------------------|----------------------|
| 9 Hzyr!n 1942 | 27 tmwz 1993 | 21 Hzyr!n 1967 |
| 8 tshrym !!!wl 1990 | 26 tmwz 1953 | 20 !'y!r 1990 |
| 7 k!nwn !!!wl 1993 | 26 shb!T 1993 | 20 tshrym !'wl 1983 |
| 6 !'y!r 1993 | 26 k!nwn !!!wl 1994 | 20 tshrym !!!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 tshrym !!!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 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 |
| 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

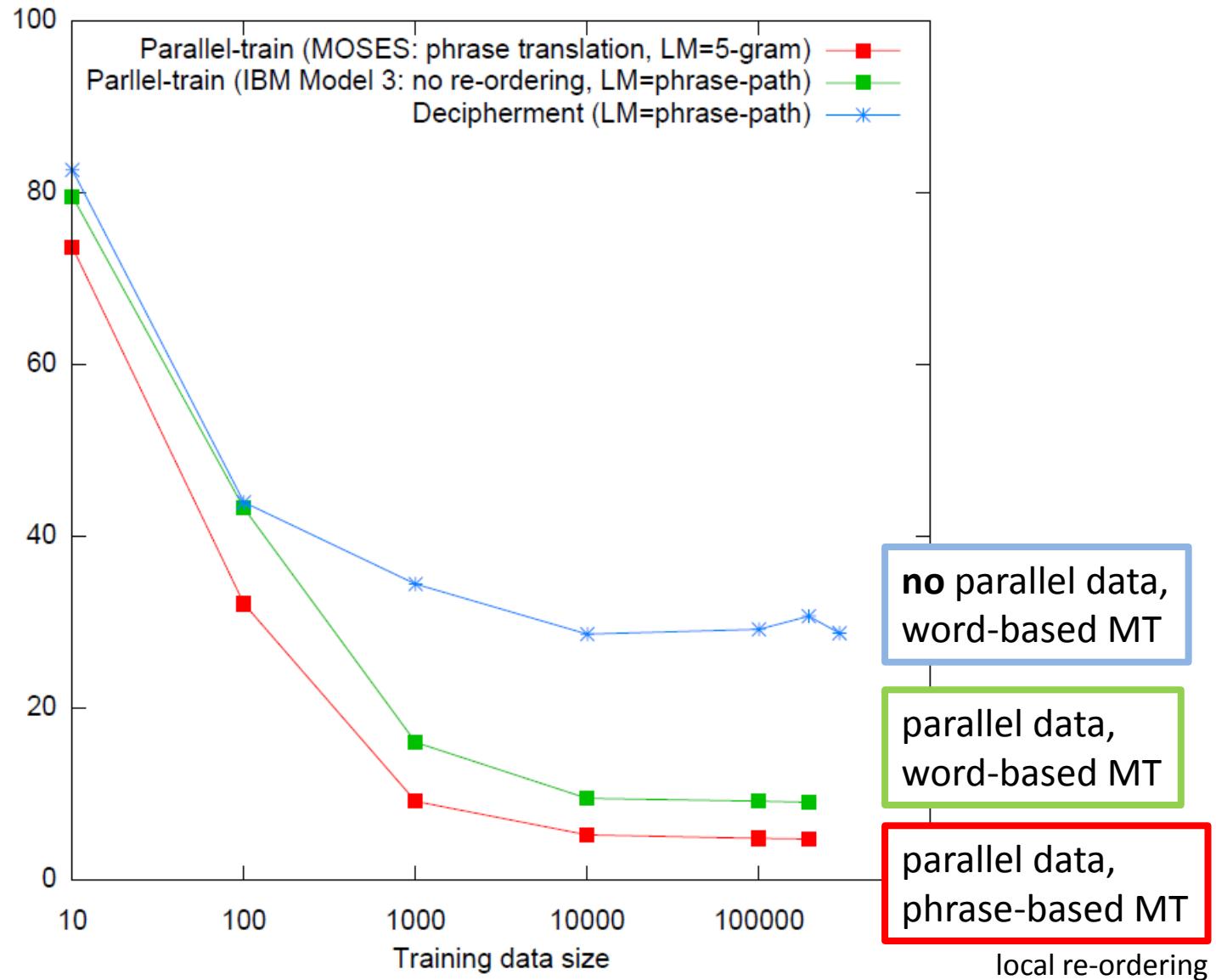
Hzyr!n

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

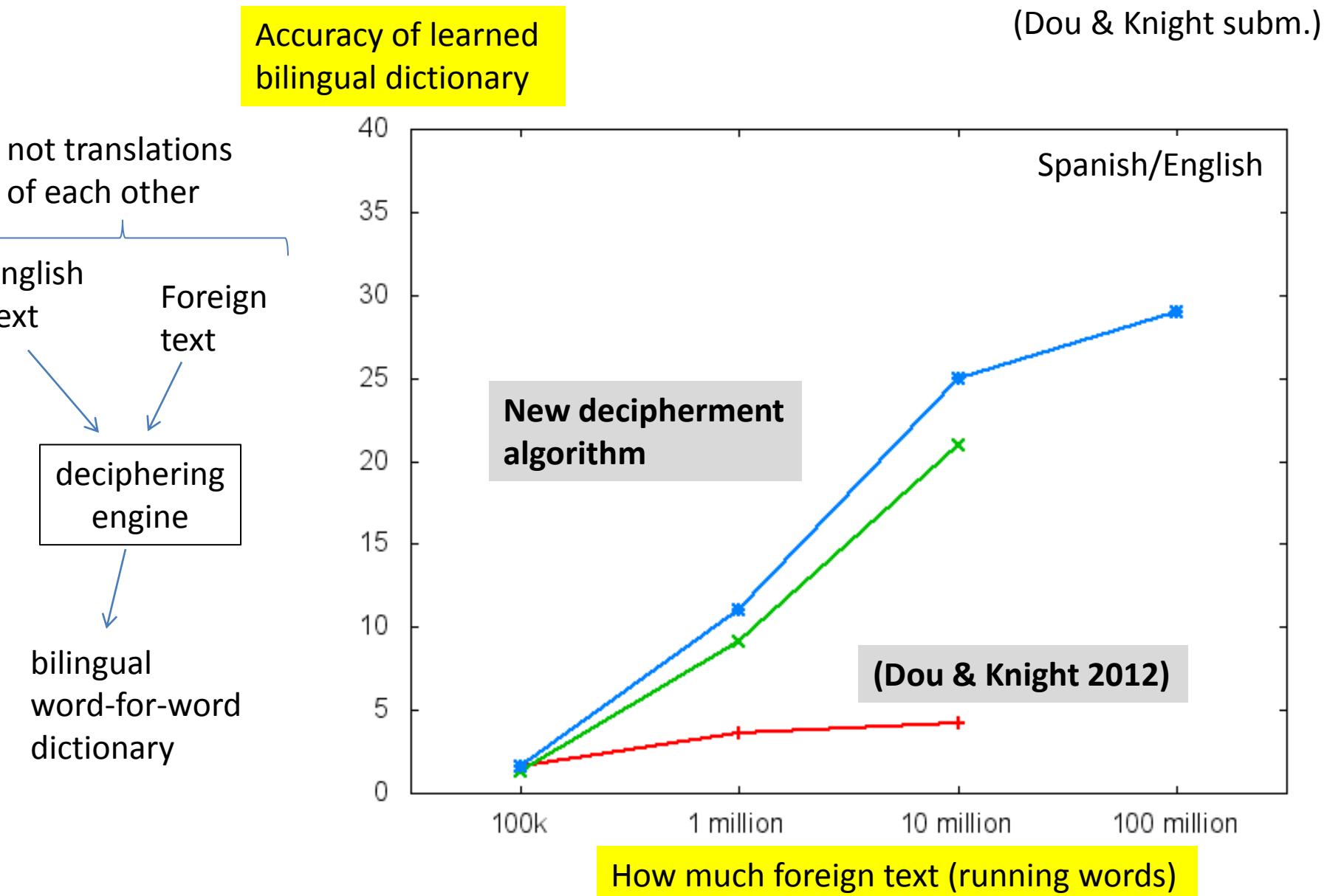
Deciphering Spanish Time Expressions

MT quality
on test set

(Edit distance,
lower is better)



Deciphering Foreign Language at Giga-Scale



Practical Value

- Scenarios where in-domain parallel data is scarce.
- Decipher large monolingual in-domain corpora to improve systems trained on small amounts of parallel text

Africa



Zero languages spoken

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

Unsolved ciphers

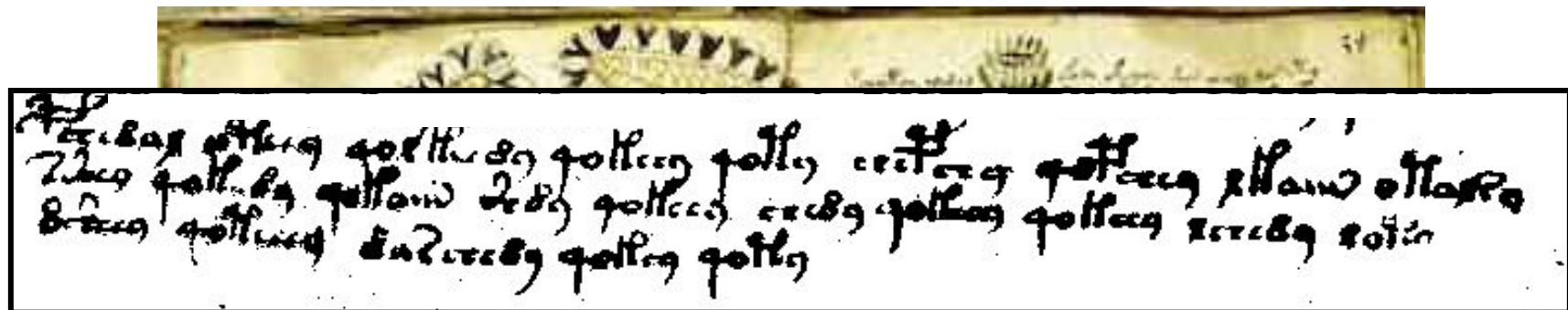
Voynich Manuscript (VMS)



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



Voynich Manuscript (VMS)



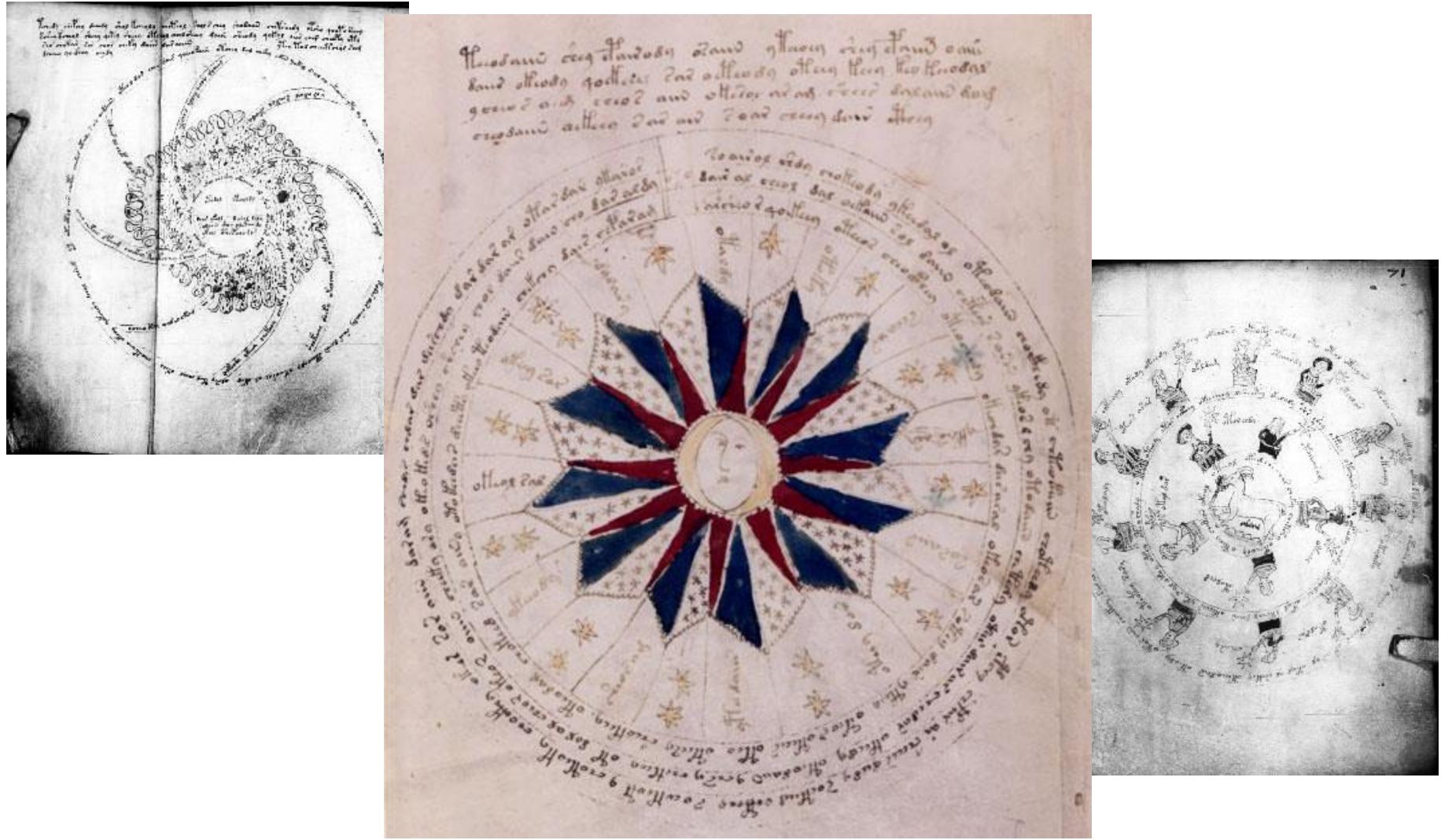
¶CC8AE 0¶CC9 4OE FCC89 4OFCC9 4OP9 SCBS9 4OBSC9 EFAM OPAE29
2ZC9 4OFC89 4OFAM Z89 4OFCC9 SC89 4OFCC9 4OFCC9 ESC89 EOP9
8ZC9 4OPCCC9 8ARSC89 4OFC9 4OP9

