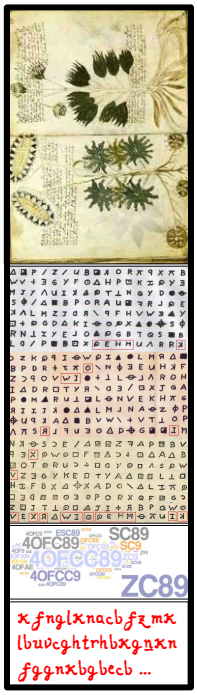


Decipherment

Kevin Knight
Information Sciences Institute
University of Southern California

includes joint work with:
S. Ravi (USC/ISI, now Google)**, Q. Dou, K. Yamada** (USC/ISI)
B. Megyesi, C. Schaefer (Uppsala Univ.)
R. Barzilay, B. Snyder (MIT)
S. Reddy (Univ. Chicago, now Dartmouth)

ACL Tutorial August 2013

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 ##### V
L 10 ##
M 6  #
N 1  .
O
P 1  .
Q 10 ######### V
R 3  .
S
T 7  ### V
U
V
W 1  .
X 5
Y 7  #### V
Z 2  .
```

```
a      e.a   .a           .e      .
KDCY LQZKTLJKX CY MDBCYJQL: "TR

.    .a     .e   a   . ee.e  .
HYD FKXC, FQ MKX RLQQIQ HYDL

 a       .  .     e .e .a
MKL DXCTW RDCDLQ JQMNKXTMB

.     .e   a .a.      e.a
PTBMYEQL K FKH CY LQZKTL TC."
```

didn't create "ae"

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

```
a      e.ao .a           .e     o.
KDCY LQZKTLJKX CY MDBCYJQL: "TR

.    .a     .e  a   . ee.e .
HYD FKXC, FQ MKX RLQQIQ HYDL

 a      o. .     e .e .a o
MKL DXCTW RDCDLQ JQMNKXTMB

.o    .e   a .a.      e.ao  o
PTBMYEQL K FKH CY LQZKTL TC."
```

don't like "ao" – back up!

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

Pattern word dictionaries

KDCY (LQZKTLJKX) CY MDBCYJQL: "TR

abcdeafdg

```
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
```

HYD FKXC MKX (RLQQIQ) HYDL

abccdc

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

MKL DXCT CDLQ JQ TMB

PTBMYEQ KH CY L L TC."

abcdefghiklm

```
consumptively
copyrightable
documentarily
lycanthropies
musicotherapy
semivoluntary
subordinately
unpredictably
```

```
OR, NORWEGIAN!