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

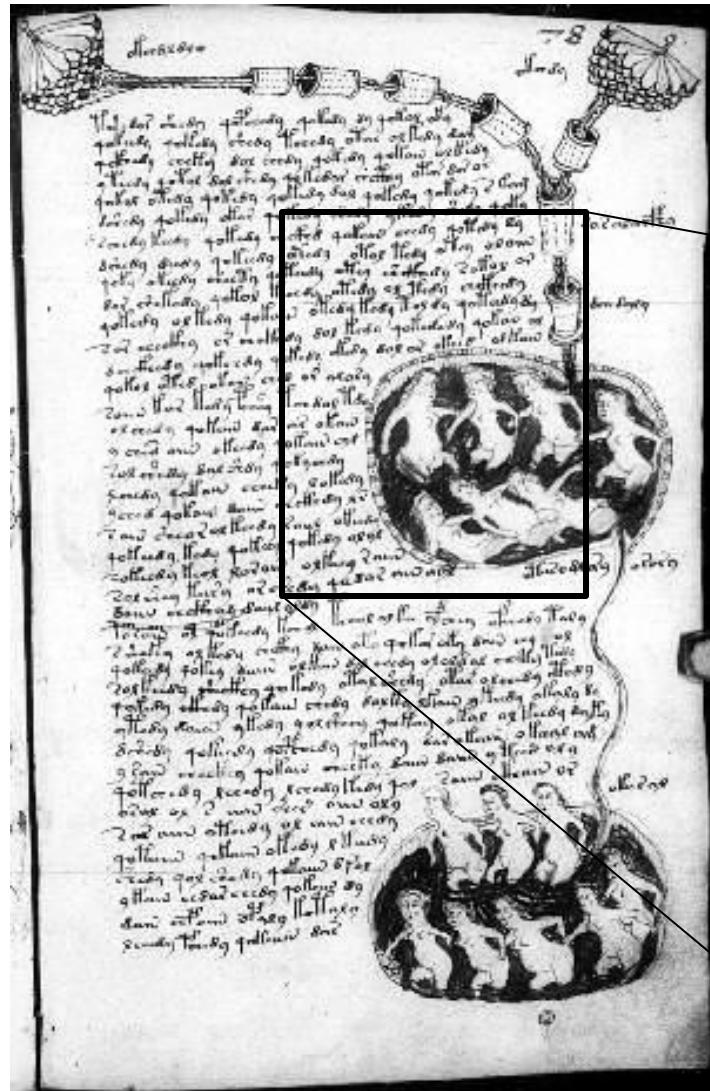
“Herbal” section



“Astrological” section



“Biological” section



Small nudes in baths

Interconnecting tubes of liquids



“Pharmacological” section



History of Voynich Manuscript (VMS)

1576-1612 Rudolf II purchases VMS

1608-1622 J. de Tepenecz signs VMS
in Bohemian court

1630s George Baresch owns VMS
sends letter to Kircher

1639 GB writes Kircher again

16xx Marci inherits VMS from GB

1665 Marci sends VMS to Kircher
with letter

1665-80 Kircher owns VMS

1680 Kircher dies

- 1864 Ethel Boole born in England
- 1865 WV born in Lithuania
- 1885 WV imprisoned, Polish nationalist
- 1890 WV & EB meet, marry in 1902
- 1898 WV publishes first book list
- 1912 WV acquires VMS in “ancient castle”
- 1914 WV moves to USA, opens bookshop
- 1919 WV sends photostatic copies of VMS
- 1919 Copying reveals de Tepenecz signature
- 1919 WV writes to Bohemian State Archvs
- 1921 WV presents VMS + inserted Marci letter
mentioning Francis Bacon, asks \$160k**
- 1921 Newbold & WV announce decipherment**
- 1930 WV dies. VMS placed in vault, \$100k
- 1931 VMS appraised at \$19,400
- 1960 Ethel dies, VMS to secretary Ann Nill
“Castle” revealed as Villa Mondragone
- 1961 NY dealer Hans Kraus buys for \$24,500
- 1969 Kraus donates VMS to Yale
- 1972 Brumbaugh finds WV letters in BSA
- 200x Zandbergen finds 1639 Baresch letter
in newly online Kircher archive



WILLIAM ROMAINE NEWBOLD
1869-1946
The Portrait by Joseph Salsky

Newbold Decipherment

Marci letter → Bacon → Cabala → “letter doubling” cipher

| A | B | C | D | E | F | G | H | I | L | M | N | O | P | Q | R | S | T | U | V | X | Z |
|---|---|---|---|---|---|---|-----|---|---|-----|-----|---|---|---|---|---|---|---|---|---|---|
| A | V | Z | B | F | G | L | M | N | N | O | ... | | | | | | | | | | |
| B | C | F | T | U | V | X | ... | | | | | | | | | | | | | | |
| C | F | B | A | Q | F | C | D | Z | Z | ... | | | | | | | | | | | |
| D | | | | | | | | | | | | | | | | | | | | | |
| E | | | | | | | | | | | | | | | | | | | | | |
| F | | | | | | | | | | | | | | | | | | | | | |
| G | | | | | | | | | | | | | | | | | | | | | |
| H | | | | | | | | | | | | | | | | | | | | | |
| I | | | | | | | | | | | | | | | | | | | | | |
| L | | | | | | | | | | | | | | | | | | | | | |
| M | | | | | | | | | | | | | | | | | | | | | |
| N | | | | | | | | | | | | | | | | | | | | | |
| O | | | | | | | | | | | | | | | | | | | | | |
| P | | | | | | | | | | | | | | | | | | | | | |
| Q | | | | | | | | | | | | | | | | | | | | | |
| R | | | | | | | | | | | | | | | | | | | | | |
| S | | | | | | | | | | | | | | | | | | | | | |
| T | | | | | | | | | | | | | | | | | | | | | |
| U | | | | | | | | | | | | | | | | | | | | | |
| V | | | | | | | | | | | | | | | | | | | | | |
| X | | | | | | | | | | | | | | | | | | | | | |
| Z | | | | | | | | | | | | | | | | | | | | | |

22x22 table

Encoding:

A → CC, OM, ...

B → ...

...

N → HA, MI, DO, NU ...

...

Z → ...

Decoding:

...

DO → N

...

Encoder has freedom to devise
a “cover text” to hide real message.

Example:

a n n ... → DO MI NU ... → DOMINU ...

Newbold System

- Too hard to assemble good “cover” text!
- **So, make cipher letter-pairs overlap:**
a n n ... → AD DB BR ... → ADBR ...
- **Then, employ anagramming:**
a n n ... → OM DO MI ... → DO OM MI ... → DOMI ...
- Now can construct a plausible looking “cover” text in Latin for our secret message (also in Latin)
- An ingenious system, to be sure!

Newbold Decipherment

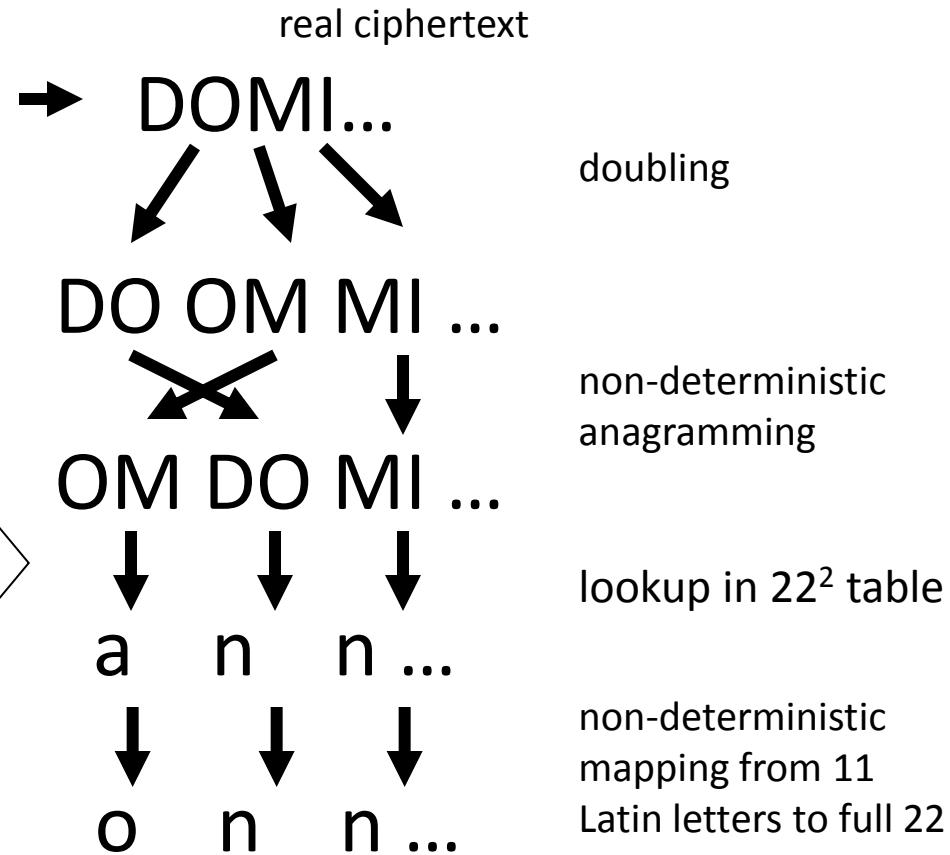
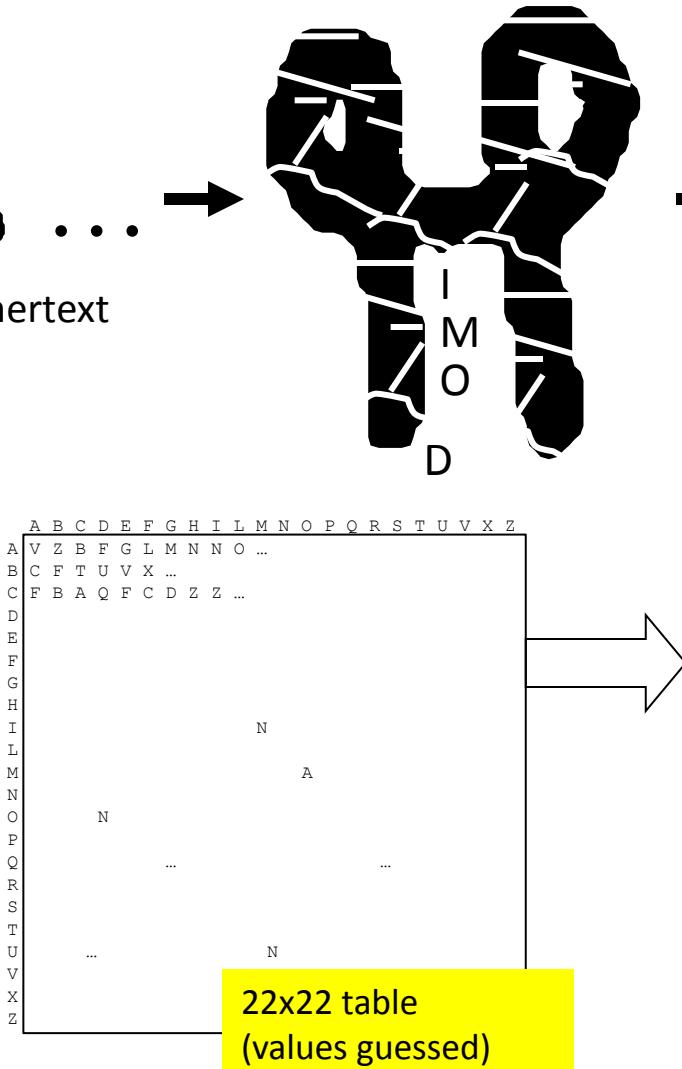
Hmm, by the method, both plaintext **and ciphertext** should be in Latin letters...

But the VMS doesn't have Latin letters...



Let's Decipher with Newbold !

Hcc8g ...
apparent ciphertext





Newbold's Results

1300 real ciphertext “letters” in first 3 lines

Decipherment of those first lines:
“I, Roger Bacon, have written this...”
(in Latin)

Anagramming sets of 55 letters is sometimes required.

Slow but steady progress... Andromeda galaxy, ovaries ... so
... Roger Bacon must have had a microscope & telescope,
hundreds of years before they were invented ... !

VMS Transcription

Figuras etiam quodlibet quod possit possit esse, ceteras quod possit esse sicut etiam figurae.

ଫେରସାଧ ୦୯୮୮୯ ୪୦୯ ମୁଁ୮୯ ୪୦୯୮୯ ୪୦୯୯ ଟେଲ୍ଫୋନ୍ ୪୦୯୮୯ ଡିବାଇୟ ୦୯୮୨୨୯
୨୮୮୯ ୪୦୯୮୯ ୪୦୯୮୯ ୨୮୯ ୪୦୯୮୯ ୪୦୯୮୯ ୪୦୯୮୯ ୪୦୯୮୯ ୧୮୮୯ ୨୦୯୯
୮୮୮୯ ୪୦୯୮୮୯ ୪୦୯୮୯ ୪୦୯୯

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

last paragraph, f103r

Alphabet: Currier/D'Imperio Transcription

ፋ ፃ ፄ
C S Z

ፅ ፆ ፇ ፈ
P F B V

ፉ ፊ ፋ ፌ
Q X W Y

ፋ ፈ ፩ ፪ ፫ ፬ ፭
J A E R O I D

፧ ፨ ፩ ፪ ፫ ፬
6 7 8 9 4 2

፪ ፫ ፬
G H 1

፩ ፪ ፪
T U O

፪ ፪ ፪
N M 3

፪ ፪ ፪
K L 5

VMS Letters

| count | letter |
|-------|--------|
| 25468 | O օ |
| 20227 | C ց |
| 17655 | 9 յ |
| 14281 | A ա |
| 12973 | 8 յ |
| 11008 | S ՛ |
| 10471 | E Շ |
| 10026 | F Ւ |
| 6716 | R Ր |
| 5994 | P Փ |
| 5423 | 4 Ք |
| 4501 | Z Հ |
| 4076 | M Վ |

| count | letter |
|-------|--------|
| 2886 | 2 ՞ |
| 1752 | N Ն |
| 1413 | B Բ |
| 1046 | J յ |
| 950 | Q Զ |
| 908 | X Ճ |
| 591 | T Ւ |
| 524 | * |
| 431 | V Յ |
| 316 | I ՚ |
| 217 | W Ճ |
| 157 | D Յ |
| 156 | 3 Վ |

| count | letter |
|-------|--------|
| 148 | U Ա |
| 96 | 6 Ֆ |
| 74 | Y Տ |
| 52 | K Ո |
| 31 | G Ե |
| 17 | L Լ |
| 14 | H Ի |
| 2 | 1 Վ |
| 1 | 5 Ր |
| 1 | 0 Վ |

Total
63k character tokens

VMS Words

count word

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

count word

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

count word

| | |
|-----|-------|
| 140 | OPCC9 |
| 138 | OFAE |
| 130 | ZO |
| 129 | OFAR |
| 119 | ESC89 |
| 118 | OFC89 |

+ many more!

Total:

8116 distinct words

VMS Word Bigrams

- Very few repeated bigrams: **Extremely troubling!**
Nothing like “of the” in English.
- 115 (out of 8116) distinct words appear doubled
... 401cc89 401cc89 ...
- 8 distinct words appear tripled
... 401cc89 401cc89 401cc89 ...
... ccoz ccoz ccoz ...
... cccoz cccoz cccoz ...
... olfaniv olfaniv olfaniv ...
... oz oz oz ...
... gHfaniv gHfaniv gHfaniv ...
... 8aniv 8aniv 8aniv ...
... 401cc89 401cc89 401cc89 ...

Substitution Cipher?

- Nope.
- Tried 80+ languages.
- For example, if we decipher assuming Latin plaintext:

quiss squm is onum pom
quuss hates s qum hatis ...



- Tried 80+ languages written without vowels.

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

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:

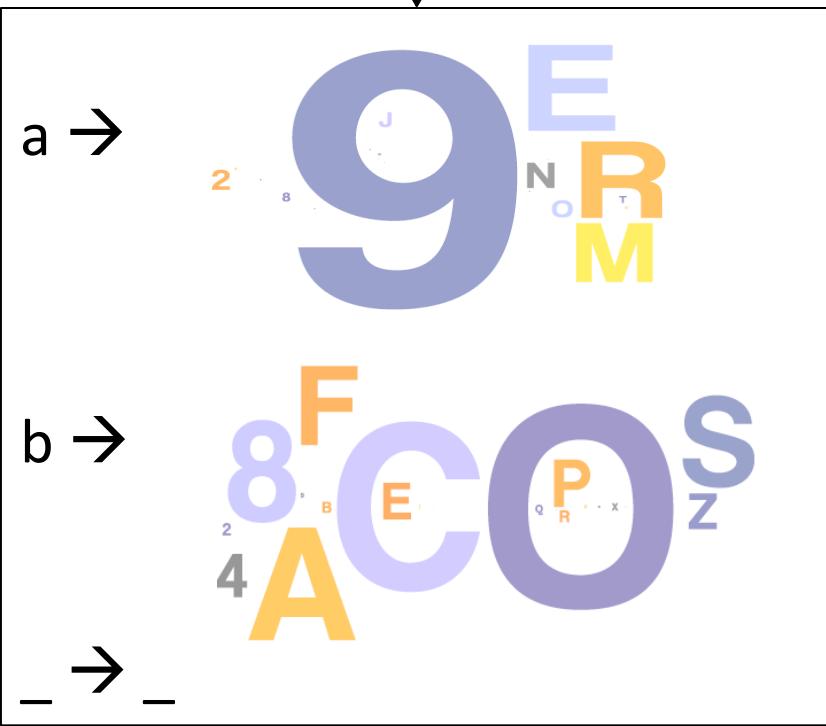
? ? ? ? ? _ ? ? ? ? _ ? ?
V A S 9 2 _ **9 F A E** _ **A R** _
? ? ? ? _ ? ? ? _ ? ? ? ? _ ...
A P A M _ **Z O E** _ **Z O R 9** _ ...

Letter Clustering

Trigram model over {a, b, _ }



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

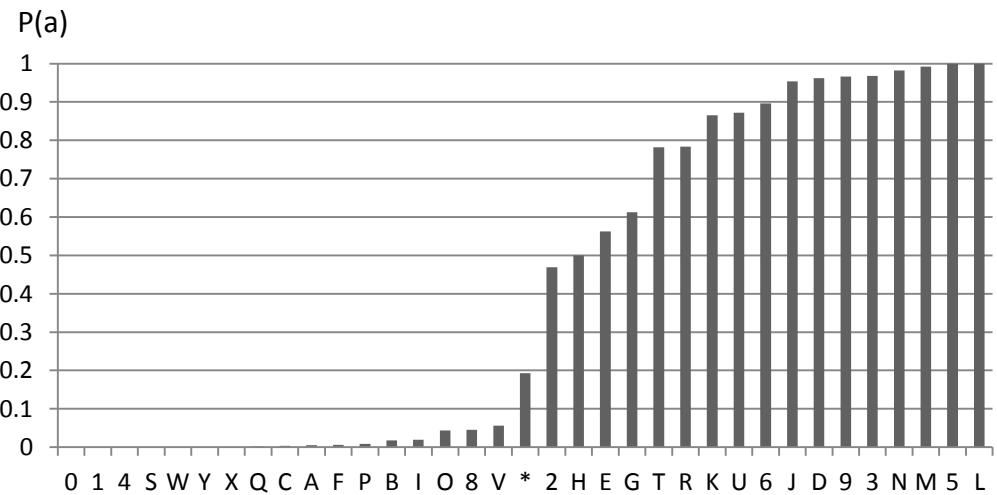
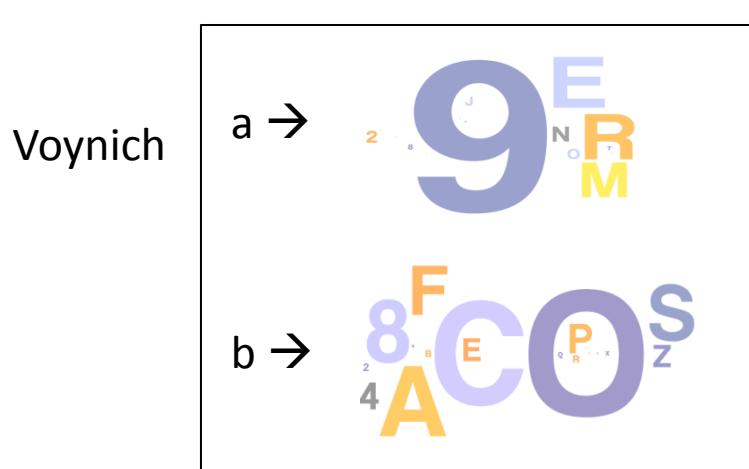
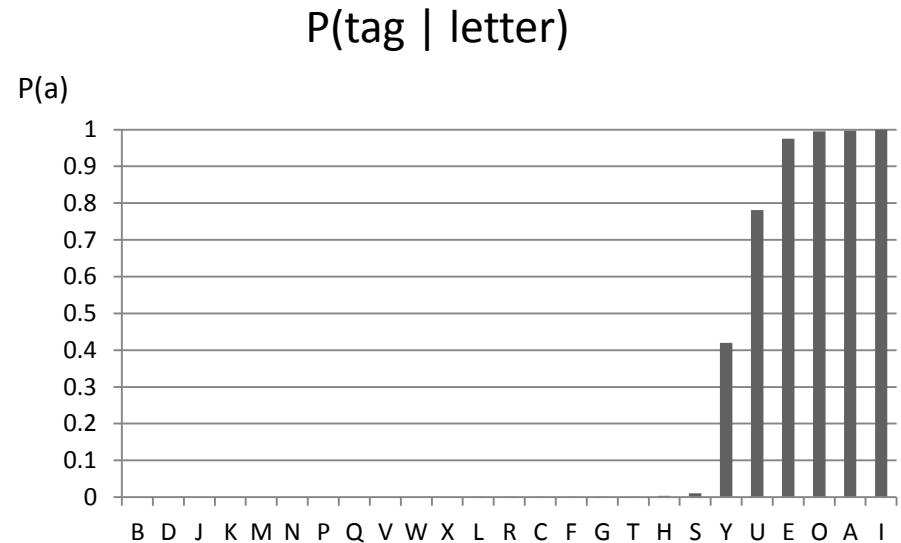
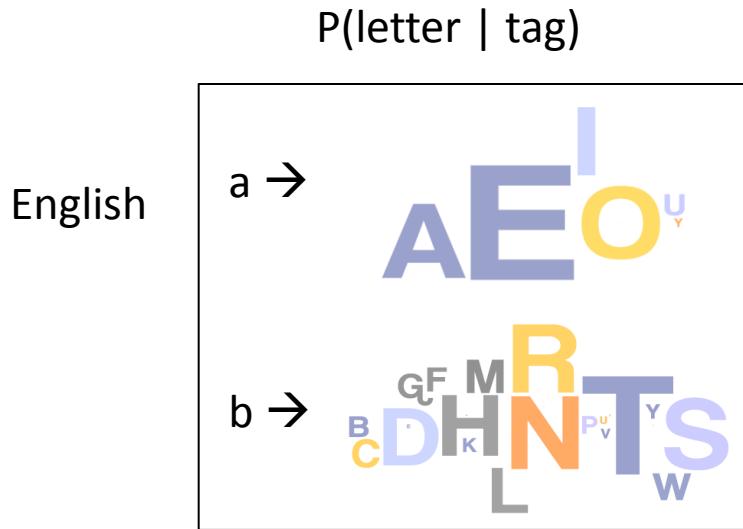


V A S 9 2 _ 9 F A E _ A R _ A P A M _ ...

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



Word Clustering

Bigram model over {a, b}

↓

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

1

a →



b →



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

Sample tagging with learned model:

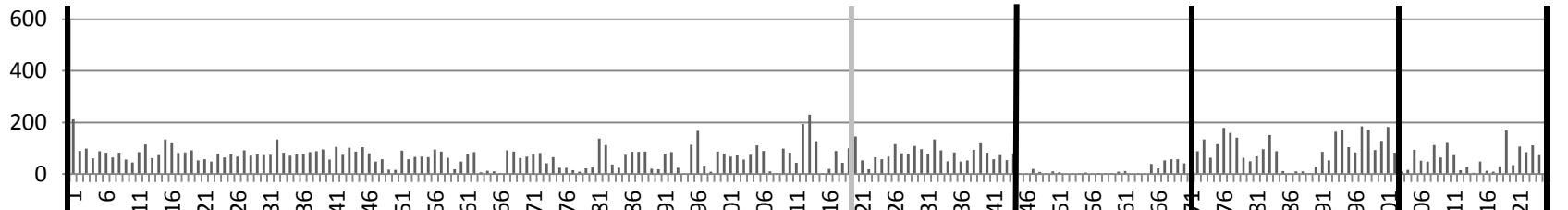
a a a a a
VAS92 9FAE AR APAM ZOE

a a a a a ...
ZOR9 ORC2 9 FOR ZOE89 ...