filmprodusent
kurspamelding
publikasjoner
stylemarginpx
uproblematisk
```

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

or perhaps:
A = ɣ ¡ l ɣ ɾ
B = û
C = ȯ ɲ
D = ₹
E = ẋ ƒ ∆ ₹ ƒ î ɬ ɔ
F = ₱
G = ÿ …

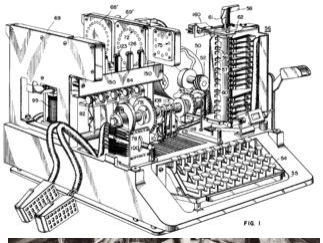
- Polyalphabetic ciphers
  - the secret key changes at each plaintext letter token
  - e.g., rotate through 10 different keys
- Transposition ciphers

Cipher Types

- http://cryptogram.org/cipher_types.html
  - documents ~70 types
- E.g., RUNNING KEY cipher
  - key = agreed-upon standard English text
  - ciphertext(i) = [plaintext(i) + key(i)] mod 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 → J

Secret key = initial rotor ordering and settings

>Billions of initial configurations.

Substitution table changes with every keystroke:

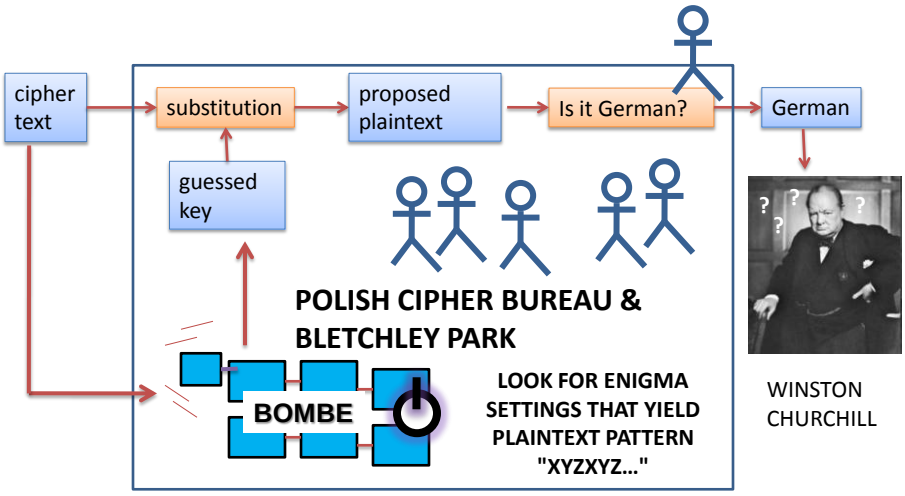
NNN → JTE

Rotates through 1000s of substitution keys.

Reversible behavior

NNN → JTE → NNN

Breaking Enigma



WINSTON CHURCHILL

Enigma

- Mathematical breakthroughs:
  - Log-odds for weight of evidence [Good, Turing]
  - Smoothing with prior [Good, Turing]
  - Information theory [Shannon]

elegant, powerful, widely-applicable mathematics

- 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)

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 c inverse rods for the middle wheel. The rods are paired off R.H.W. couplings, i.e. M.W. output. This has been done for fx, ep which arose in the DANZIGVON crib in Fig 55, assum in the middle. The pairs in each column of this set up give possible M.W. couplings. We have now to find out whether these coupli Our procedure is rather different according as the U.K.W. rotate. In the case that the U.K.W. does not rotate it will b 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

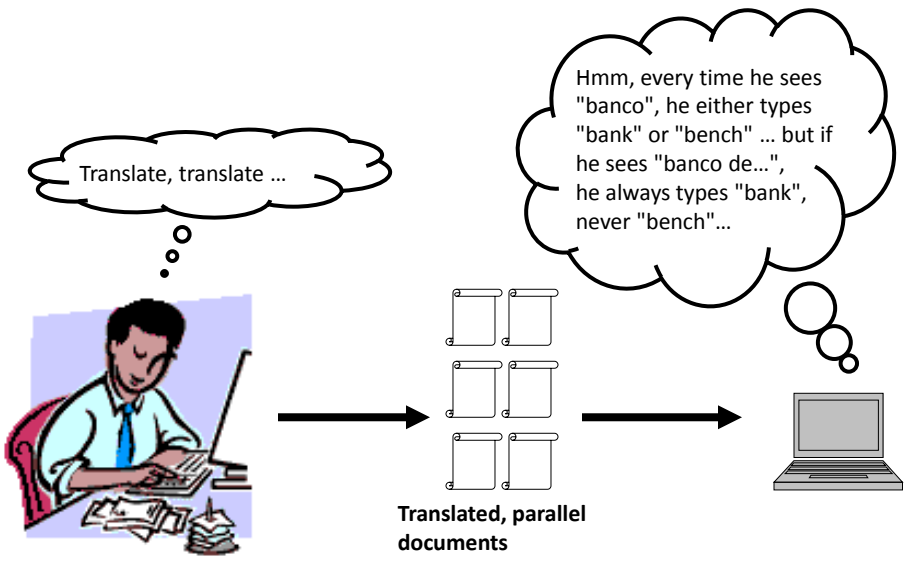
"When I look at an article in Russian, I say: this is really written in English, but it has been coded in some strange symbols. I will now proceed to decode." -- Warren Weaver (1947)

OUR HERO

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

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

Statistical Machine Translation



Parallel Corpus

12 English sentences in English and Spanish.

1a. Garcia and associates . 1b. Garcia y asociados .	7a. the clients and the associates are enemies . 7b. los clients y los asociados son enemigos .
2a. Carlos Garcia has three associates . 2b. Carlos Garcia tiene tres asociados .	8a. the company has three groups . 8b. la empresa tiene tres grupos .
3a. his associates are not strong . 3b. sus asociados no son fuertes .	9a. its groups are in Europe . 9b. sus grupos estan en Europa .
4a. Garcia has a company also . 4b. Garcia tambien tiene una empresa .	10a. the modern groups sell strong pharmaceuticals . 10b. los grupos modernos venden medicinas fuertes .
5a. its clients are angry . 5b. sus clientes estan enfadados .	11a. the groups do not sell zenzanine . 11b. los grupos no venden zanzanina .
6a. the associates are also angry . 6b. los asociados tambien estan enfadados .	12a. the small groups are not modern . 12b. los grupos pequenos no son modernos .

Parallel Corpus

12 English sentences in Centauri and Arcturan.

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

Centauri/Arcturan

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

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

Centauri/Arcturan

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

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

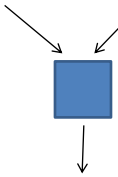
Learn Translation Knowledge from Non-Parallel Text?

English/Albanian Parallel text

Translation model

English text

Albanian text



Translation model

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

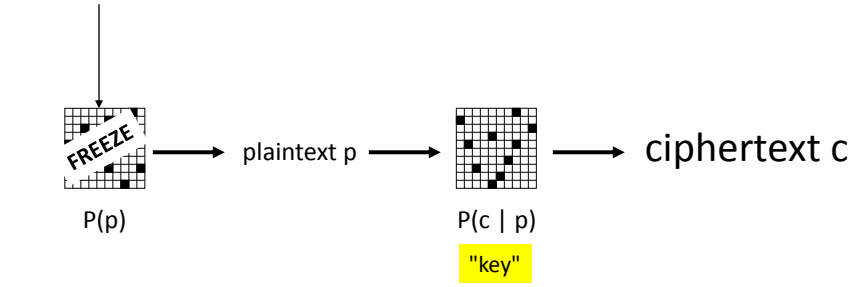
Automatic decipherment

Letter Substitution Cipher

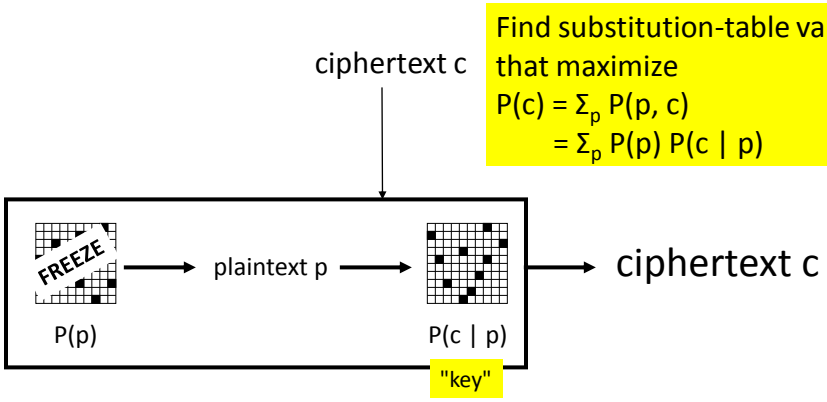
ciphertext c

Letter Substitution Cipher

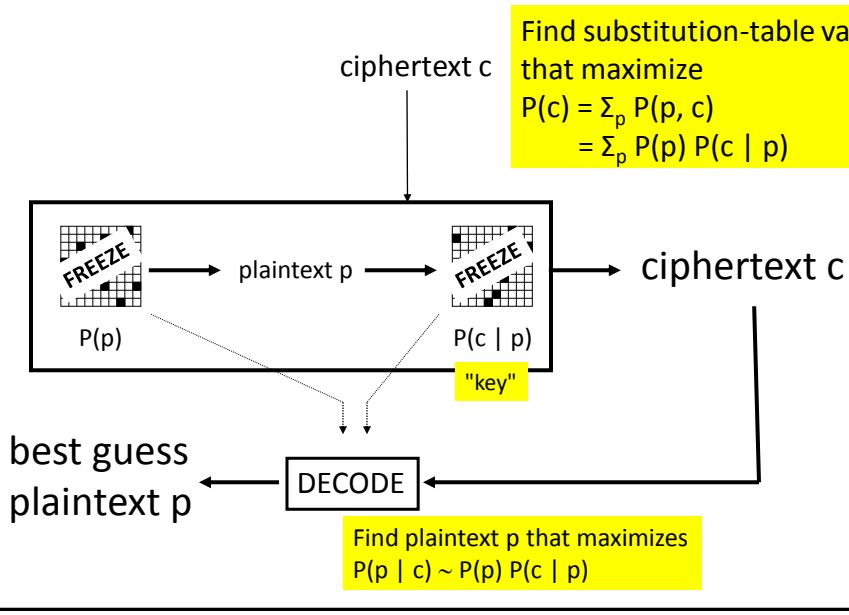
plaintext samples,
unrelated to ciphertext



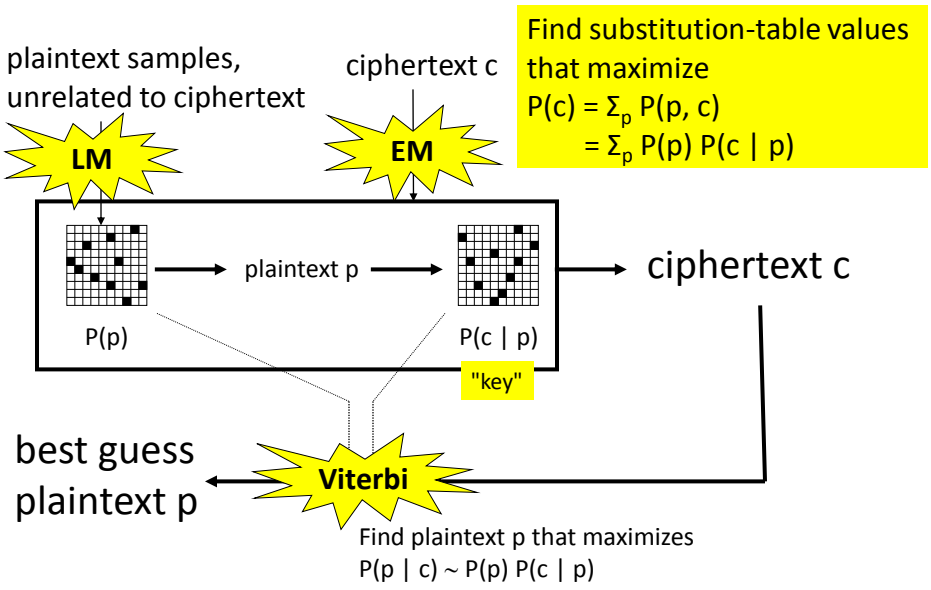
Letter Substitution Cipher



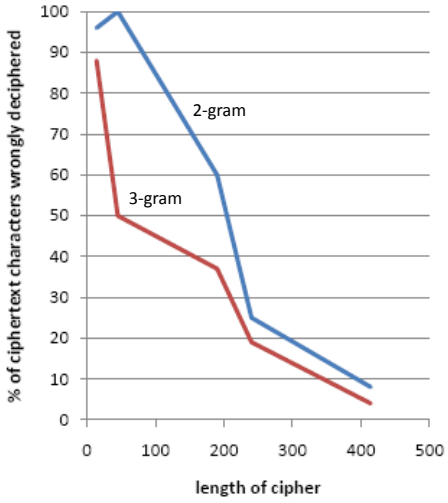
Letter Substitution Cipher



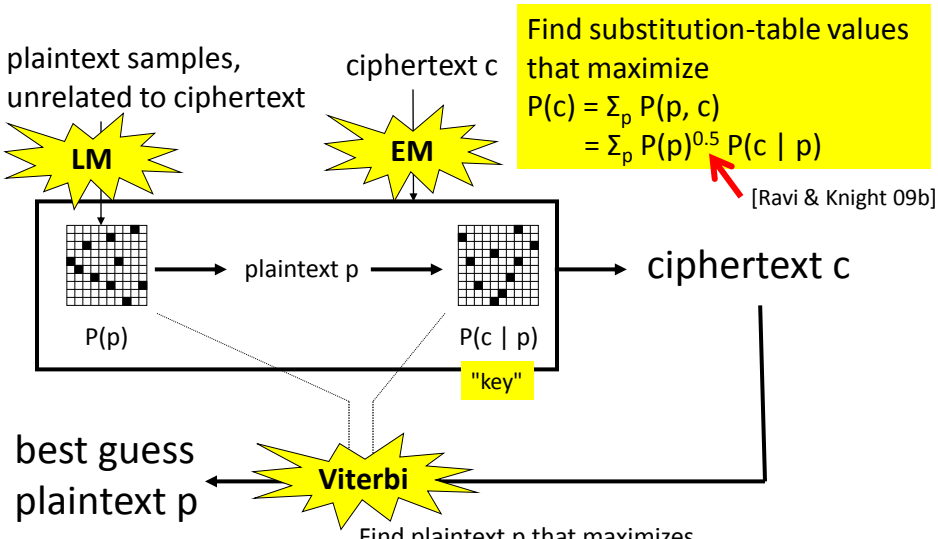
Letter Substitution Cipher



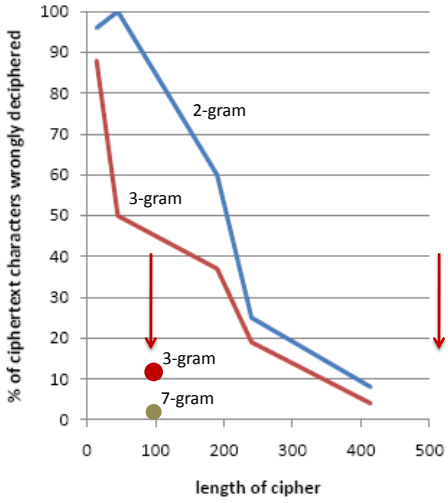
Decipherment Accuracy vs. Cipher Length



Letter Substitution Cipher



Reducing LM Weight During EM



Set EM to maximize

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

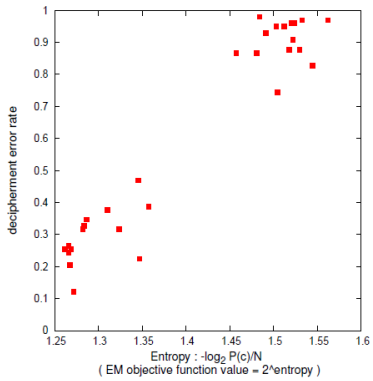
instead of

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

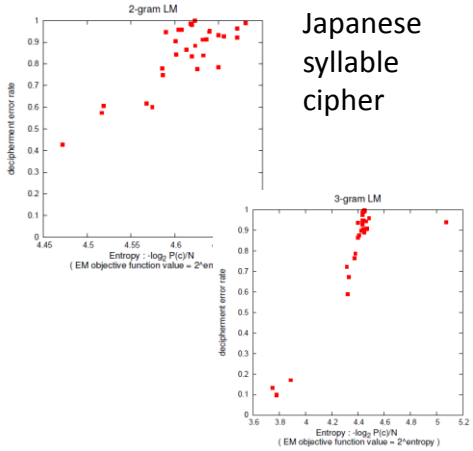
[Ravi & Knight 09b]

Random Restarts are Critical

English 98-letter cipher, 3-gram LM



Japanese syllable cipher

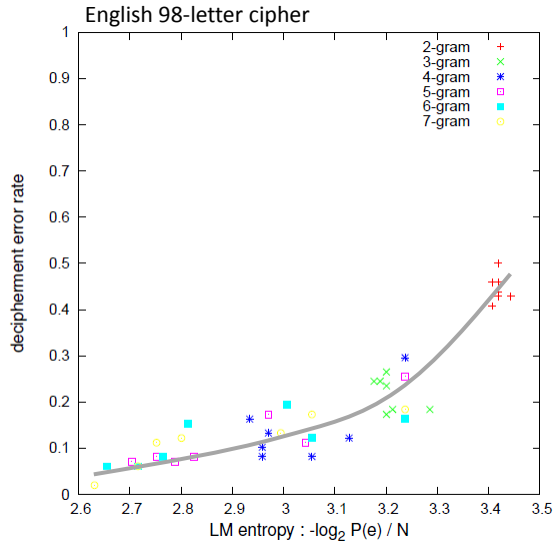


even people do restarts!

[Ravi & Knight 09b]

Good Language Models are Critical

English 98-letter cipher



[Ravi & Knight 09b]

Searching for Deterministic Keys

- Peleg & Rosenfeld, 1979
  - relaxation search

...

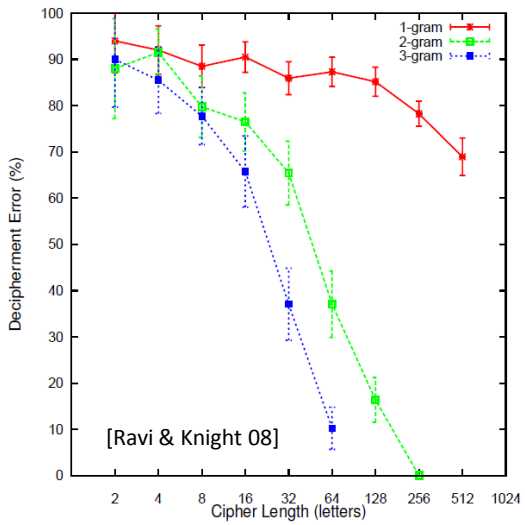
- Ravi & Knight, 2008
  - ILP, exact search
- Corlett & Penn, 2010
  - A* exact search
- Nuhn, Schamper, and Ney, 2013
  - beam search

Deterministic Keys

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

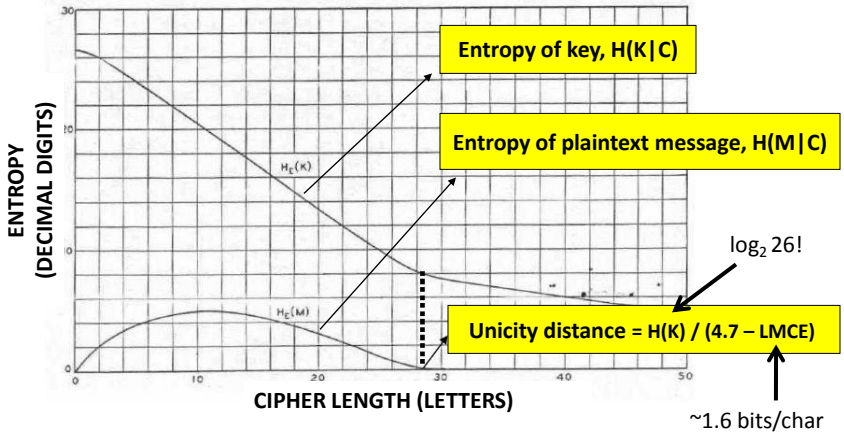
Cipher Length	EM error	ILP error
52	85 %	**21 %**
98	45 %	**12 %**
414	10 %	**0.5 %**

Using 2-gram letter-based LM

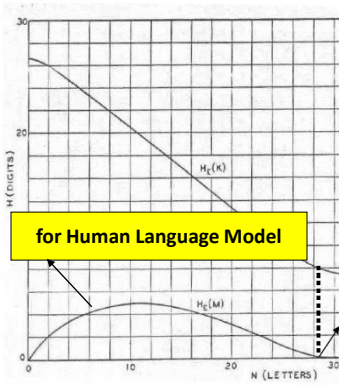


[Shannon 46, 49]
"Communication Theory of Secrecy Systems"

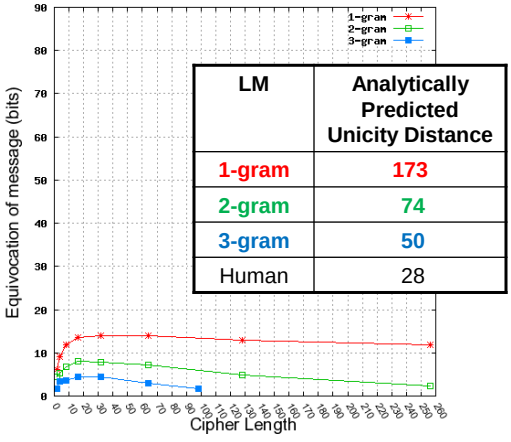
- Shannon analytically predicted uncertainty about key and message
- Graphed it for a human-level language model



Verifying Shannon's Prediction of Plaintext Message Uncertainty



ANALYTIC CURVES (Shannon)



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

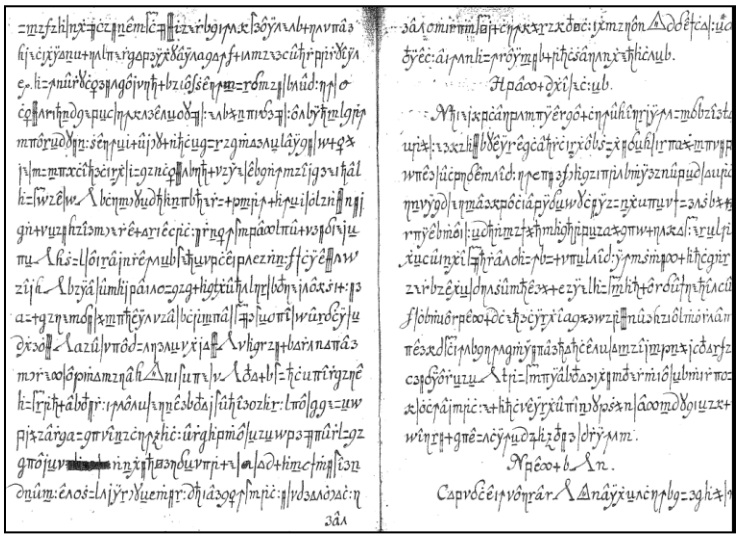
ACTUAL CURVES

[Ravi & Knight 08]

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



[Knight, Megyesi, Schaefer 11]

Copiale Cipher

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

Section headers

Lines ≈ equal length

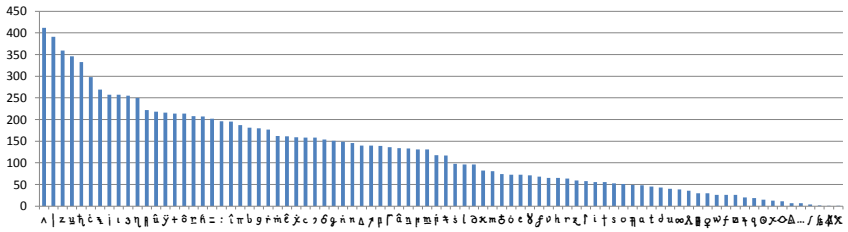
Paragraphs and section titles always begin with **capitalized Roman letters**.

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

Some scratch-outs, rare

Preview text fragments ("catchwords")

Letter Frequencies



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

tendencies:

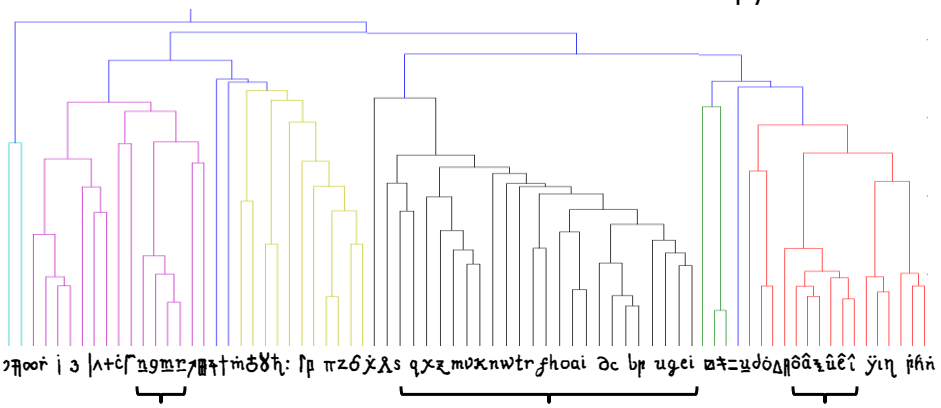
â, ê, î, ô, û followed by ȝ and ¡

â, ê, î, ô, û preceded by z and ᴨ

Clustering of Cipher Letters

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

Scipy software



underlined letters

unaccented Roman letters

circumflexed vowels

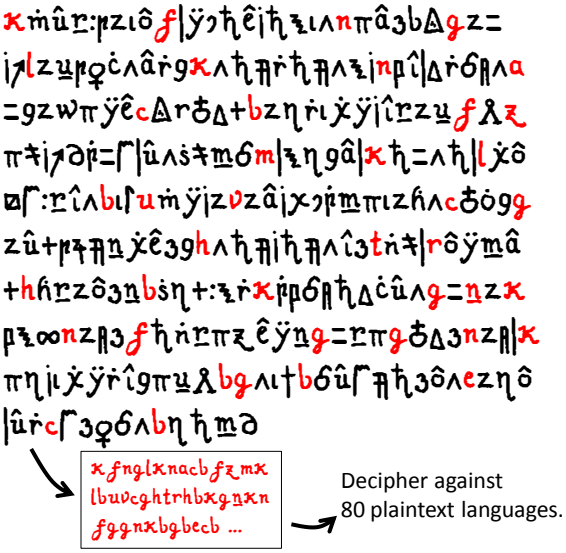
thanks Jon Graehl

First Decipherment Approach

unaccented Roman letters that cluster:

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

most common letter = 12%
least common = very small

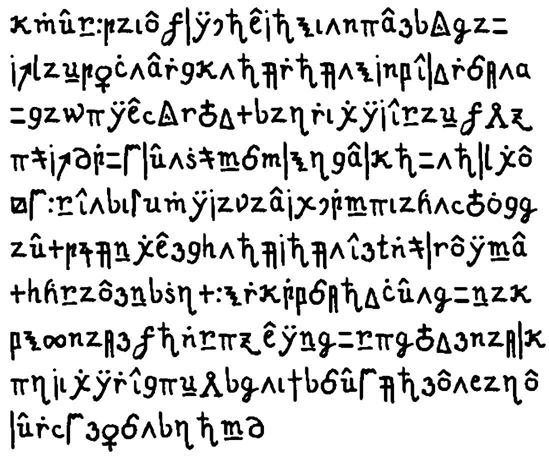


Decipher against 80 plaintext languages.

Second Decipherment Approach

Homophonic cipher, e.g.:

A = ȣ ¡ l ɥ ṙ
B = û
C = ȯ ṉ
D = ₳
E = ẋ ʄ ∆ ≠ ƒ î ₹ ȝ
F = ꝑ
G = ÿ
etc.



Homophonic Cipher

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

FAIL

But, slight numerical preference for German

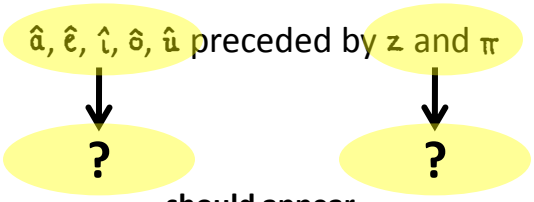
Cipher Characteristics

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

tendencies:

â, ê, î, ô, û followed by ȝ and ¡

â, ê, î, ô, û preceded by z and π



should appear adjacent in German text

Make full digraph table for cipher and for German

Key Observation #1

In Copiale, ɔ almost always followed by ɬ

In German, C almost always followed by H

(German CH is like English QU)

So guess: ɔ = C, ɬ = H

One Thing Leads to Another

ϽϦ = CH → ϽϦʌ = CHT → ʌ = T ?

Each step is guesswork.

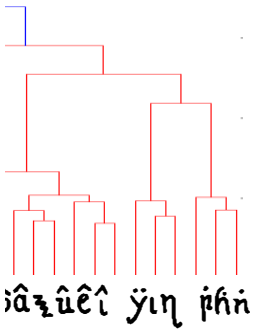
Must be willing to retract.

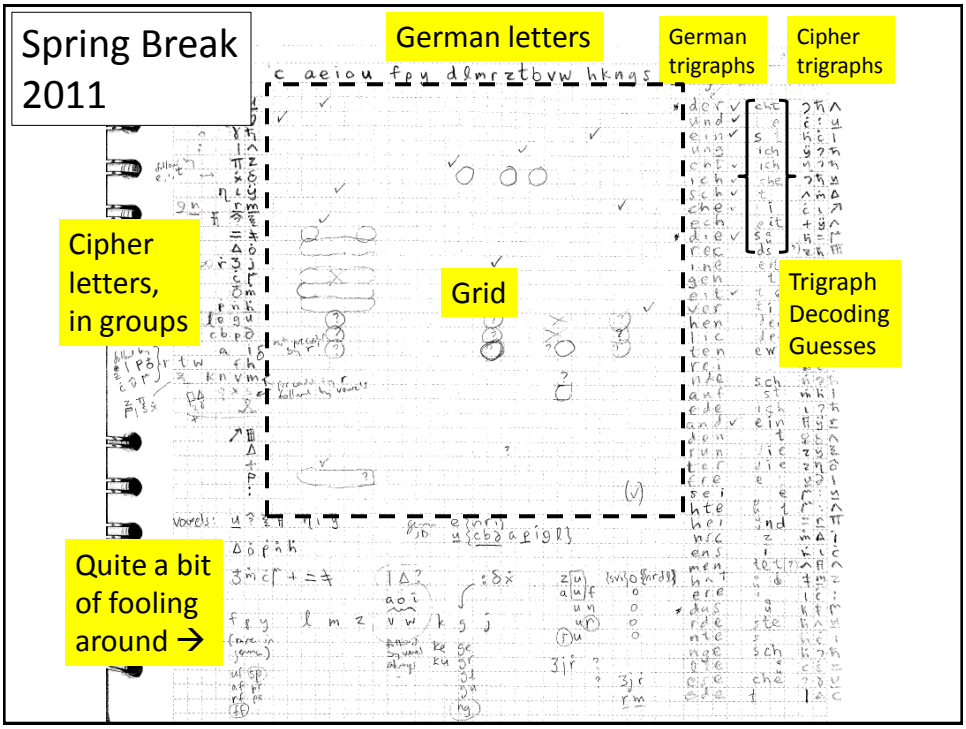
Weird task, not knowing German.

No longer care what the book says.

Cluster diagram crucial:

ÿ = I → ι = I , η = I

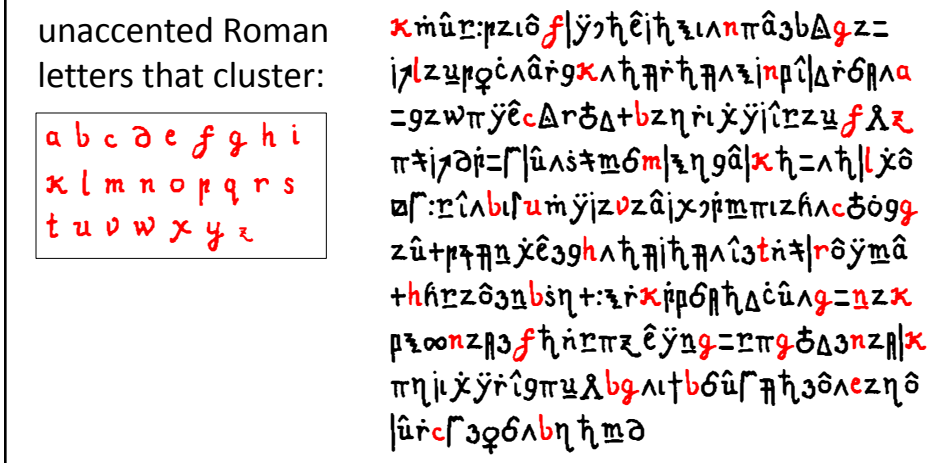




Key Observation #2

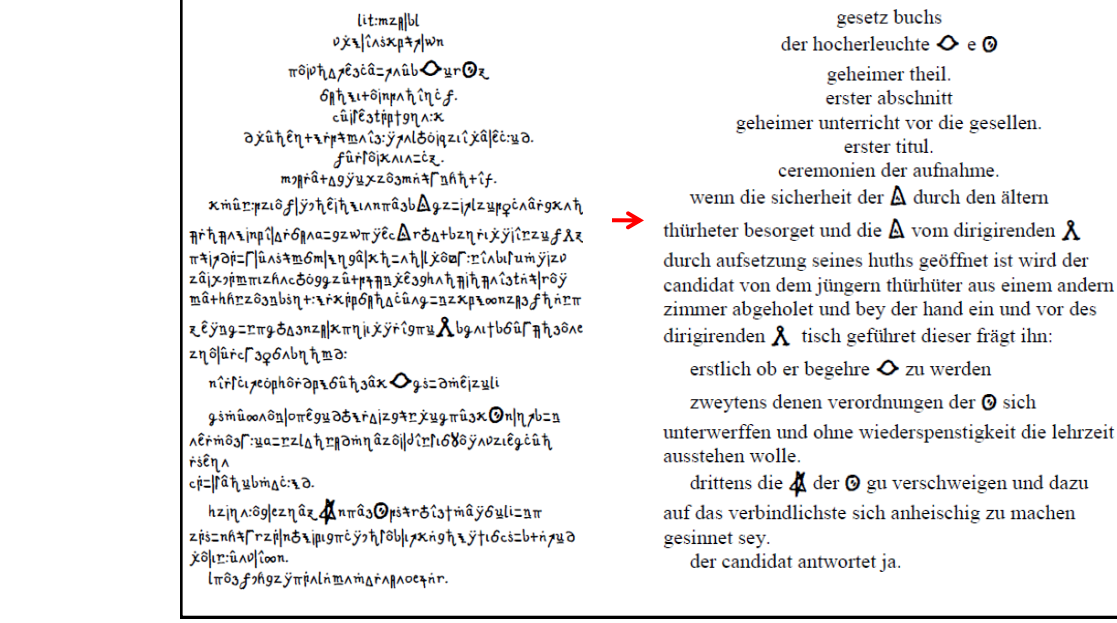
unaccented Roman letters that cluster:

a b c ə e ƒ g h i
ĸ l m n o ƥ q r s
t u ʋ w x y ƨ



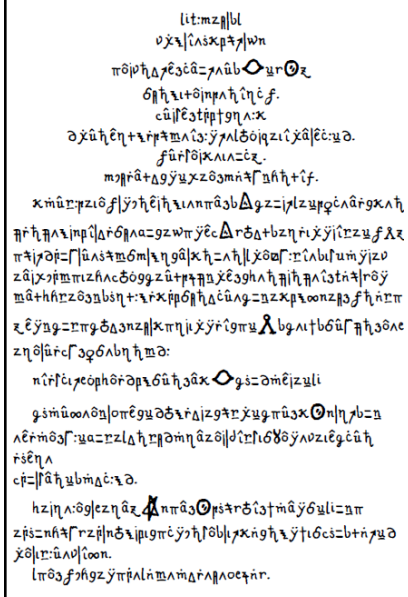
Actually, those are space bars

Copiale Decipherment



gesetz buchs
der hocherleuchte ◇ e ⊙
geheimer theil.
erster abschnitt
geheimer unterricht vor die gesellen.
erster titul.
ceremonien der aufnahme.
wenn die sicherheit der △ durch den ältern thürheter besorget und die △ vom dirigirenden ⅄ durch aufsetzung seines huths geöffnet ist wird der candidat von dem jüngern thürhüter aus einem andern zimmer abgeholet und bey der hand ein und vor des dirigirenden ⅄ tisch geführet dieser frägt ihn:
erstlich ob er begehre ◇ zu werden
zweytens denen verordnungen der ⊙ sich unterwerffen und ohne wiederspenstigkeit die lehrzeit ausstehen wolle.
drittens die ⅄ der ⊙ gu verschweigen und dazu auf das verbindlichste sich anheischig zu machen gesinnet sey.
der candidat antwortet ja.

Copiale Decipherment



First lawbook
of the ⌔ e ⓪

Secret part.
First section
Secret teachings for apprentices.
First title.
Initiation rite.

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

First, if he desires to become ⌔.

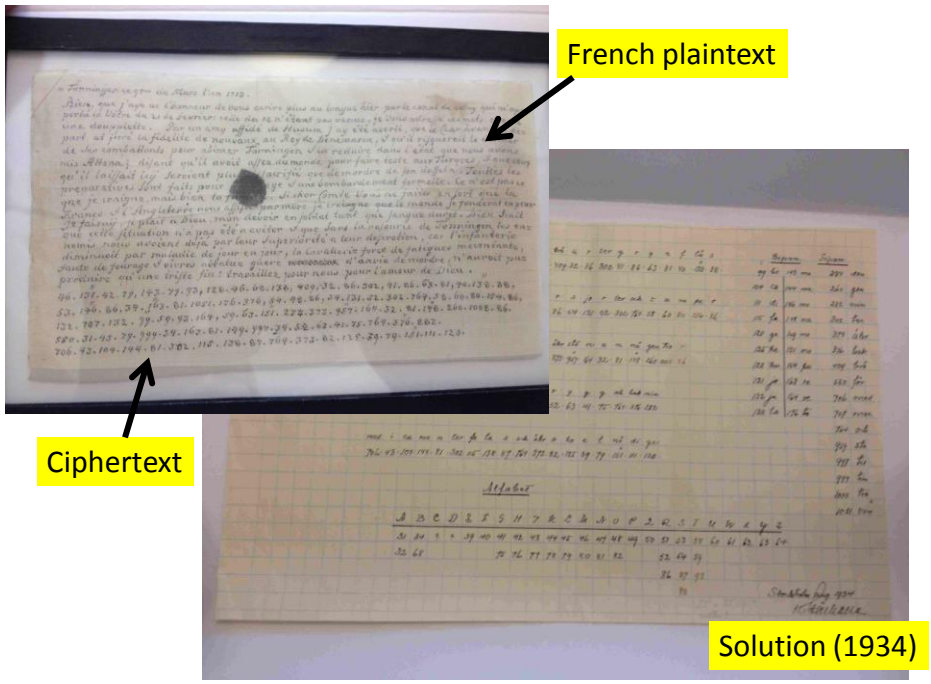
Secondly, if he submits to the rules of the ⓪ and without rebelliousness suffer through the time of apprenticeship.

Thirdly, be silent about the 🜏 of the ⓪ and furthermore be willing to offer himself to volunteer in the most committed way.

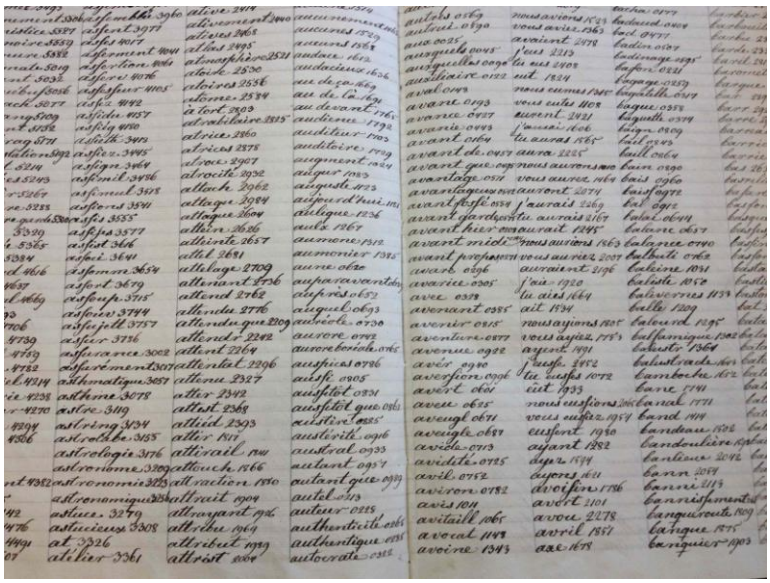
The candidate answers yes.

Historical Archives

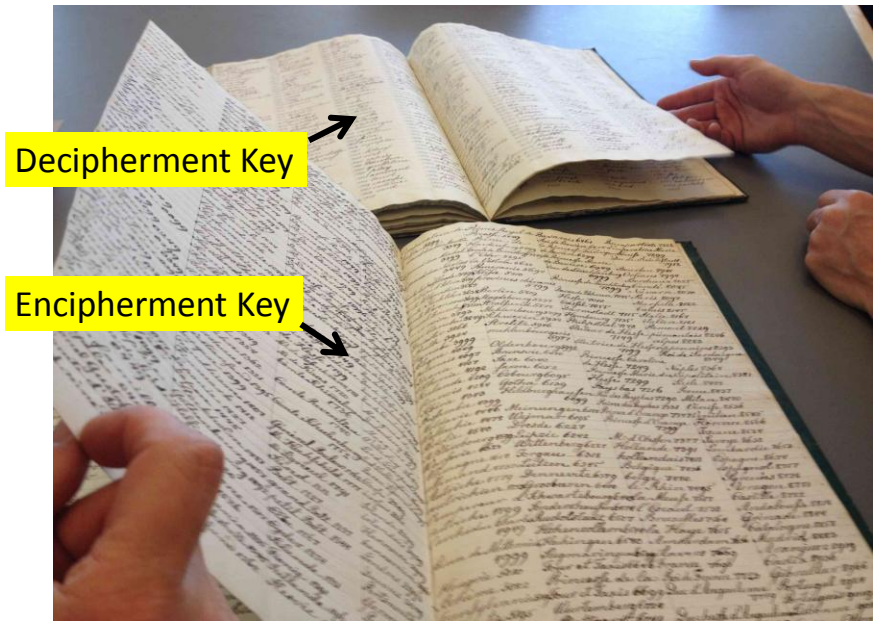
?



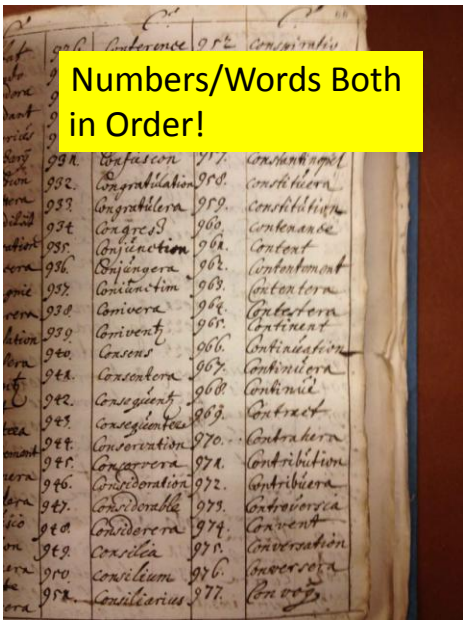
Word Substitution Encipherment Key

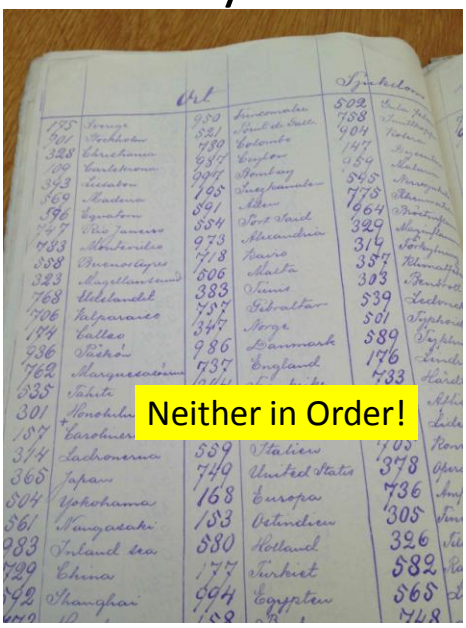


Word Substitution Keys



Word Substitution Keys

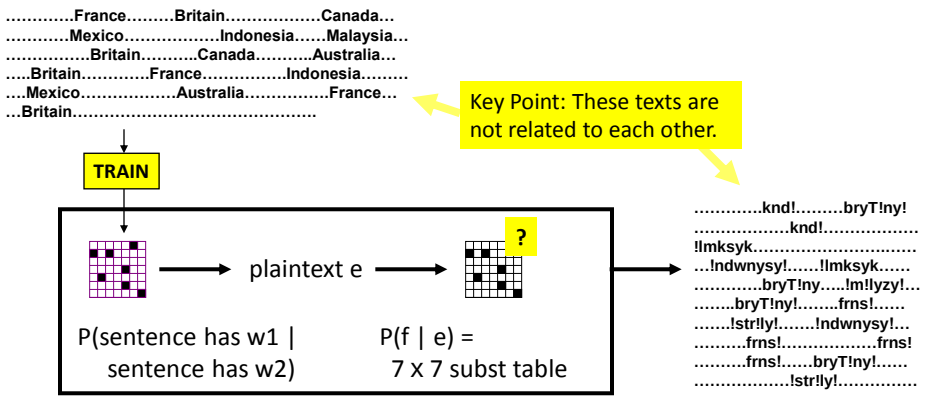




Word Substitution

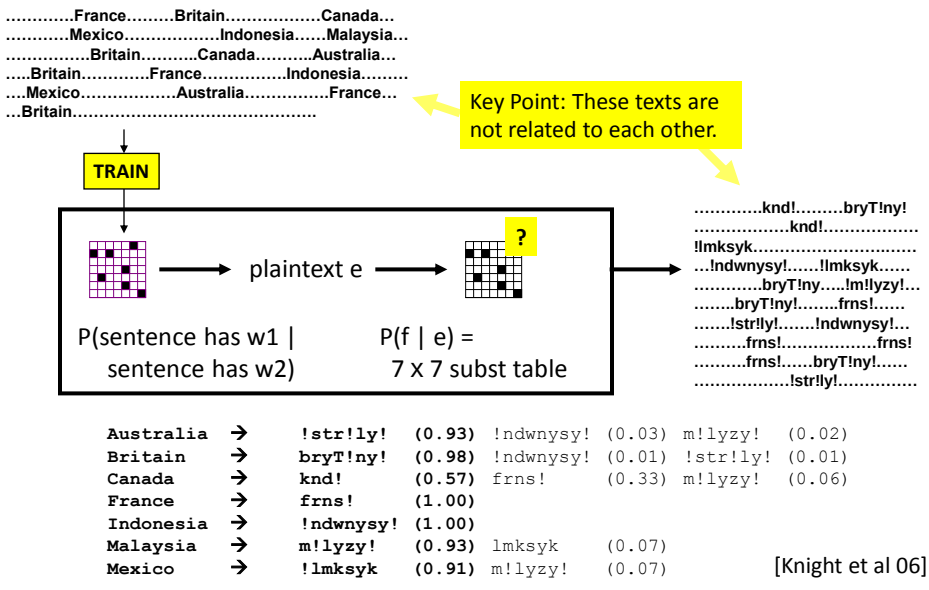
- Interesting for NLP
- Language translation can be viewed as word substitution (and transposition)
- Certainly, that is how IBM models 1-5 view it

Word Substitution (Small-scale)



[Knight et al 06]

Word Substitution (Small-scale)



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 x 100,000

Word Substitution (Giga-scale)

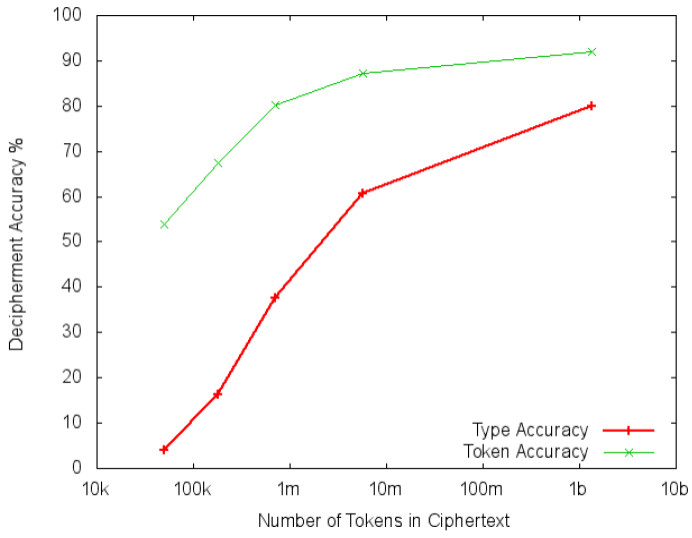
- Gibbs sampling fixes memory problem