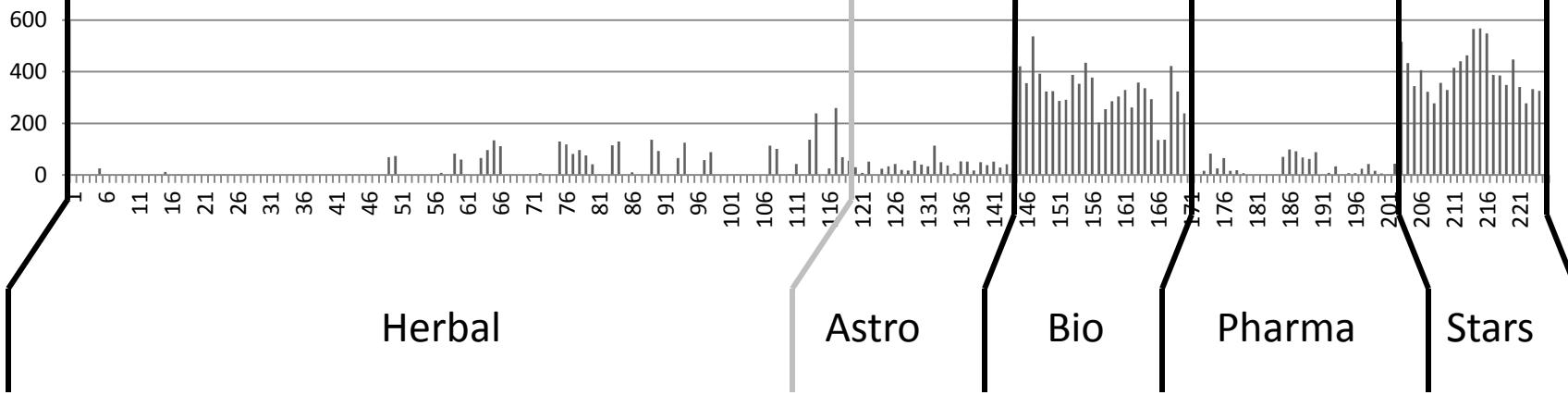
Word Clustering

Voynich words tagged as "a"



← pages →

Voynich words tagged as "b"



Voynich sections, per drawings observed.
Captain Currier's "two languages" (1976).

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), paper in the mmer. Mr. King on his

1970s National Security Agency report recently declassified!

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.

I chose PTAH

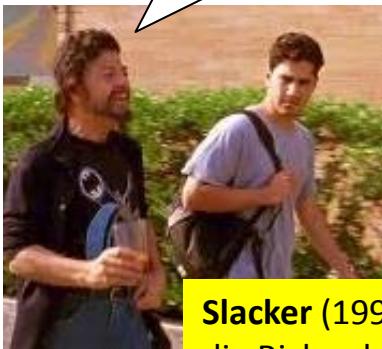
for the present study for two main reasons: first, because of the applications of PTAH to book codes, and second, because I wished to learn more about PTAH itself [REDACTED]

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.



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

Slacker (1991)
dir. Richard Linklater

Association for Computational Linguistics

1950s

1960s

1970s

1980s

1990s

2000s

2010s

2020s

1970s paper on
HMM Voynich

1993 paper
on Statistical
Machine
Translation

2011 paper on
HMM Voynich

ACL applies statistics
to language processing
problems

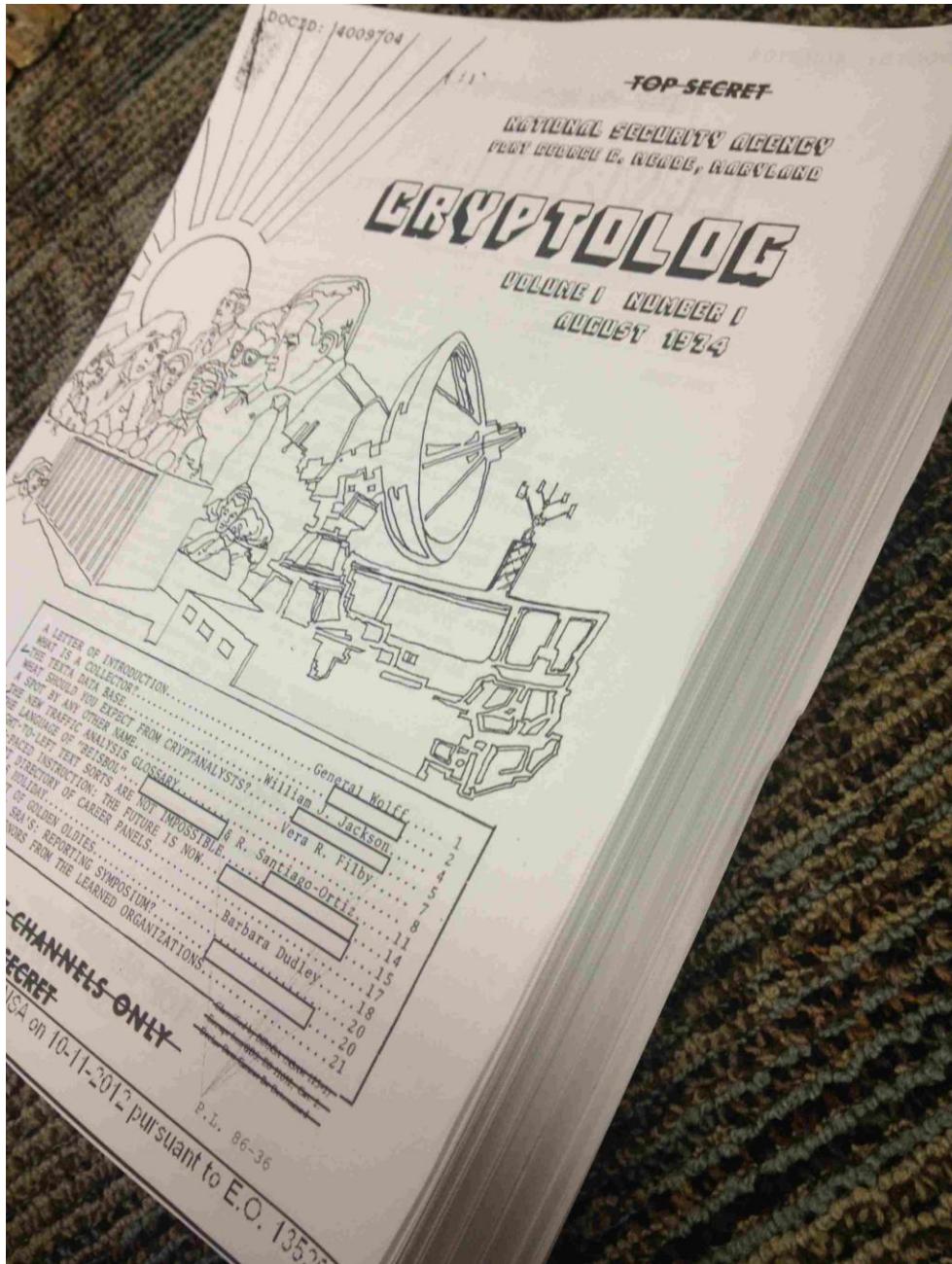
CRYPTOLOG

NSA newsletter declassified
in 2013.

4400 pages (1974-1997).
238 Mbyte PDF file.

Covers intelligence gathering,
linguistics, military, cryptography,
office space, pay grades, human
factors, etc.

Heavily redacted.



CRYPTOLOG: Voynich

DOCID: 4009723

UNCLASSIFIED

The Voynich Manuscript

When a newspaper editor needs a filler, he can always fall back on the Loch Ness Monster or the Abominable Snowman. For the editor of a cryptologic magazine the obvious device is another blurb on the subject here discussed. So, evidently, thought a former editor, among whose effects the following paragraphs were found.

Is the Voynich manuscript "real"? No. Is it a hoax? No. What is it, then? A make-believe--an elaborate fantasy produced purely for the satisfaction of the maker.

That was my reaction the first time I looked at it closely, but faced with all the profound theories about it I lacked the courage to say so. However, a recent rereading of Elizabeth Friedman's article in the Washington Post (August 5, 1962) and of Brigadier Tiltman's paper in the NSA Technical Journal (Summer 1967), plus some phenomena I have seen in the meantime, have emboldened me to give the world the benefit of my thoughts.

The Voynich Manuscript, an object of interest off and on since the seventeenth century, contains over 200 pages written in a partially cursive alphabet which has proved indecipherable. Equally enigmatic are the large number of drawings -- of plants, few of which are identifiable, and of naked women sitting in tubs or emerging from pipes (one writer has called the latter a "plumber's nightmare").

The history of the manuscript, which has been detailed in other places, needs only passing mention since it does not throw any light on the content. Dating from about 1500, it was said by Joannes Marci, mathematician and orientalist at the University of Prague, to have belonged at one time to Emperor Rudolf II (1576-1612). Marci writes in 1666 to the Jesuit Athanasius Kircher, in Rome, that he was making a present to the latter of the manuscript, the author of which, he had heard from another source, was the great medieval scholar Roger Bacon. (How Marci came into possession of it, I do not know.)

Marci himself withheld judgment on the attribution, but at least one scholar since his time became intrigued with the notion of Baconian authorship. Professor William Newbold of the University of Pennsylvania was convinced that it

(FOUO) An example of all of these problems is the Voynich manuscript, a unique European manuscript thought to date most probably from the 15th or 16th century, which has resisted solution, not only by philologists early in this century, but by NSA cryptanalysts as well.

The Voynich Manuscript Revisited

P

was an enciphered text prepared by Bacon and he worked on this assumption from 1919 until his death in 1926. He thought he had deciphered some of it, including an occurrence of "R. Baconi" on the last page¹. His "solution" has been convincingly refuted by other scholars, who however have not offered anything better.

I now rush in where angels fear to tread. Although not a specialist in Old Norse, I am convinced that the manuscript is a text in fifteenth century Danish or Norwegian -- not a cipher, and not an artificial language, as has also been suggested. For reasons too complex to go into here, I have tentatively ruled out Old East Norse (that is, Old Swedish) and rejected altogether the second branch of Old West Norse, Old Icelandic. The reasoning which suggested Dano-Norwegian is given below.

Most of the manuscript has a depressing number of repeated words and phrases, of little help unless collateral information is available, suggesting that these are prayers, incantations, or formulas of a specific character. This is

¹The information in this paragraph and the preceding paragraph was taken from *Horizon*, January 1963 (Vol. V, No. 3).

(UNCLASSIFIED)

CRYPTOLOG: Machine Translation

is machine translation. Machine translation is actually having a computer prepare a translation. There was to have been no difference in quality or style between a translation done by a machine and one done by a person. Georgetown University was very active in the field for some time. Progress wasn't as easy and rapid as had been anticipated, however, and in 1966 the Automatic Language Processing Advisory Committee published a report recommending that research along machine-translation lines be cut back. This report sharply curtailed federal funding. There is still, however, research being done both here and abroad, and there are several machine-translation systems that claim to be operational. One is the METEO project in Canada, which developed a system that translates weather reports from English into French. CULT (Chinese University Language Translator) in Hong Kong translates two periodicals into English. And a system was developed by a U.S. company for FTD and was adapted for use by NASA during the Apollo-Soyuz Test Project. These systems differ a great deal in their approach and in the amount of pre-editing and postediting that is necessary, but all are true machine-translation efforts.

At present, NSA has a rather limited machine-translation effort.

As machine translation stands today, we haven't reached the stage where we can feed a "source" (foreign-language) text into a computer and produce a text in the "target" (in our case, English) language which is as good as the human product, not without extensive pre-editing or postediting. But in the science and technology world, current machine translation has a place. Some scientists prefer it to the

Partial Machine Translation: A Final Report (U)

P16
and
P16
P.L. 86-36

Partial Machine Translation (PMT) is a word-for-word or phrase-by-phrase "translation" from one language to another. The quotation marks are placed around the word "translation" to show that PMT is not what most people consider a true (or full) translation, but the quotes are inserted reluctantly. Although it may be difficult to read the

EO 1.4. (c)
P.L. 86-36

DOCID: 4010113

TOP SECRET UMBRA

CRYPTOLOG
Fall 1995

MACHINE TRANSLATION:

What can it do for us?

by [redacted]

EO 1.4. (c)
P.L. 86-36

TOP SECRET UMBRA

CRYPTOLOG: Evaluating Translations

An Objective Approach to SCORING TRANSLATIONS

Reprinted from *QRL (Quarterly Review for Linguists)*, November 1973

Author's note: The philosophy underlying the translation grading system described in this paper has been developed and applied by Emery Tetrault and myself, with many valuable suggestions from our colleagues on Professional Qualification Examination (PQE) Committees and from other Agency linguists. My use of the pronoun "we" reflects this collaboration. I personally take full responsibility for presenting our findings here.

*

Translation as an intellectual activity has been practiced since antiquity for practical as well as aesthetic reasons, but even today we

tuitive judgments across lang in source language-to-English

Over the past 2 or 3 years I have developed a way to sco which may obviate this proble tent even though our results been far from perfect (total grading any kind of connected impossible). Our first large system, which I will describe the Russian QE. We have sub in a number of other PQEs inv languages, mainly Indo-European other families. The results aging enough in both instance mend its use in the QE Handb

CRYPTOLOG: Linguists

LET'S GIVE LINGUISTS A BIGGER PIECE OF THE PIE!

• Recognition

Most linguists specified that they want recognition above all else. A number felt that lack of recognition of the worth of linguists is evident in the inability of Agency linguists to compete successfully with managers or others for promotion. Despite almost unanimous complaints about lack of recognition, few specific suggestions were made regarding how that recon-

At the
rests the m
foreign lan
their [redacted]

own

12. PUBLICATIONS (List titles; do not confuse this with reports prepared as a regular part of the job)

SOME TIPS ON GETTING PROMOTED

Article based on talk given in April 1978 to WIN (Women in NSA)

Promotion. The word inevitably stirs response of some kind in every red-blooded NSA employee: hope, pleasure, challenge; despair, frustration, disappointment; even inertia, resentment, resignation. Despite disparate views on promotion,

[redacted] manager."

TEACHING COMPUTER SCIENCE TO LINGUISTS

by [redacted] Pl6

serving on the Agency Grade 14 [redacted]. In my experience there has simply been no time to learn the language. It is so foreign to me that I have never been able to hold impressions and reinforced them. The critical importance of the language is covered in this article.

Personnel Summaries

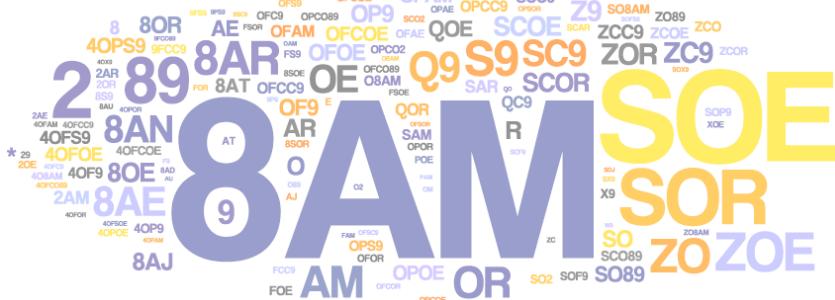
cou
sep
lin
sys
see
inc
doo
ord
axie
pro
fail
ext

Back to Word Clustering

Bigram model over {a, b}

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

a →



b →



VAS92 9FAE AR APAM ZOE ZOR9 ORC2 9 ...

Let's try 10 clusters.

Let's limit ourselves to the more homogenous Bio + Stars sections.

10-Class Word Clustering: English

etc

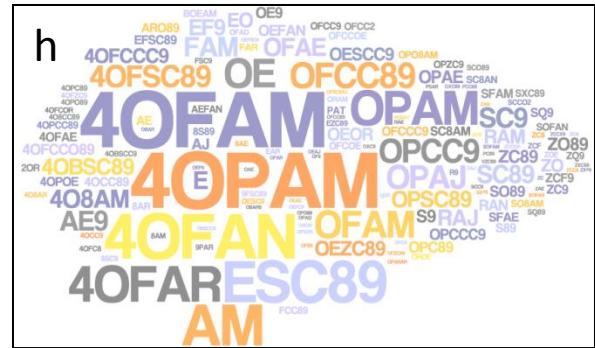
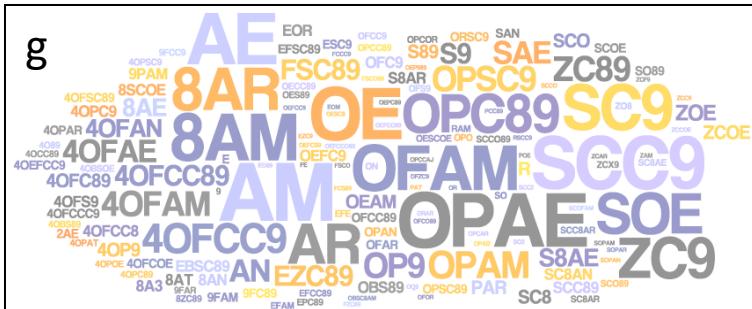
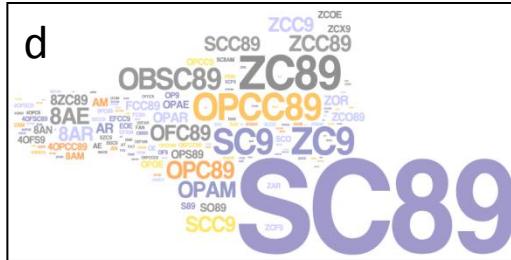
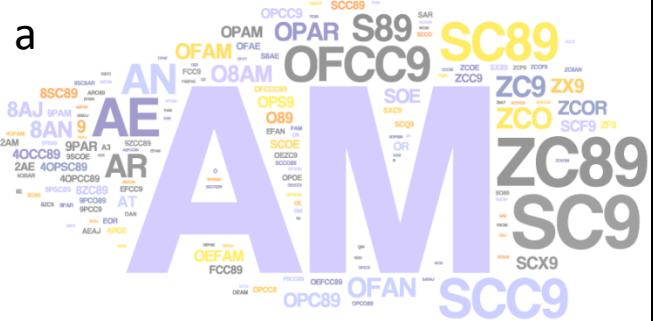
etc

etc

etc

etc

etc

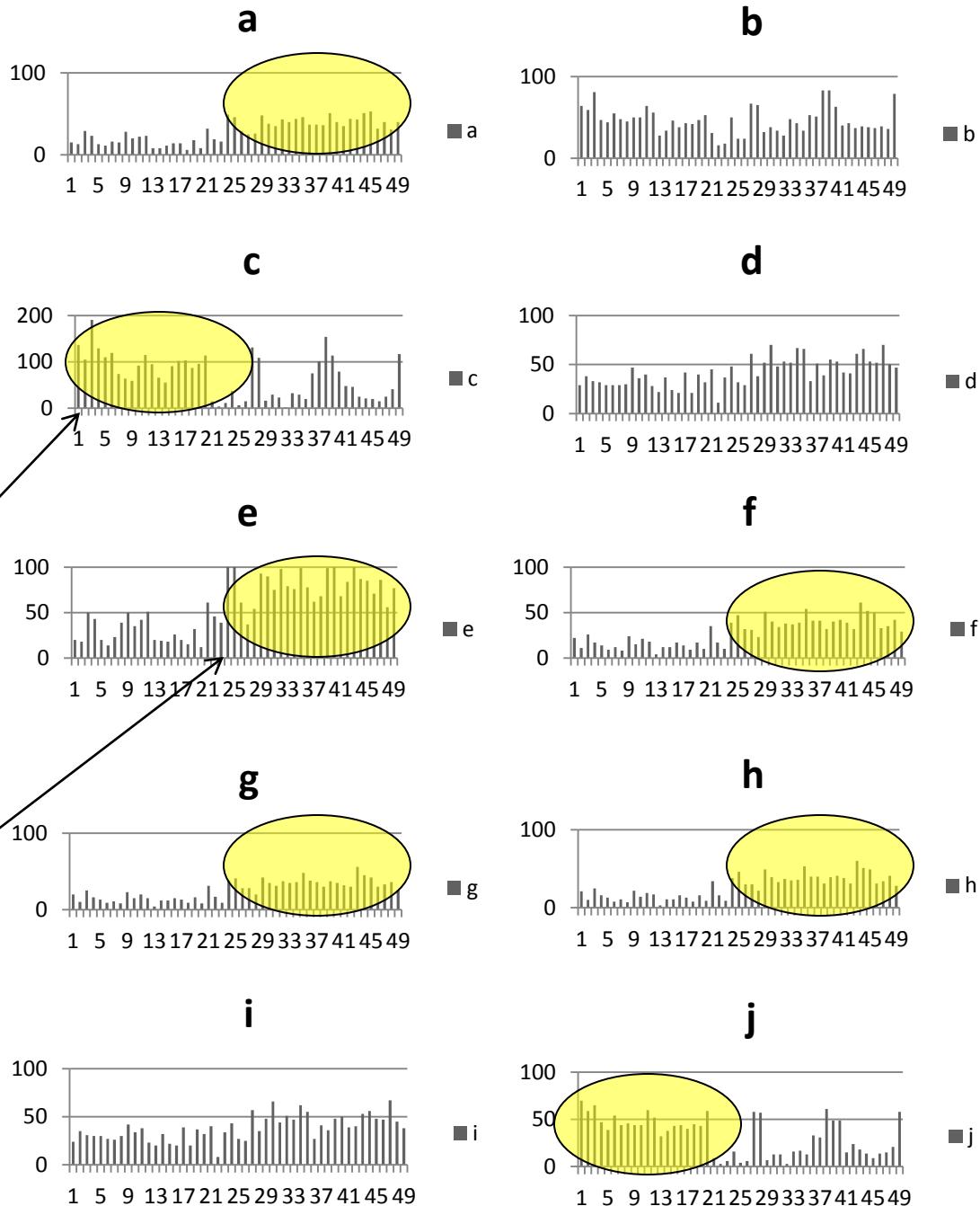


10 Classes:
Voynich-B

10 clusters: Voynich-B

Tags per
page.

“Bio” words vs.
“Stars” words



Does VMS Have Content Words?

Measure the saliency of a word in a page
with TF-IDF

$$\text{TF-IDF}(w, d) = \text{TF}(w, d) \times \log \frac{N}{\text{DF}(w)}$$

times that word w
occurs in page d

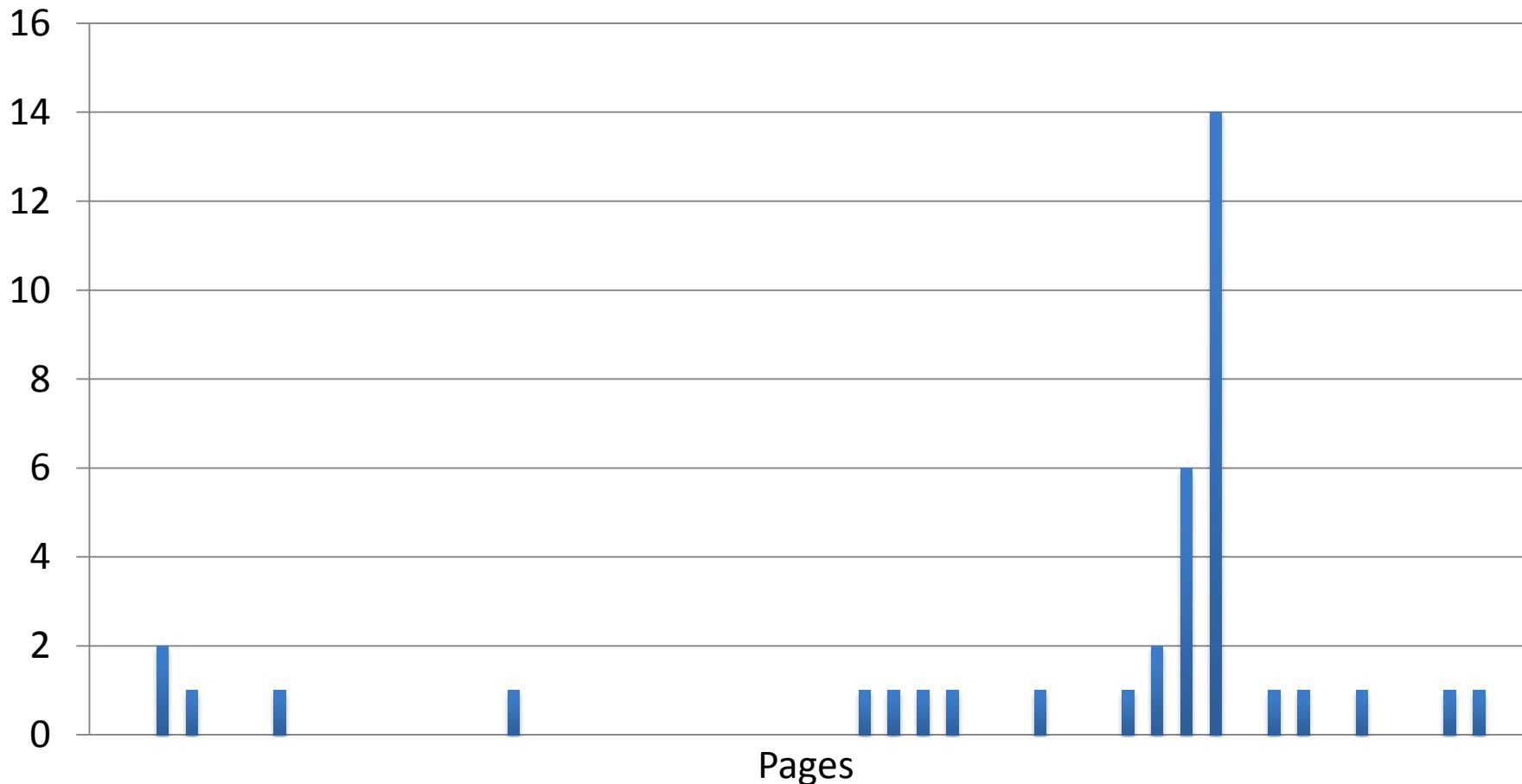
pages that
contain word w

Does VMS Have Content Words?