```
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)

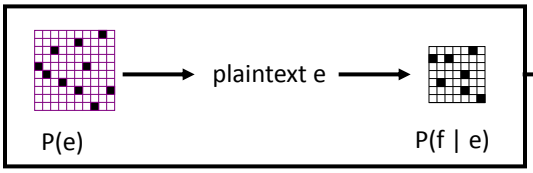


(Dou & Knight 2011)

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

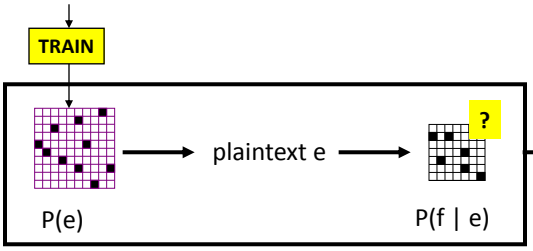


رفض رئيس السلطة الفلسطينية محمود عباس مجددا تصريحات وزير الخارجية الإسرائيلي سيلفان شالوم التي قال فيها إنه يتعين على إسرائيل إعادة النظر في انسحابها من غزة، المقرر أن يتم الصيف المقبل إذا فازت حركة المقاومة الإسلامية حماس في الانتخابات التشريعية. وقال عباس في مؤتمر صحفي على هامش مشاركته في القمة العربية-الثلاثينية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