OFCC9 ZC89 8AN ZC9 SCFCC89 ROR 4OFAN SAE FAN ZOFC9 **EPAN** OR 9FCC9 EZC9 FAE 40FCC89 FAR SC9 R **AN** ZCC89 40FCC9 4OF9 **FCCOR** BAR OFC9 SQ9 BSC89 **SX9** EZC89
OFCC89 40FC89 OESC89 2AE S89 OPC9 **ESC89** 40FAE OFAR FCC89 2AN OPAR **2AM** OPZC89 **BSC8AR** 8AE ZAE OPAN SAR AR ESC9 EOR AJ OEFCC9 **SCQ9** ZX9 OPC89 8AT ZCQ9 E
SFAE 40FC9 SC9 **SQC89** EPC89 OFAE OP9 40FCS89 OPCC89 4OPCC9 40PCC89 OFC89 OPAJ 2ZC89 2 O 8AM EFAR 8E ES89 FC9 SCC89 8AR9 ZCOR 40FCOR AM ZC9 SE FCC9
4OPAN OFAN 4OPS89 OPCC9 SXC9 **EFC89** 40FCC9 AE SC8AN **EFC89** ZC8 SOE OEAN ZC8AE 40FAM 9 SCOE EFC9 4OBSC9 OFS9 **SCAE** AEOE ZCAR ORAM
40XC9 RAM OFC8AE SCQC9 4OPCC2 ZC8AJ OFCCC89 40XC89 SCXC9 OPCOE 2AEFCC89 PCC89 SXC89 **PCC9** ZCFCC9 ZCCOE ZCAE ON 40FCAR SR BAJ
EAJ 8ZCC89 8SCC9 ZAR SQ*9 EOFCC89 ER SXAE 4OPSC89 ESCOE SC8AR ORAN 4OPSC9 AT OIF*9 4OSC9 OEEFAE 8ZC9 PAR EFO **EFC9** ZC9 AEOJ FCCOE
40FCSC89 08AM ZCFAE **8SC8** 40VSC89 SO89 40FCOE ESC8 SFCC9 8SCC9 EOFAJ 4OF OFCCOE SCBSC89 8OM ZCCX9 PC9 OPC8AR OPC2 40FCS9 OPCO89
SCD89 **40FCO89** OPCC8 AK EVS9 SAM PAT CCC2 OJ EFAJ **FCCC9** SCO2 8AK ZCCF EOP9 ZCFAN 40FCCS9 OFCAE EFCCO89 ZC08 ZC8AN BSAE OPCAE
OPCCAJ 4OAN 4O8 SCCO SCO SCOFCC9 AEAJ **OFCCO** EFCSC9 EFCSC89 SCCFAN OPCCC9 EFAT ESAE OPCCOE SO FCSC89 OEFCCC89 40FC08
40SC89 OPCO ZCC08AR SCOJ OPARAE OAM OEFCCO EFCCOE EFAE EFCCC9 PC8AJ OFC8AN ZOF **40FCCOE** 40FCC02 OPC8AN EFCCC89 SCAJ
SCXC8 OSC9 FSOES8AR AIIB SCPAEZ9 4CCAE 40FCCA2 SCOFCC89 ZCCFZ9 SOFSC9 SX*9 9SCCO8AN 9FCC8AM RCCC9 OEA3 AIF*9 ZFAM 8ZCCO OPSC8C9
OPCOEAT 4OZCO 8SC8AR 9FCCCO2 BSC8C9 OPCO80 BS8AJ OFCO8AN ZCP ZCOPAJ ZPAR OESCO89 ZCQC9 ESCOCFAJ 40FCC08OR SCQ89 4OPC8E
EOC89 SC8C9 EFAK O*OR F98CC89 ZCCP SCOP989 ZZCO PSC8 FCOQC89 40FCZC9 FCCZ089 **OFCSC89** ESR 20 AEEA O*AR 20AM OPCO8AM
40PCC8AM PCC8AN ZCOFAR RF9 SPAR 4OTAN ZCOPSC89 ZFC9 **40FCAN** ZFC089 8ZCCOPCC9 PCAR SOEFCCC89 ESCS89 40CCCO SC8A FCC08AE
PCO OFCOJ 40FCC08 EFC8EFC9 **PCC8** 9SC8E BOEAE B9FCOR SCCV9 OBSC8AE EVSC89 ESCCOE OPCON SCAJAR 40FCOFC89 OPCCOEFCC9 EAN
40FCCAE 40FCCE SC89PCOFAN EFC8AR EFCC8AN OFCAJ PCOEFC8AN EPCCAE U OFCCC2C9 OESAE 4CCAR BOCOFCC9 PC8AN SCBSAJ 40FCCOFAN
FCCE BOEFCOC SCOFCAN I*AR *AN 9ZC OFZ89 ZFCC9 SXAM

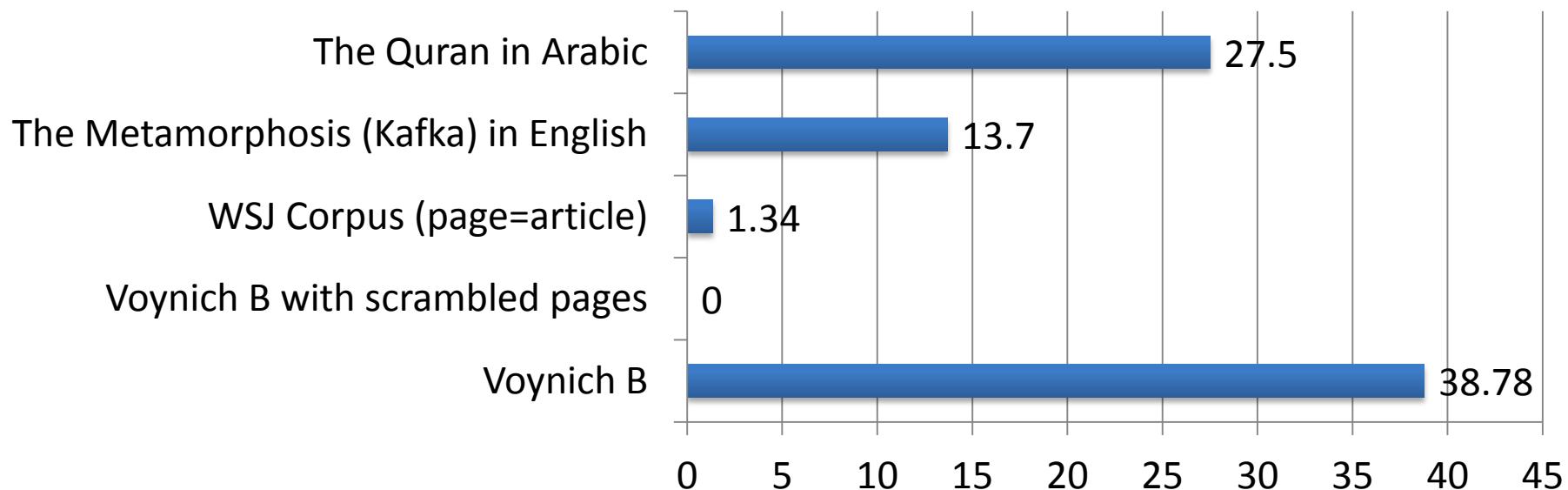
Do Content Words Indicate Topics?

Frequency of EFCC89 in Voynich B pages



Are VMS Pages in Order?

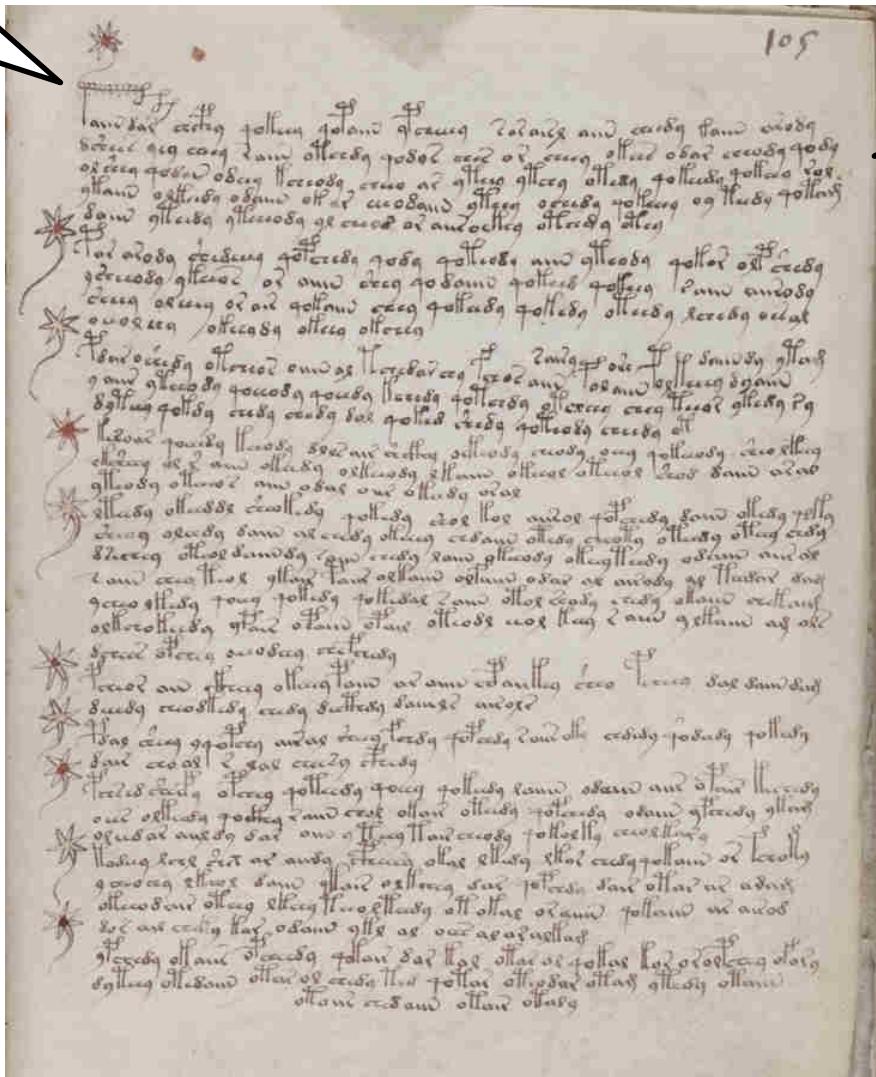
- Measure similarity between a pair of pages using cosine similarity (with bag-of-words)
- Count the % of pages P where the most similar page to P is adjacent to it



Special ligatures
at beginning of
“paragraphs”

Is VMS Prose?

Looks like
paragraph
structure



BUT:
Lines begin and end
disproportionately
often with certain
characters!

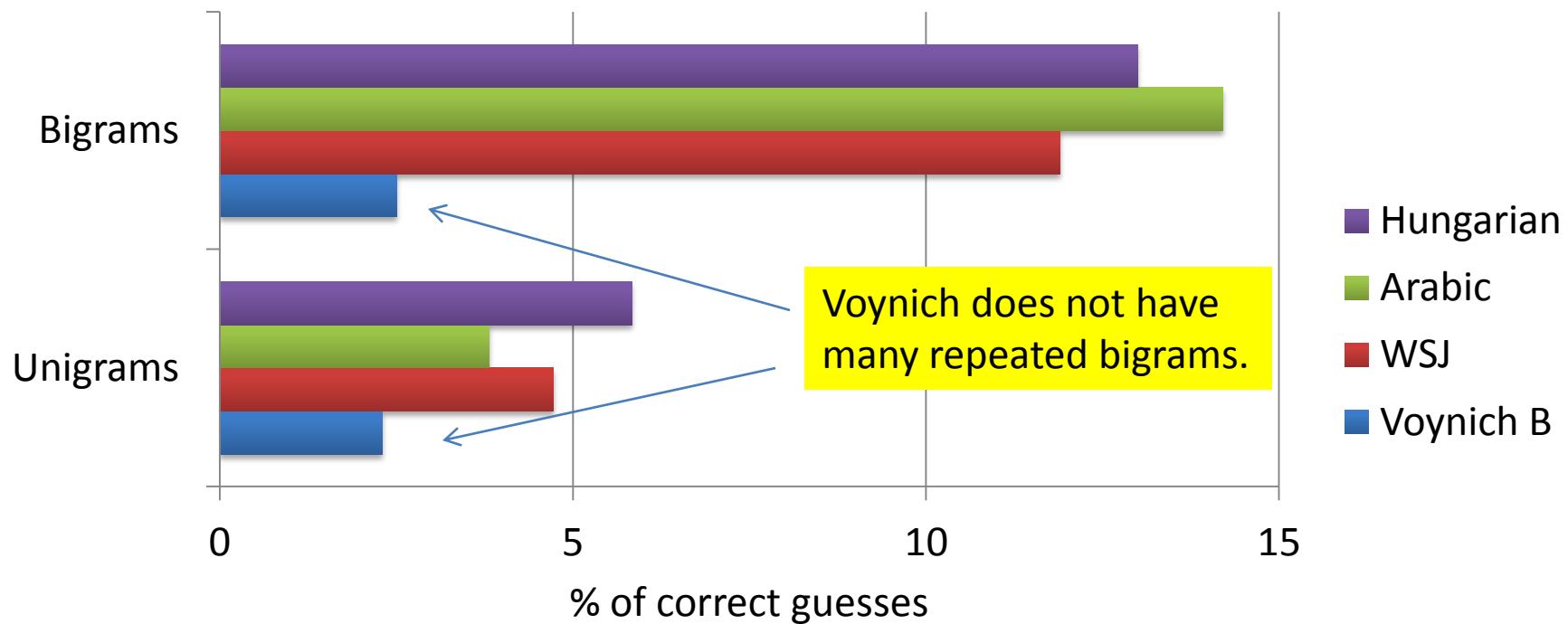
The line is a
functional entity...



Prescott H. Curnier

Are VMS Word Sequences Predictable?

- Guess most likely word to follow current word
- Simulate game from bigram probabilities
90-10 train-test splits



Zodiac Killer Ciphers

Zodiac 408 (solved, 1969)

Δ □ P / Z / 4 B □ K O R A X 9 X K B
W V + E G Y F O D H R □ K I R Y E
M J U L A N I K A R T N Q Y D O O
S F / D □ B R O R A U □ F R A R E
K A L M Z D J A \ 9 F H V W E A Y
□ + @ G D A K I - O R X A O S F
R N I I Y U E L O A R G B T Q S □ B
L O / P □ B □ X R E H M U A R R K

C Z K Q P I B I W O I A L M R A D □
B P D R + T A K □ O \ N F E E N K H F
Z C P O V W I □ O + T L O H O R A L
I A D R O T Y R Y / E D / O X J Q A
P O M A R U T □ L O N V E K H K
J I I K O D A L M N A O Z F P
Φ O K P U A D □ B V W \ + V T O P
A L S □ R A E U F L □ O A D F G □ I M
N K N S O K N / A D Z F A P B V
P X Q W O D F □ A C + O A A D B
O T O R U D + D U D Y O D O S O W
V Z E G Y K E D T Y A D D L T D
H J F B X A D X A D C \ A L I K
D E □ O E O P O R X Q F □ G
Z T J □ G T A J + I J R B P Q W O
V E X R A D W I O P E H M O K I N



Zodiac 340 (still unsolved)

H E R > P L A V R K I O L T G O D
N P + B F □ O □ D W Y . < □ K F □
B Y I C M + U Z G W F S L □ H J
S P P A D V A L □ V O P O + + R K O
D M + T M □ I D T + F P + R O K /
P □ R A L F > O L D □ D C F > O D F
■ + K Q □ I O C H G V . L I
Φ G O □ T F J D O + O N Y F + O L A
D < M + 8 + Z R O F B C Y A O O K
- □ U V + L J + O 9 A < F B Y -
U + R / O L E I D Y B 9 8 T M K O
O < J R J I □ O T O M . + P B F
Φ O A S Y □ + N I O F B C F I □ R
J G F N V F O B O C . B O C .
Y B X O E O D C E > V U Z O - +
I O F B K F O 9 L . F M Q G O
R C T + L E O C < + F J W B I O L
+ O W C F W S O H T / F O B
I F W < A D Y O B T A D C - C
T E I A O Z S K P N H M >

A circle with a crosshair symbol is drawn near the bottom right of the cipher text.

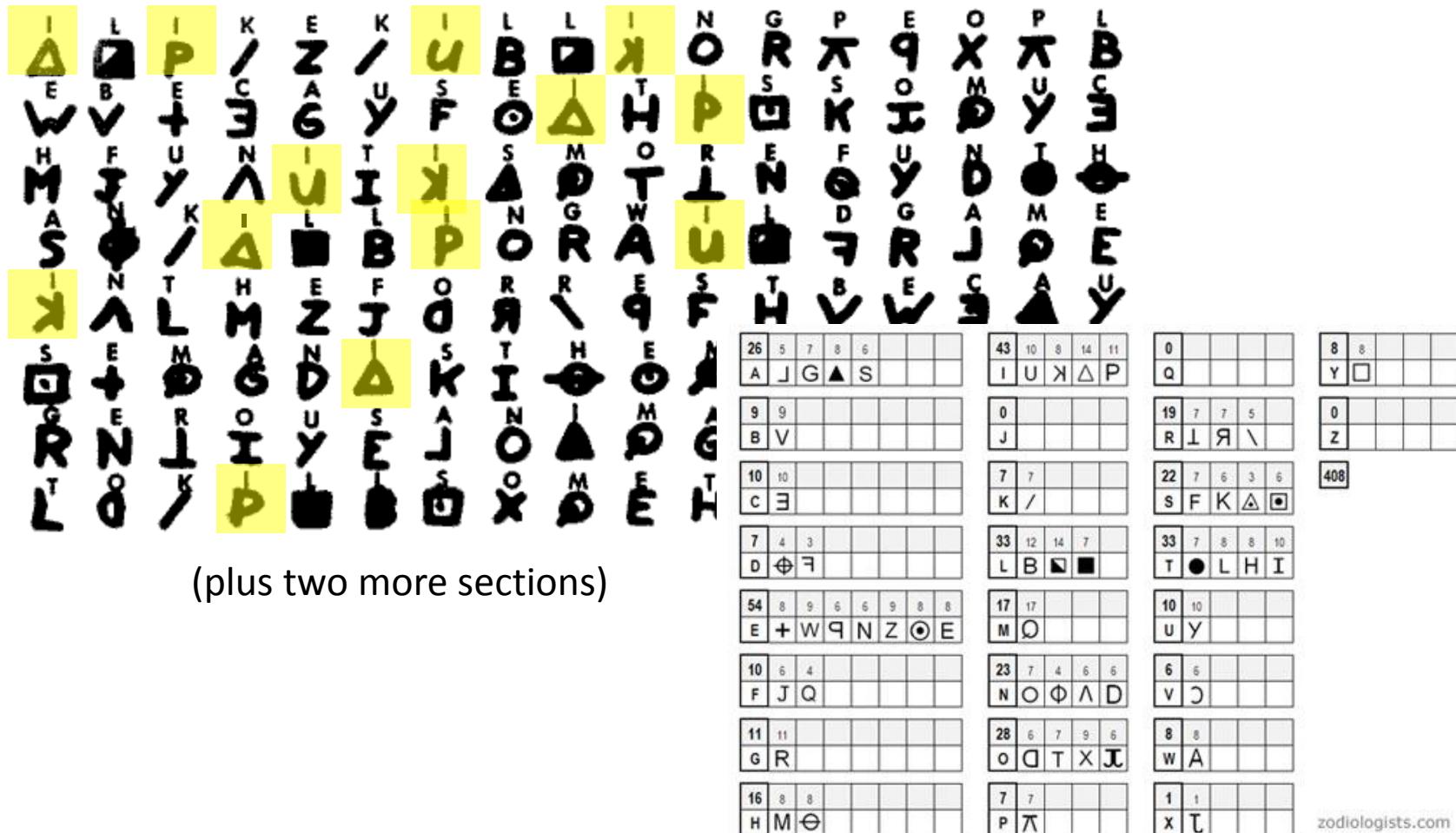
COOP-SFPO
1596-78
11-4-78 GWL
7-7-78

#2 11-4-69

1 2 3 4 5 6 7 8 9 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

Zodiac Serial Killer

408-letter cipher (solved):



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 DANGEROUUE 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 THAЕ 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

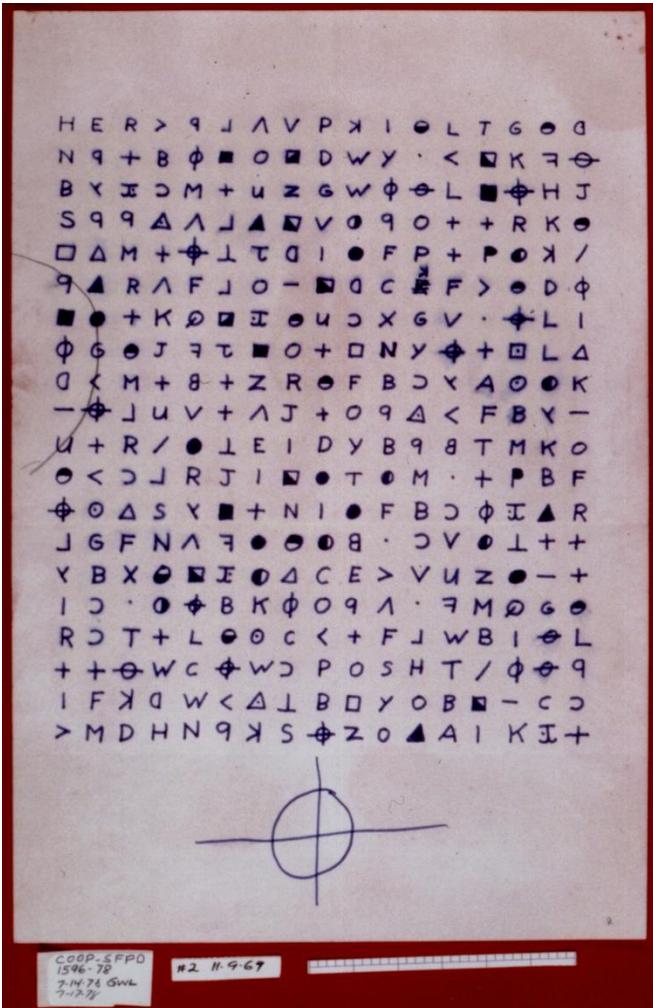
Bayesian models & Gibbs sampling

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

[Ravi & Knight 11]

See also Malte Nuhn's
paper at ACL 2013!

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.

Other Unsolved Ciphers

Beale (1885)

, 111, 95, 84, 341, 975,
14, 40, 64, 27, 81, 159, 215, 65, 90, 1120, 8, 15, 3, 120, 2018, 40, 74, 758, 485,
604, 230, 436, 664, 582, 150, 251, 284, 308, 231, 124, 211, 486, 225, 401, 370,
11, 101, 305, 139, 189, 17, 33, 88, 208, 193, 145, 1, 94, 73, 416, 918, 263, 28, 500,
538, 356, 117, 136, 219, 27, 176, 130, 10, 460, 25, 485, 18, 436, 65, 84, 200, 283,
118, 320, 138, 36, 416, 280, 15, 71, 224, 961, 44, 16, 401, 39, 88, 61, 304, 12, 21,
24, 283, 134, 92, 63, 246, 486, 682, 7, 219, 184, 360, 780, 18, 64, 463, 474, 131,
160, 79, 73, 440, 95, 18, 64, 581, 34, 69, 128, 367, 460, 17, 81, 12, 103, 820, 62,
116, 97, 103, 862, 70, 60, 1317, 471, 540, 208, 121, 890, 346, 36, 150, 59, 568,
614, 13, 120, 63, 219, 812, 2160, 1780, 99, 35, 18, 21, 136, 872, 15, 28, 170, 88, 4,
30, 44, 112, 18, 147, 436, 195, 320, 37, 122, 113, 6, 140, 8, 120, 305, 42, 58, 461,
44, 106, 301, 13, 408, 680, 93, 86, 116, 530, 82, 568, 9, 102, 38, 416, 89, 71, 216,
728, 965, 818, 2, 38, 121, 195, 14, 326, 148, 234, 18, 55, 131, 234, 361, 824, 5,
81, 623, 48, 961, 19, 26, 33, 10, 1101, 365, 92, 88, 181, 275, 346, 201, 206, 86,
36, 219, 324, 829, 840, 64, 326, 19, 48, 122, 85, 216, 284, 919, 861, 326, 985,
233, 64, 68, 232, 431, 960, 50, 29, 81, 216, 321, 603, 14, 612, 81, 360, 36, 51, 62,
194, 78, 60, 200, 314, 676, 112, 4, 28, 18, 61, 136, 247, 819, 921, 1060, 464, 895,
10, 6, 66, 119, 38, 41, 49, 602, 423, 962, 302, 294, 875, 78, 14, 23, 111, 109, 62,
31, 501, 823, 216, 280, 34, 24, 150, 1000, 162, 286, 19, 21, 17, 340, 19, 242, 31,
86, 234, 140, 607, 115, 33, 191, 67, 104, 86, 52, 88, 16, 80, 121, 67, 95, 122, 216,
548, 96, 11, 201, 77, 364, 218, 65, 667, 890, 236, 154, 211, 10, 98, 34, 119, 56,
216, 119, 71, 218, 1164, 1496, 1817, 51, 39, 210, 36, 3, 19, 540, 232, 22, 141, 617,
84, 290, 80, 46, 207, 401, 150, 29, 38, 46, 172, 85, 194, 39, 261, 543, 897, 624, 18,
212, 416, 127, 931, 19, 4, 63, 96, 12, 101, 418, 16, 140, 230, 460, 538, 19, 27, 88,
612, 1431, 90, 716, 275, 74, 83, 11, 426, 89, 72, 84, 1300, 1706, 814, 221, 132,
40, 102, 34, 868, 975, 1101, 84, 16, 79, 23, 16, 81, 122, 324, 403, 912, 227, 936,
447, 55, 86, 34, 43, 212, 107, 96, 314, 264, 1065, 323, 428, 601, 203, 124, 95, 216,
814, 2906, 654, 820, 2, 301, 112, 176, 213, 71, 87, 96, 202, 35, 10, 2, 41, 17, 84,
221, 736, 820, 214, 11, 60, 760.

Taman Shud (1948)



FBI (1999)

Dorabella (1897)

July 14. 97

Kryptos (1990)



UOXOGHULBSOLIFBBWFLRVQQPRNGKSS
TWTQSJQSSEKZZWATJKLUDIAWINFB**NYP**
VTTMZFPKWGDKZXTJCDIGKUHUAUEKCAR

NYPVTT = BERLIN (2011 clue)

AKPNT E GLSC - SE ERTE
VLSC MTSE-CTSE-LSE-FRTSE
PUKTRSEONPSCHWLD NCSE
WLD XCRCHMSP NEDLSE

(2 pp total)

Collected Ciphers

Mina ein geita, af perer-oli,
ni lara obet pera, menu mina
lurus quon var all in der
ujos. Je an lora
af estra SONORAM
nek o oli men.

72 - wyn 4-a e-rn-3 et'a 4-a 37 zbm r-n
5m y-r-E7a EH ym BWTAS km2n 37a crt 0E0Z
-5 37 zbm Ky u2r-E7a, SW7-2 vBn amnkrtan
SAS krm 37a

1640. 20 July
Zifra 211202042A043304
4302 01

+ many more!