Foreign Language as a Cipher

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

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



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

Time Expressions

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

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

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

Time Expressions

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

9 Hzyr!n 1942
8 tshryn !l!wl 1990
7 k!nwn !l!wl 1993
6 !'y!r 1993
6 !~Adh!r 1991
5 shb!T 1950
4 Hzyr!n 1989
30 !~Adh!r 1944
29 !y!r 1945
29 !~Adh!r 1993
28 k!nwn !l!'wl 1994

27 tmwz 1993
26 tmwz 1953
26 shb!T 1993
26 k!nwn !l!wl 1994
25 !ylwl 1926
24 !~Adh!r 1993
22 !ylwl 1957
22 tshryn !l!wl 1948
22 tmwz 1952
21 !y!r 1994
21 k!nwn !l!wl 1988

21 Hzyr!n 1967
20 !'y!r 1990
20 tshryn !'wl 1983
20 tshryn !l!'wl 1921
1 !y!r 1994
17 Hzyr!n 1972
16 !ylwl 1919
16 Hzyr!n 1984
16 !~Ab 1929

Time Expressions

<n> Hzyr!n <n>

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

Time Expressions

<n> Hzyr!n <n>

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

Time Expr

<n> Hzyr!n <n>

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

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

Time Expressions

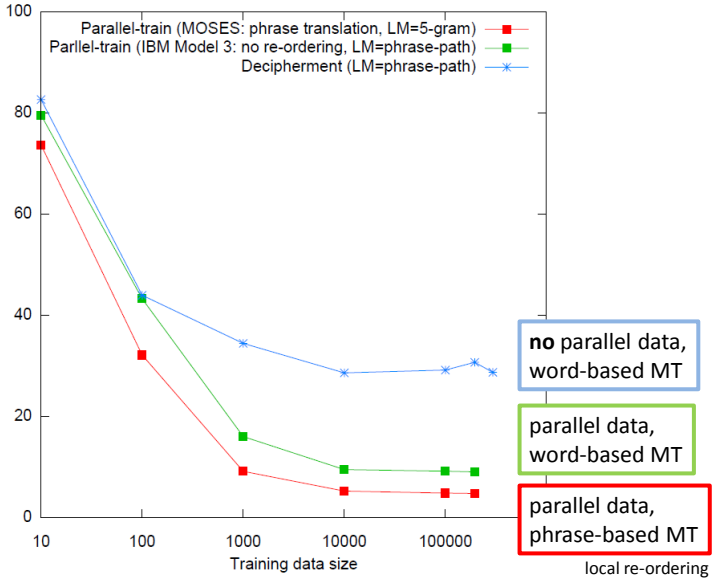
Hzyr!n

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

Deciphering Spanish Time Expressions

MT quality on test set

(Edit distance, lower is better)



[Ravi & Knight 11]

Deciphering Foreign Language at Giga-Scale

(Dou & Knight subm.)

Accuracy of learned bilingual dictionary

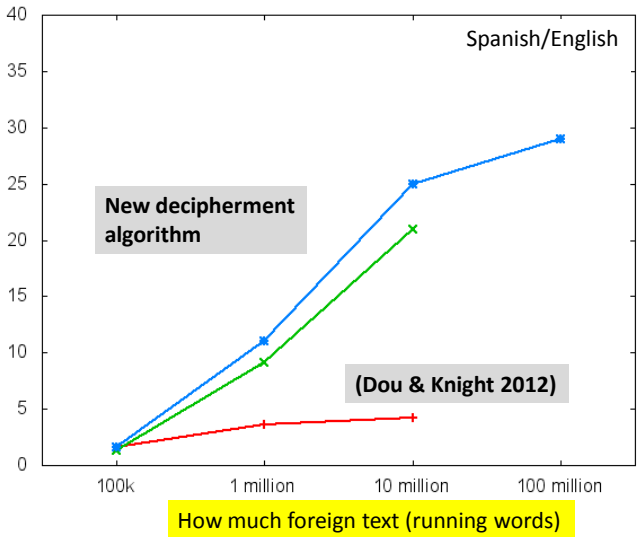
not translations of each other

English text

Foreign text

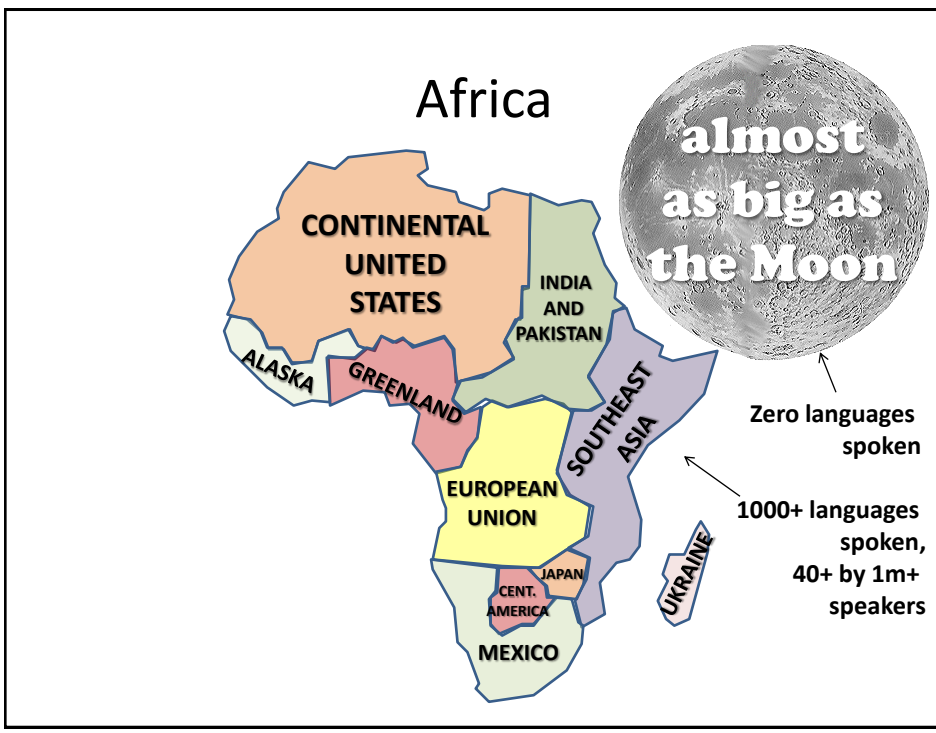
deciphering engine

bilingual word-for-word dictionary



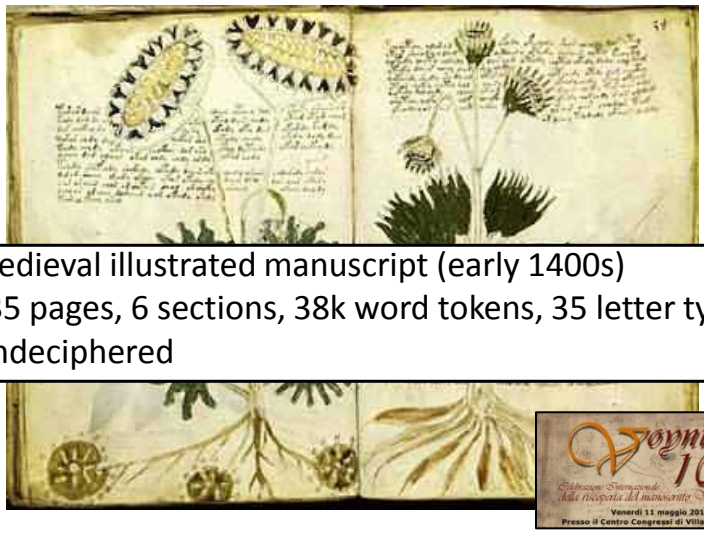
Practical Value

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



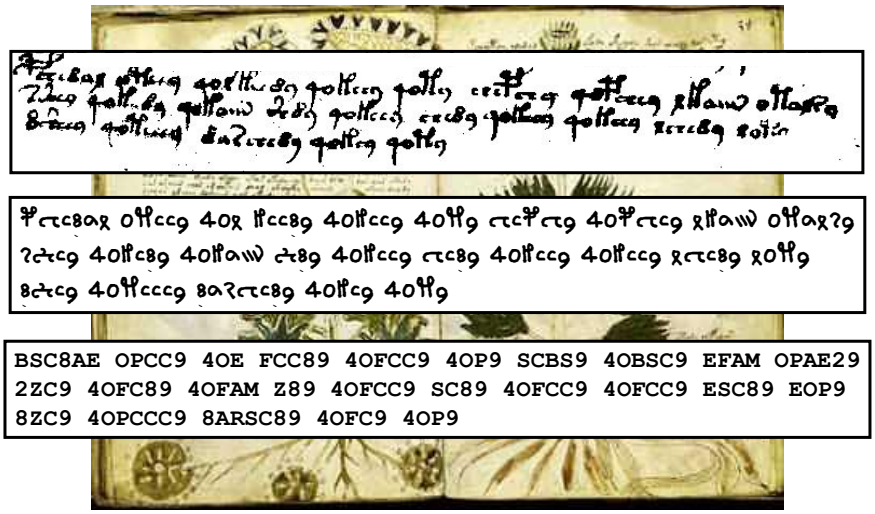
Unsolved ciphers

Voynich Manuscript (VMS)



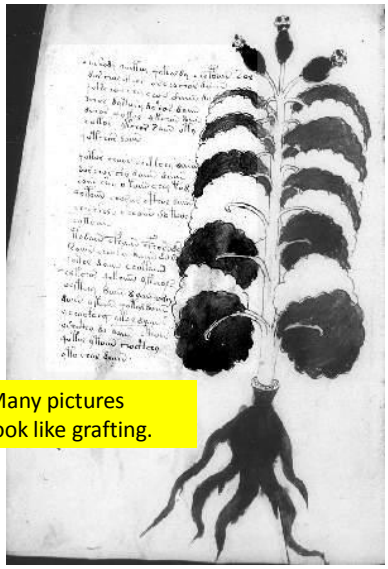
- Medieval illustrated manuscript (early 1400s)
- 235 pages, 6 sections, 38k word tokens, 35 letter types
- Undeciphered

Voynich Manuscript (VMS)



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

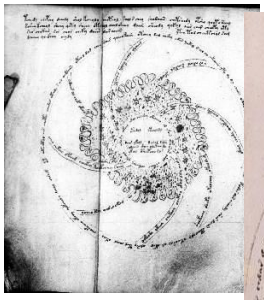
"Herbal" section

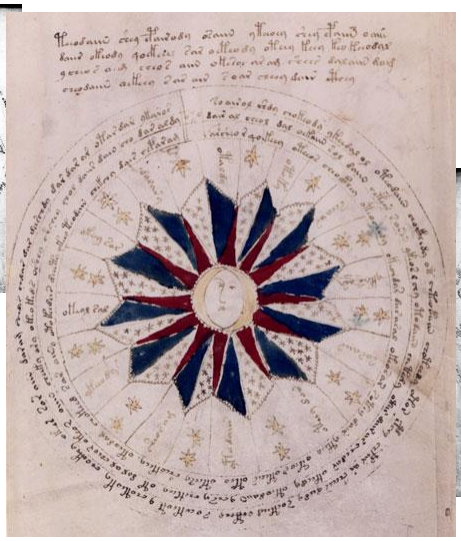


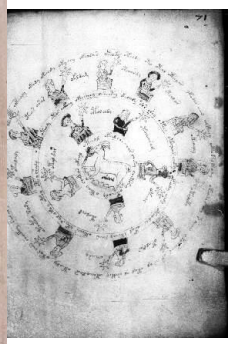
Many pictures look like grafting.

Sunflower? Would date VMS as post-1492.