Writing as a code for speech

Archaeological Decipherment

ciphertext

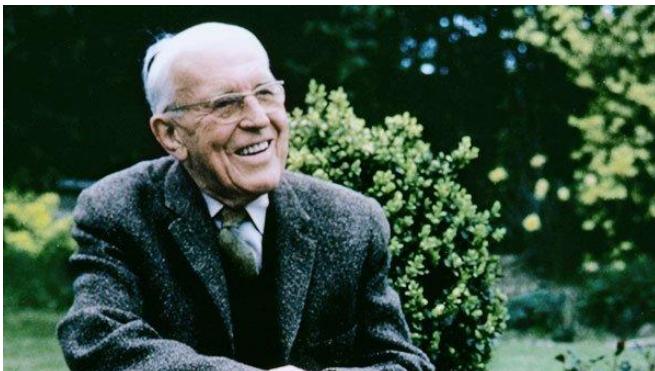


Mayan
glyphs

Archaeological Decipherment

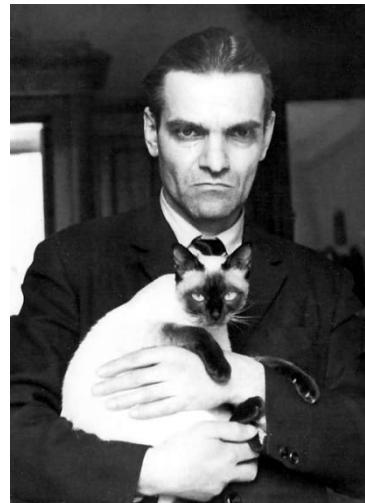
Thinks Mayan decipherment should be based on ideographic rather than linguistic principles.

Resists notion that the glyphs have a phonetic component.



J. Eric S. Thompson

It's phonetic.



Yuri Knorozov

ciphertext



Mayan
glyphs

Archaeological Decipherment

- Mayan glyphs
 - Egyptian glyphs (Rosetta Stone)
 - Linear B
- etc

Computer did not play much of a role in these successful decipherments

Archaeological Decipherment

ciphertext

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

Archaeological 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..."



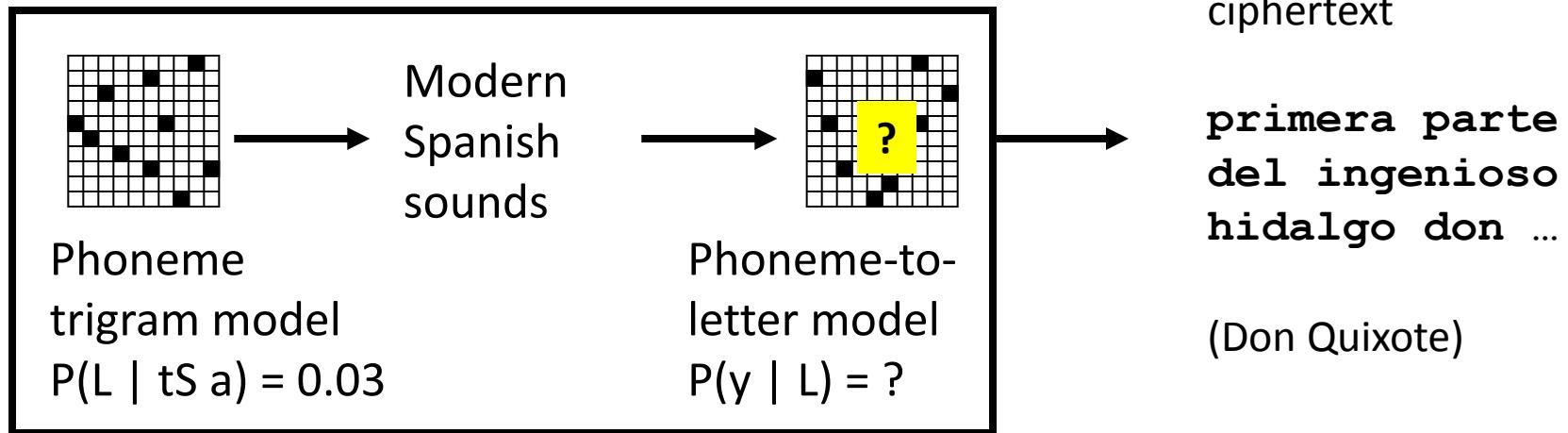
OUR HERO

ciphertext

primera parte
del ingenioso
hidalgo don ...

(Don Quixote)

Archaeological 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

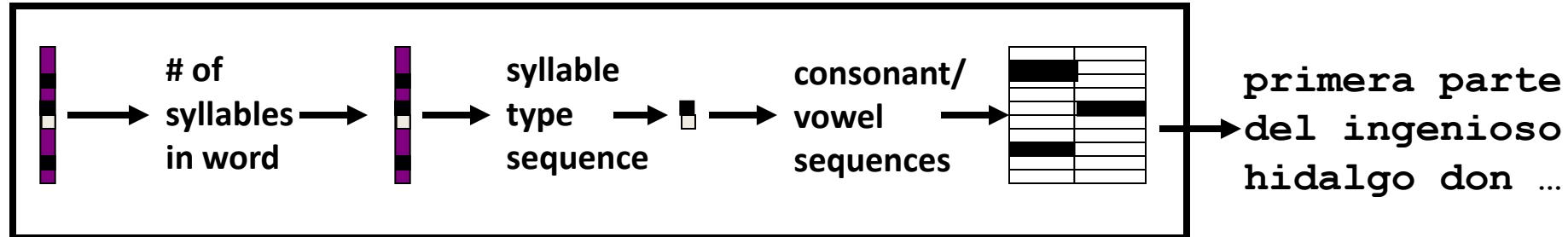
EM approach = 93% accurate phonetic decipherment

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 $P(c | p)$ 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

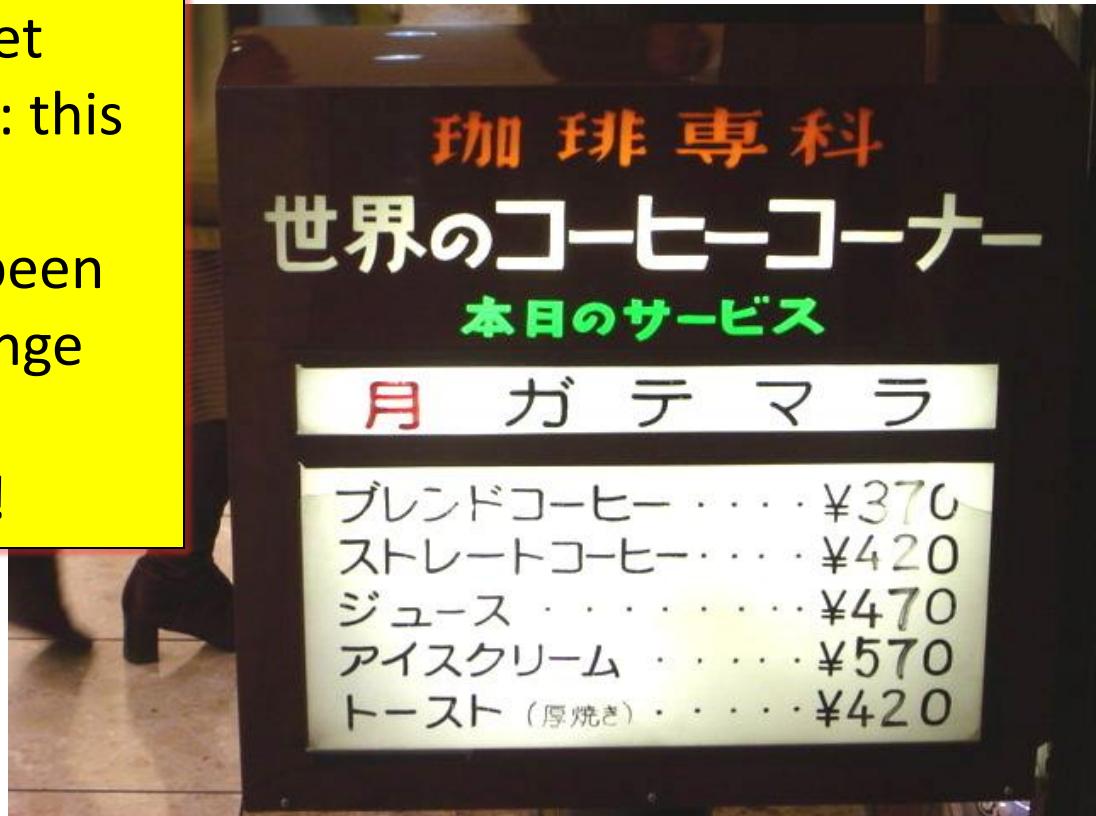
V = vowel phoneme

[Knight et al 06]

See Y. Kim & B. Snyder's
ACL 2013 paper addressing
100s of human languages!

Practical Detour: 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]

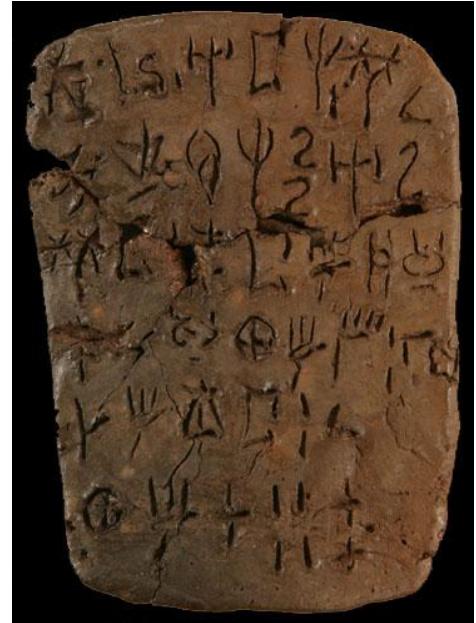
Undeciphered Writing Systems

Undeciphered writing systems

Indus Valley
Script
(3300BC)



Linear A
(1900BC)



Rongorongo (1800s?)

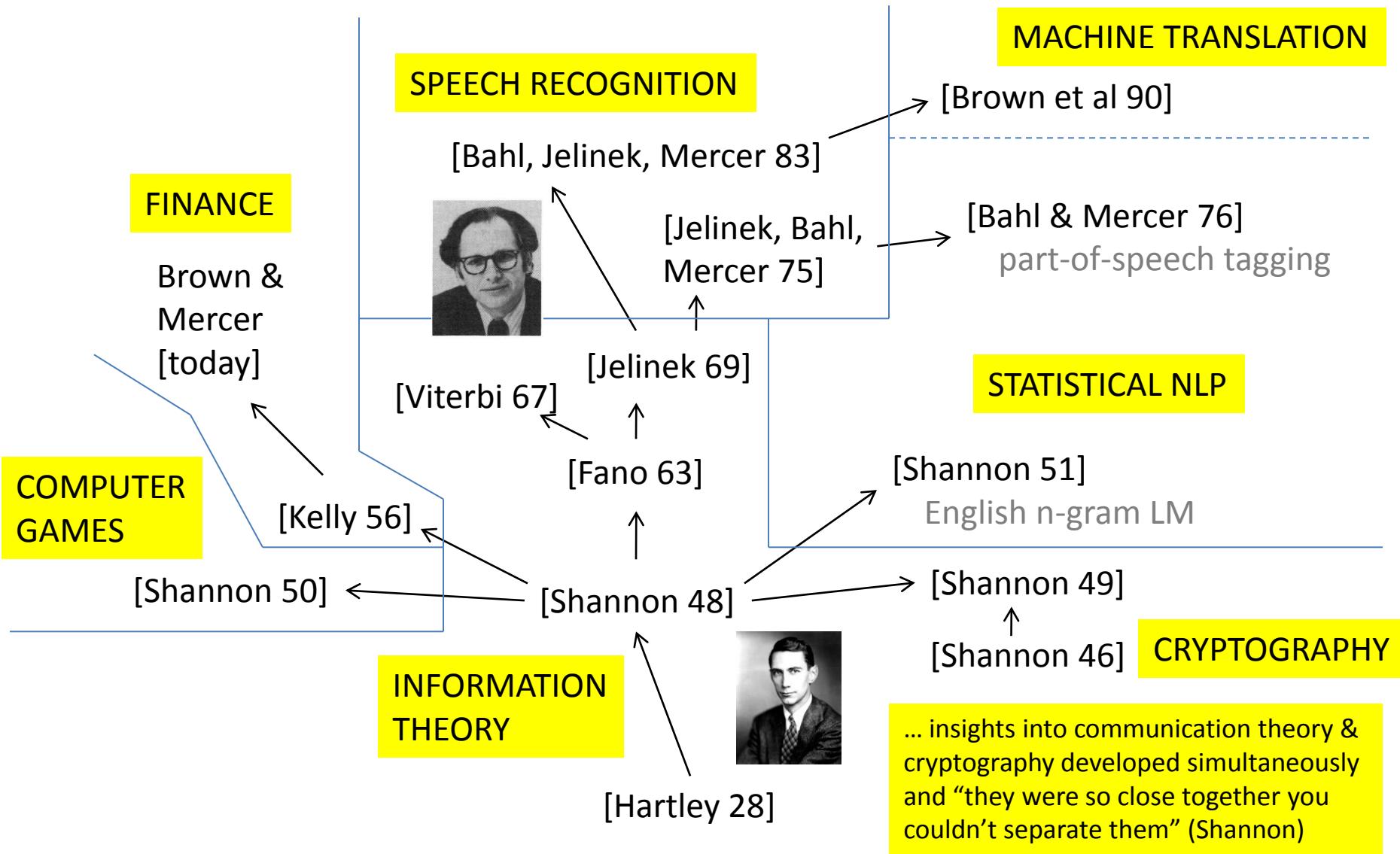


Phaistos Disc (1700BC?)

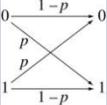


Conclusions

Decipherment and NLP



Decipherment and NLP

| | 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