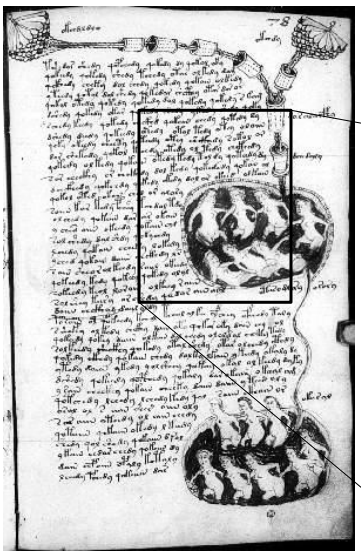
“Astrological” section







“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

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												N										
L																						
M													A									
N																						
O					N																	
P																						
Q									…								…					
R																						
S																						
T																						
U					…							N										
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

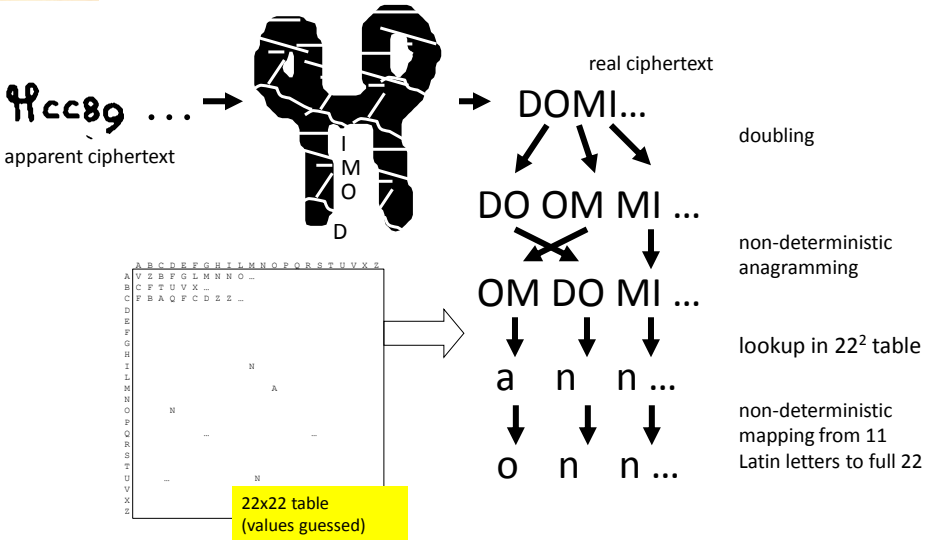
- Too hard to assemble good "cover" text!
- **So, make cipher letter-pairs overlap:**
  a n n … → AD DB BR … → ADBR …
- **Then, employ anagramming:**
  a n n … → OM DO MI … → DO OM MI … → DOMI …
- Now can construct a plausible looking "cover" text in Latin for our secret message (also in Latin)
- An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



Newbold's Results

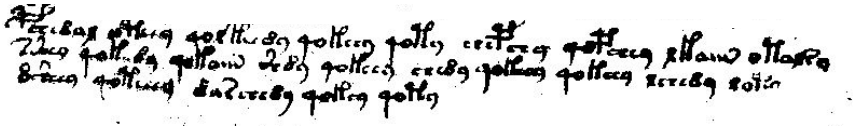
1300 real ciphertext "letters" in first 3 lines

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 … !

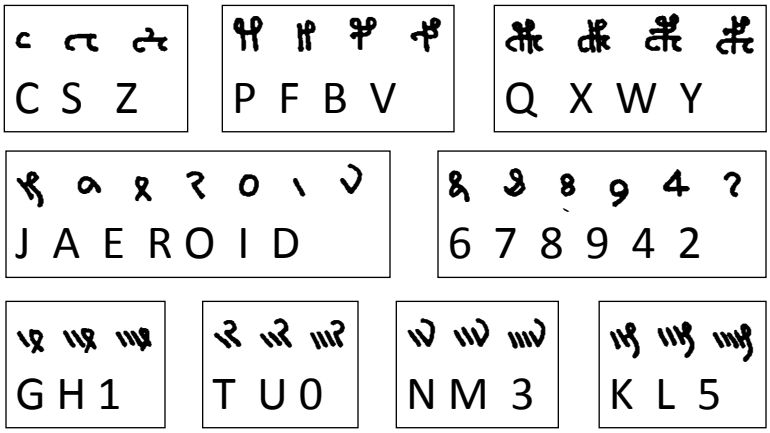
VMS Transcription



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



VMS Letters

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

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

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

Total
63k character tokens

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
  … 4OFCC89 4OFCC89 …
- 8 distinct words appear tripled
  … 4OFC89 4OFC89 4OFC89 …
  … SOR SOR SOR …
  … ZCOR ZCOR ZCOR …
  … OFAM OFAM OFAM …
  … OR OR OR …
  … 9FAM 9FAM 9FAM …
  … 8AM 8AM 8AM …
  … 4OFCC89 4OFCC89 4OFCC89 …

Substitution Cipher?

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

quiss squm is onum pom
quss 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 _ …

a →

b →

_ → _

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 _ …

a →

b →

_ → _

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

Sample tagging with learned model:

b b b b a _ a b b a _ b a _
V A S 9 2 _ 9 F A E _ A R _

b b b a _ b b a _ b b b a _ …
A P A M _ Z O E _ Z O R 9 _ …

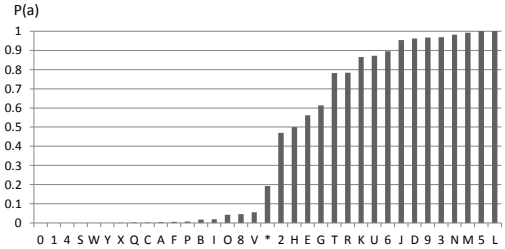
Letter Clustering

P(letter | tag)

P(tag | letter)

English

Voynich



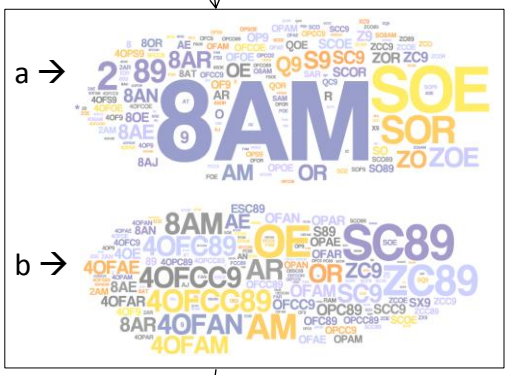
Word Clustering

Bigram model over {a, b}

↓

a a b a b a b a a …

↓



↓

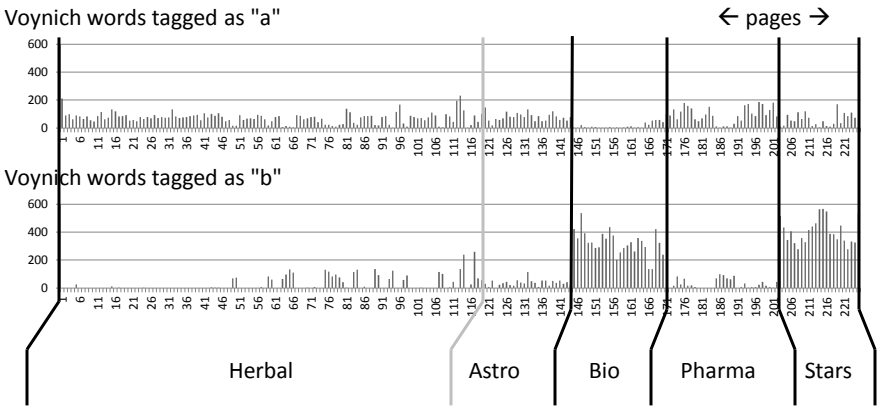
VAS92 9FAE AR APAM ZOE ZOR9 QRC2 9 ...

Sample tagging with learned model:

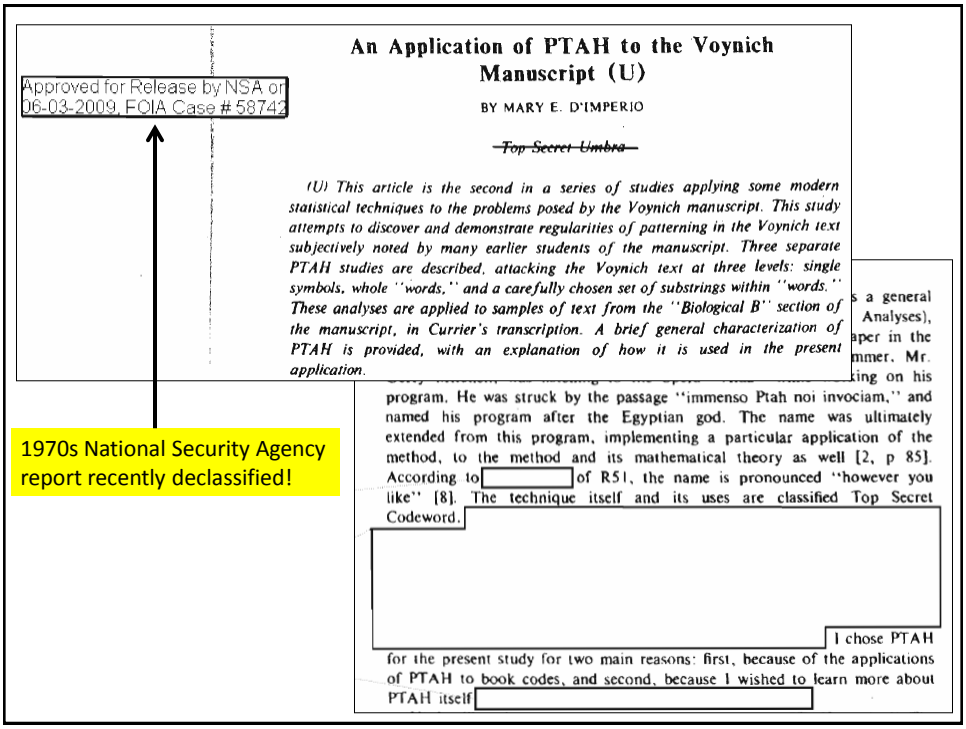
a a a a a
VAS92 9FAE AR APAM ZOE

a a a a a …
ZOR9 QRC2 9 FOR ZOE89 …

Word Clustering



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



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

An Application of PTAH to the Voynich Manuscript (U)

BY MARY E. D'IMPERIO

~~Top Secret Umbra~~

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

s a general
Analyses),
aper in the
mmer. Mr.
king on his

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

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

1970s National Security Agency report recently declassified!

National Security Agency

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

NSA employs more mathematicians and linguists than any other organization.

NSA has more computers than any other organization.

1970s paper on HMM Voynich

1950s
1960s
1970s
1980s
1990s
2000s
2010s
2020s

Association for Computational Linguistics

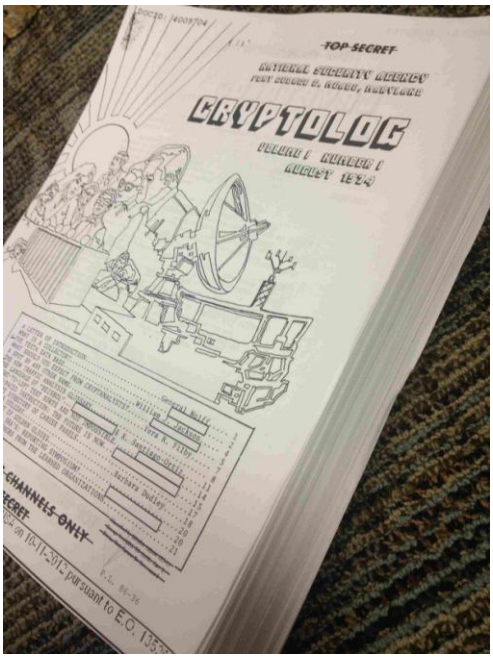
1993 paper on Statistical Machine Translation

2011 paper on HMM Voynich

ACL applies statistics to language processing problems

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

Slacker (1991)
dir. Richard Linklater



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 efforts 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.

~~(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

[redacted] P16

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 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 page[1]. 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

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

April 76 * CRYPTOLOG * Page 11

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 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.

[redacted]

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)

[redacted] P16

and

[redacted] P16

E.O. 1.4.(c)
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 to show that a PMT is not exactly what most people consider a true (or full) translation, but the quotes are inserted reluctantly. Although it may be difficult to read the

DOCID: 4010113 ~~TOP SECRET UMBRA~~

CRYPTOLOG P48 1985

MACHINE TRANSLATION:

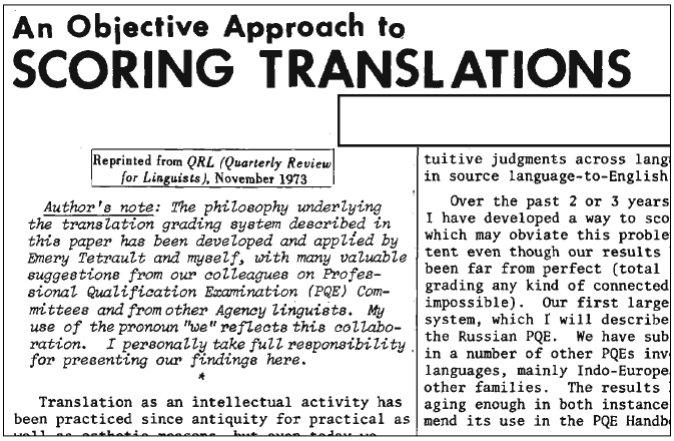
What can it do for us?

[redacted]

E.O. 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.

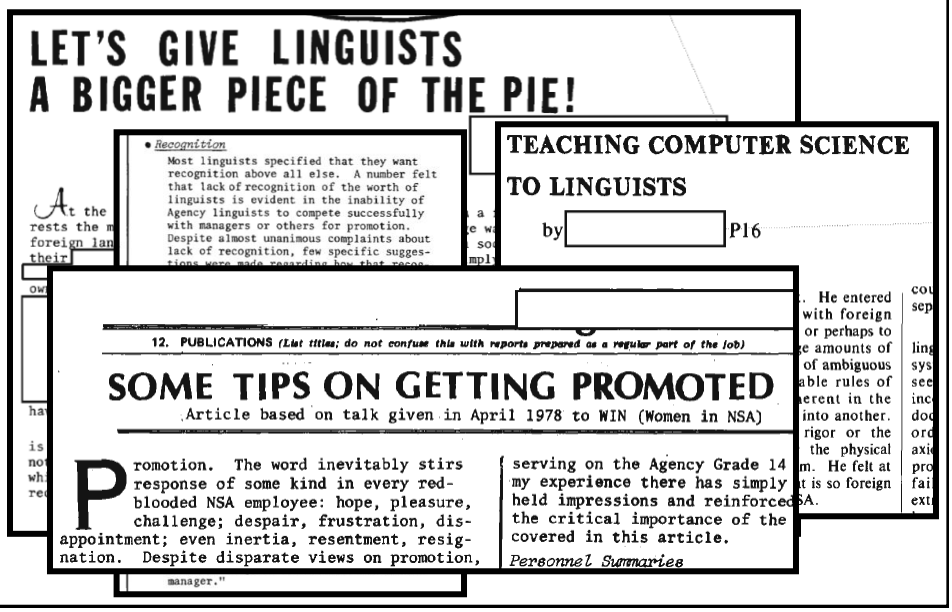
*

Translation as an intellectual activity has been practiced since antiquity for practical as

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 PQE. We have sub
in a number of other PQEs inv
languages, mainly Indo-Europe
other families. The results
aging enough in both instance
mend its use in the PQE 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 sugges-

manager."

TEACHING COMPUTER SCIENCE
TO LINGUISTS
by P16

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,

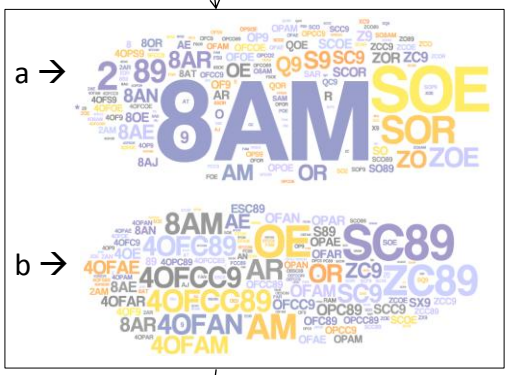
serving on the Agency Grade 14
my experience there has simply
held impressions and reinforced
the critical importance of the
covered in this article.

Personnel Summaries

Back to Word Clustering

Bigram model over {a, b}

a a b a b a b a a …

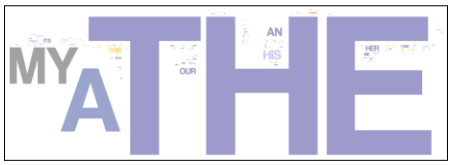


VAS92 9FAE AR APAM ZOE ZOR9 QRC2 9 ...

Let's try 10 clusters.

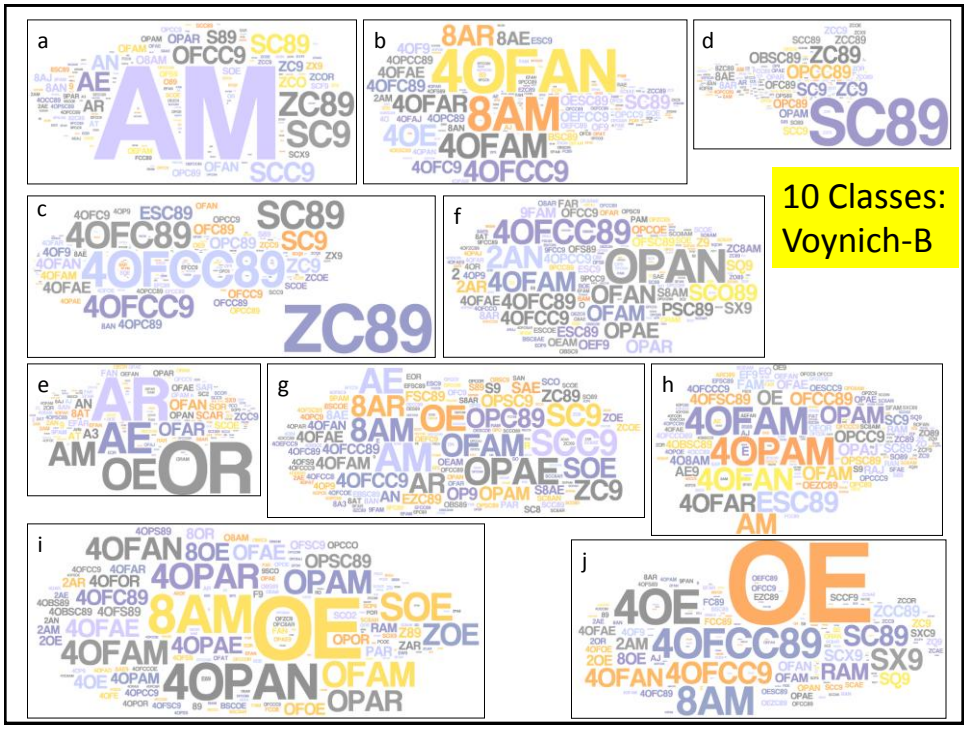
Let's limit ourselves to the more homogenous Bio + Stars sections.

10-Class Word Clustering: English



etc | etc | etc

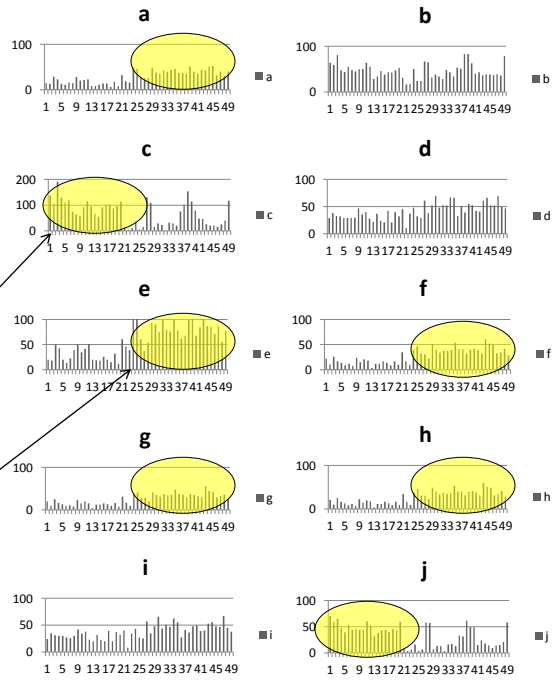
etc | etc | etc



10 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*

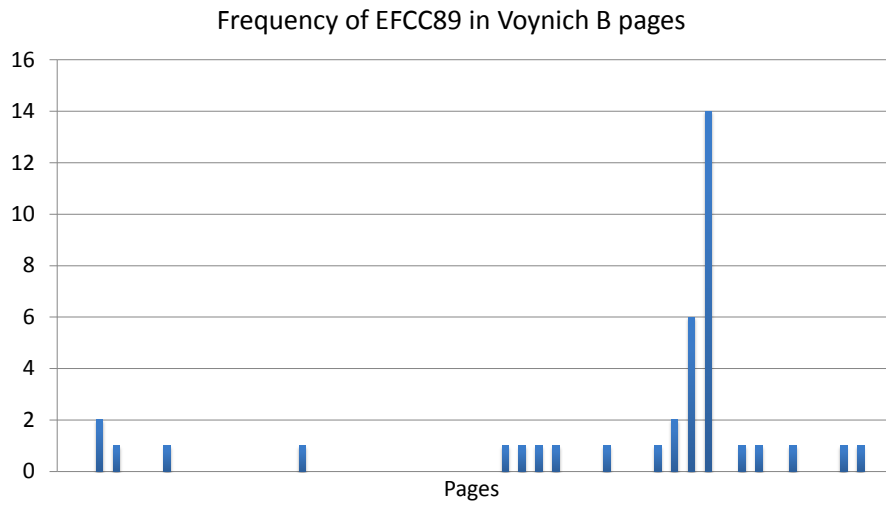
pages that contain word *w*

(Reddy & Knight, 2011)

Does VMS Have Content Words?

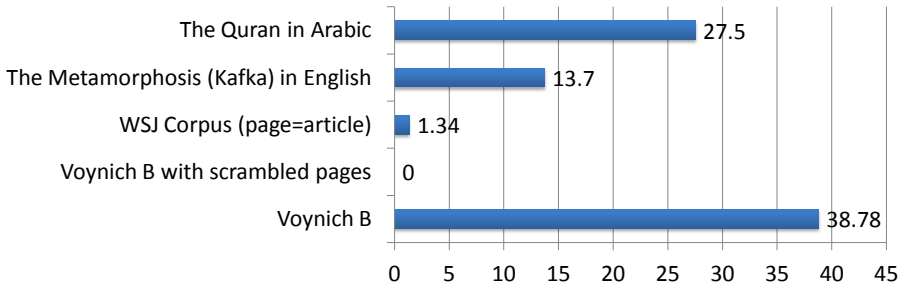
OFCC9 ZC89 8AN ZC9 SCFCC89 ROR 4OFAN SAE FAN ZCCF9 EPAN OR 9FCC9 EZC9 FAE 4OFCC89 FAR SC9 R AN ZCC89 4OFCC9 4OF9 FCCOR 8AR OFC9 SQ9 BSC89 SX9 EZC89 OFCC89 4OFC89 OESC89 2AE S89 OPC9 ESC89 4OFAE OFAR FCC89 2AN OPAR 2AM OPZC89 BSC8AR 8AE ZAE OPAN SAR AR ESC9 EOR AJ OEFCC9 SCQ9 ZX9 OPC89 8AT ZCQ9 E SFAE 4OFC9 SCF9 SQC89 EPC89 OFAE OP9 4OFCS89 OPCC89 4OPCC9 4OPCC89 OFC89 OPAJ 2ZC89 2 O 8AM EFAR 8E ES89 FC9 SCC89 8AR9 ZCOR 4OFCOR AM ZCC9 SE FCC9 4OPAN OFAN 4OPS89 OPCC9 SXC9 EFC89 4OFCCO AE SC8AN EFCC89 ZC8 SOE OEAN ZC8AE 4OFAM 9 SCOE EFC9 4OBSC9 OFS9 SCAE AEOE ZCAR ORAM 4OXC9 RAM OFC8AE SCQC9 4OPCC2 ZC8AJ OFCCC89 4OXC89 SCXC9 OPCOE 2AEFCC89 PCC89 SXC89 PCC9 ZCFCC9 ZCCOE ZCAE ON 4OFCAR SR 8AJ EAJ 8ZCC89 9SCC9 ZAR SQ*9 EOFCC89 ER SXAE 4OPSC89 ESCOE SC8AR ORAN 4OPSC9 AT OIF*9 4OSC9 OEFAE 8ZC9 PAR EFO EFCC9 ZCO AEOJ FCCOE 4OFCSC89 08AM ZCFAE 8SC8 4OVSC89 SO89 4OFCOE ESC8 SFCC9 8SCC9 EOFAJ 4OF OFCCOE SCBSC89 8OM ZCCX9 PC9 OPC8AR OPC2 4OFCS9 OPCO89 SCO89 4OFCO89 OPCC8 AK EVS9 SAM PAT CCC2 OJ EFAJ FCCC9 SCO2 8AK ZCCF EOP9 ZCFAN 4OFCCS9 OFCAE EFCCO89 ZCO8 ZC8AN BSAE OPCAE OPCCAJ 4OAN 4O8 SCCO SCO SCOFCC9 AEAJ OFCCO EFCSC9 EFCSC89 SCCFAN OPCCC9 EFAT ESAE OPCCOE SO FCSC89 OEFCCC89 4OFCO8 4OSC89 OPCO ZCCO8AR SCOJ OPARAE OAM OEFCCO EFCCOE EFAE EFCCC9 PC8AJ OFC8AN ZOF 4OFCCOE 4OFCCO2 OPCC8AN EFCCC89 SCAJ SCXC8 OSC9 FSOES8AR AIIB SCPAEZ9 4CCAE 4OFCCA2 SCOFC89 ZCCFZ9 SOFSC9 SX*9 9SCCO8AN 9FCC8AM RCCC9 OEA3 AIF*9 ZFAM 8ZCCO OPSC8C9 OPCOEAT 4OZCO 8SC8AR 9FCCCO2 BSC8C9 OPCO8O BS8AJ OFCO8AN ZCP ZCOPAJ ZPAR OESCO89 ZCQC9 ESCOCFAJ 4OFCCO8OR SCQ89 4OPC8E EOC89 SC8C9 EFAK O*OR F98CC89 ZCCP SCOP989 2ZCO PSC8 FCOQC89 4OFCZC9 FCCZO89 OFCSC89 ESR 2O AEAE O*AR 2OAM OPCO8AM 4OPCC8AM PCC8AN ZCOFAR RF9 SPAR 4OTAN ZCOPSC89 ZFC9 4OFCAN ZFCO89 8ZCCOPCC9 PCAR SOEFCCC89 ESCS89 4OCCCO SC8A FCCO8AE PCO OFCOJ 4OFCCO8 EFC8EFC9 PCC8 9SC8E BOEAE B9FCOR SCCV9 OBSC8AE EVSC89 ESCCOE OPCON SCAJAR 4OFCOFC89 OPCCOEFCC9 EAN 4OFCCAE 4OFCCE SC89PCOFAN EFC8AR EFCC8AN OFCAJ PCOEFC8AN EPCCAE U OFCCC2C9 OESAE 4CCAR BOCOFCC9 PC8AN SCBSAJ 4OFCCOFAN FCCE BOEFCCO SCOFCAN I*AR *AN 9ZC OFZ89 ZFCC9 SXAM

Do Content Words Indicate Topics?



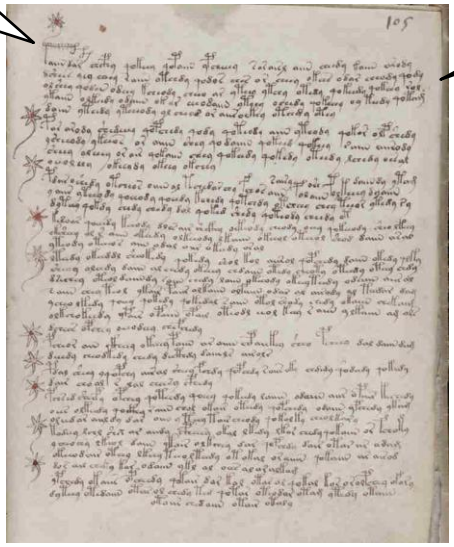
Are VMS Pages in Order?

- 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



Is VMS Prose?

Special ligatures at beginning of "paragraphs"



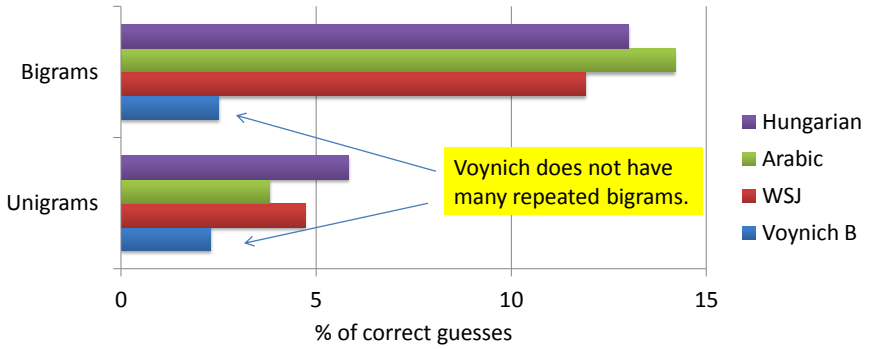
Looks like paragraph structure

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

The line is a functional entity…

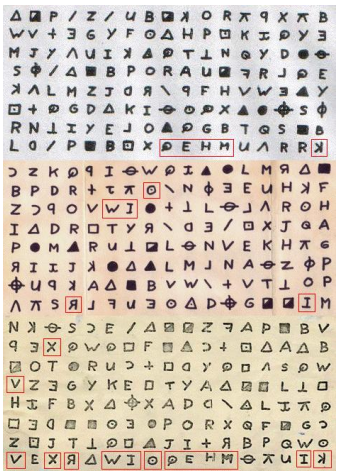
Are VMS Word Sequences Predictable?

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

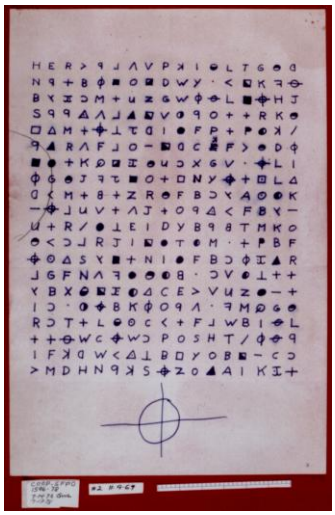


Zodiac Killer Ciphers

Zodiac 408 (solved, 1969)

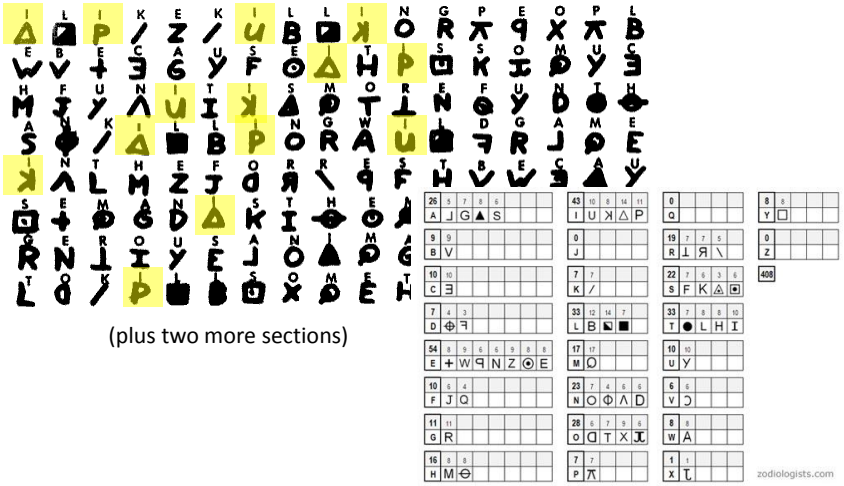


Zodiac 340 (still unsolved)



Zodiac Serial Killer

408-letter cipher (solved):



(plus two more sections)

Zodiac Serial Killer

Plaintext solution

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

Plaintext has many misspellings

Final 18 plaintext characters of 408 are "junk"

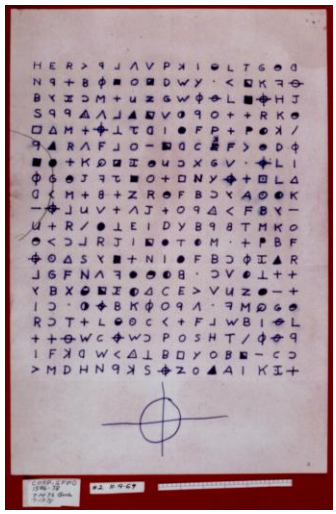
Deciphering Zodiac 408
Bayesian models & Gibbs sampling

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!
Interpolated 5-gram and word 1-gram	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



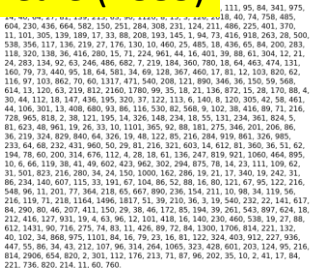
Has no obvious reading order bias:

% 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

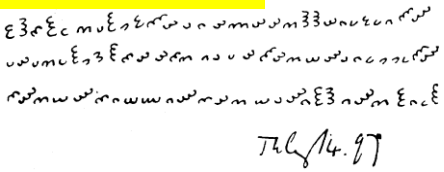
Could be nonsense … or maybe bigrams are smoothed out via more careful substitutions.

Other Unsolved Ciphers

Beale (1885)



Dorabella (1897)



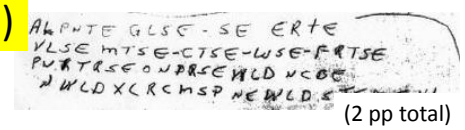
Kryptos (1990)

OBKR
UOXOGHULBSOLIFBBWFLRVQQPRNGKSSO
TWTQSJQSSEKZZWATJKLUDIAWINFB**NYP**
VTTMZFPKWGDKZXTJCDIGKUHUAUEKCAR

NYPVTT = BERLIN (2011 clue)

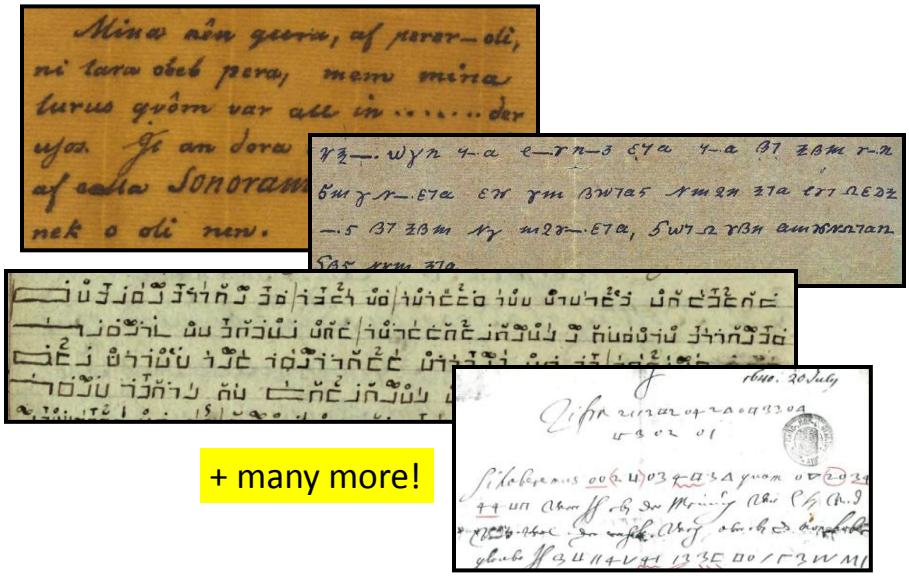
Taman Shud (1948)

FBI (1999)



(2 pp total)

Collected Ciphers



+ 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

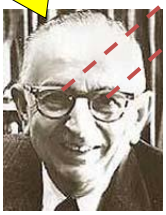
ciphertext

```
primera parte
del ingenioso
hidalgo don …
```

[Knight & Yamada 99]

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

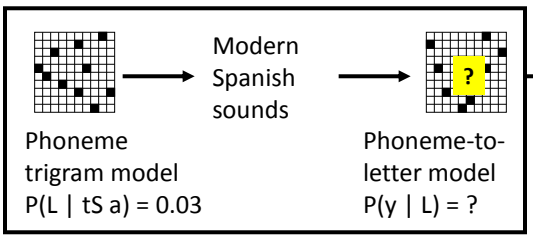
ciphertext

primera parte del ingenioso hidalgo don …

(Don Quixote)

[Knight & Yamada 99]

Archaeological Decipherment



ciphertext

primera parte del ingenioso hidalgo don …

(Don Quixote)

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

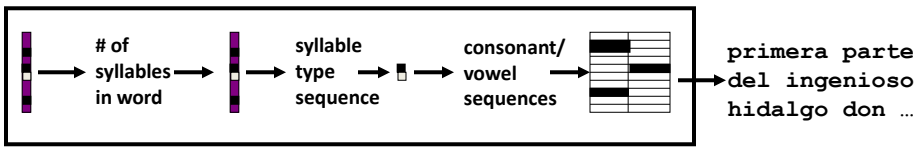
[Knight & Yamada 99]

What if Spoken Language Behind Script is Unknown?

- Build a universal model P(p) of human phoneme sequence production
  - human might generally say: K AH N AH R IY
  - human won't generally say: R T R K L K
- Find a P(c | p) table
  - such that there is a decoding with a good universal P(p) score
- Phoneme & syllable inventory
  - if z, then s
  - all have CV syllables; if VCC, then also VC
- Syllable sonority structure
  - dram, lomp, ? rdam, ? lopm
- Physiological preference constraints
  - tomp, tont, ? tomk, ? tonp

[Knight et al 06]

Unknown Source Language



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

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

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

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

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

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

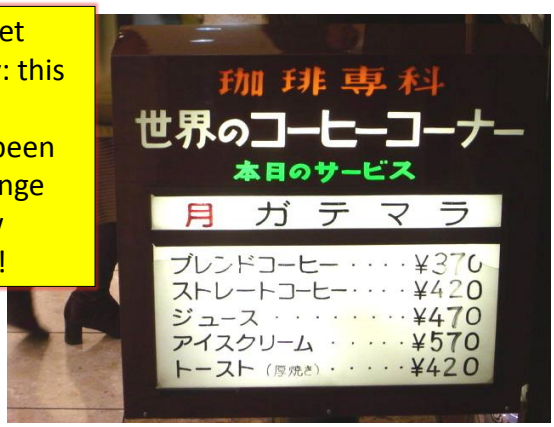
[Knight et al 06]

See 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)

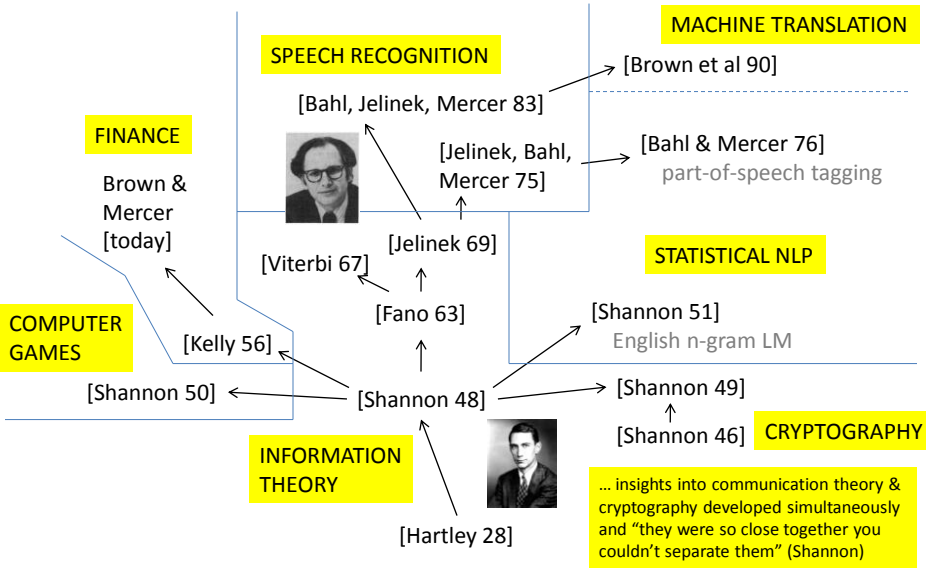
Linear A (1900BC)

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