Language Translation and Code-Breaking

Kevin Knight
Information Sciences Institute
University of Southern California

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

LXMLS Lisbon July 2013
美国关岛国际机场及其办公室均接获一名自称沙地阿拉伯富商拉登等发出的电子邮件，威胁将会向机场等公众地方发动生化袭击后，关岛经保持高度戒备。

Kowane mutum na da hakkin ya sami yancin yin tunani da na sanin yakamata da na bin addini; saboda haka yana da yancin sake addini ko ra'ayin da ya bada gaskiya gare shi, da kuma yancin nuna addininsa ko ra'ayinsa, shi daya ko a cikin taro kuma a fili ko a boye ta hanyar koyarwa ko yin ibada, ko bauta wa abin da ya bada gaskiya gare shi da yin abubuwan da abin da yake bauta wa din ya nuna masa.
The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport.

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

 Hmm, every time he sees “banco”, he either types “bank” or “bench” ... but if he sees “banco de...”, he always types “bank”, never “bench”...

 Translate, translate ...
<table>
<thead>
<tr>
<th>English</th>
<th>Spanish</th>
</tr>
</thead>
</table>
| 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 pequeños no son modernos . |
<table>
<thead>
<tr>
<th>Centauri/Arcturan</th>
<th>Arcturan</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a. ok-oon ororok sprok .</td>
<td>farok crrrok hihok yorok clok kantok ok-yurp</td>
</tr>
<tr>
<td>1b. at-oon bichat dat .</td>
<td></td>
</tr>
<tr>
<td>2a. ok-drubel ok-oon anok plok sprok .</td>
<td></td>
</tr>
<tr>
<td>2b. at-drubel at-oon pippat rrat dat .</td>
<td></td>
</tr>
<tr>
<td>3a. erok sprok izok hihok ghirok .</td>
<td></td>
</tr>
<tr>
<td>3b. totat dat arrat vat hilat .</td>
<td></td>
</tr>
<tr>
<td>4a. ok-oon anok drok brok jok .</td>
<td></td>
</tr>
<tr>
<td>4b. at-oon krat pippat sat lat .</td>
<td></td>
</tr>
<tr>
<td>5a. wiwok farok izok stok .</td>
<td></td>
</tr>
<tr>
<td>5b. totat jjat quat cat .</td>
<td></td>
</tr>
<tr>
<td>6a. lalok sprok izok jok stok .</td>
<td></td>
</tr>
<tr>
<td>6b. wat dat krat quat cat .</td>
<td></td>
</tr>
<tr>
<td>7a. lalok farok ororok lalok sprok izok enemok .</td>
<td></td>
</tr>
<tr>
<td>7b. wat jjat bichat wat dat vat eneat .</td>
<td></td>
</tr>
<tr>
<td>8a. lalok brok anok plok nok .</td>
<td></td>
</tr>
<tr>
<td>8b. iat lat pippat rrat nnat .</td>
<td></td>
</tr>
<tr>
<td>9a. wiwok nok izok kantok ok-yurp .</td>
<td></td>
</tr>
<tr>
<td>9b. totat nnat quat oloat at-yurp .</td>
<td></td>
</tr>
<tr>
<td>10a. lalok mok nok yorok ghirok clok .</td>
<td></td>
</tr>
<tr>
<td>10b. wat nnat gat mat bat hilat .</td>
<td></td>
</tr>
<tr>
<td>11a. lalok nok crrrok hihok yorok zanzanok .</td>
<td></td>
</tr>
<tr>
<td>11b. wat nnat arrat mat zanzanat .</td>
<td></td>
</tr>
<tr>
<td>12a. lalok rarok nok izok hihok mok .</td>
<td></td>
</tr>
<tr>
<td>12b. wat nnat forat arrat vat gat .</td>
<td></td>
</tr>
<tr>
<td>Centauri/Arcturan</td>
<td>Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp</td>
</tr>
<tr>
<td>-------------------</td>
<td>-----------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>1a. ok-voon ororok sprok .</td>
<td>7a. lalok [farok] ororok lalok sprok izok enemok .</td>
</tr>
<tr>
<td>1b. at-voon bichat dat .</td>
<td>7b. wat jjat bichat wat dat vat eneat .</td>
</tr>
<tr>
<td>2a. ok-drubel ok-voon anok plok sprok .</td>
<td>8a. lalok brok anok plok nok .</td>
</tr>
<tr>
<td>2b. at-drubel at-voon pippat rrat dat .</td>
<td>8b. iat lat pippat rrat nnat .</td>
</tr>
<tr>
<td>3a. erok sprok izok hihok ghirok .</td>
<td>9a. wiwok nok izok kantok ok-yurp .</td>
</tr>
<tr>
<td>3b. totat dat arrat vat hilat .</td>
<td>9b. totat nnat quat oloat at-yurp .</td>
</tr>
<tr>
<td>4a. ok-voon anok drok brok jok .</td>
<td>10a. lalok mok nok yorok ghirok clok .</td>
</tr>
<tr>
<td>4b. at-voon krat pippat sat lat .</td>
<td>10b. wat nnat gat mat bat hilat .</td>
</tr>
<tr>
<td>5a. wiwok [farok] izok stok .</td>
<td>11a. lalok nok crrrok hihok yorok zanzanok .</td>
</tr>
<tr>
<td>5b. totat jjat quat cat .</td>
<td>11b. wat nnat arrat mat zanzanat .</td>
</tr>
<tr>
<td>6a. lalok sprok izok jok stok .</td>
<td>12a. lalok rarok nok izok hihok mok .</td>
</tr>
<tr>
<td>6b. wat dat kratquat cat .</td>
<td>12b. wat nnat forat arrat vat gat .</td>
</tr>
</tbody>
</table>
Your assignment, translate this to Arcturan:

<table>
<thead>
<tr>
<th>Centauri/Arcturan</th>
<th>Arcturan</th>
</tr>
</thead>
<tbody>
<tr>
<td>Your assignment, translate this to Arcturan:</td>
<td>farok crrrok hihok yorok clok kantok ok-yurp</td>
</tr>
<tr>
<td>1a. ok-voon ororok sprok .</td>
<td>7a. lalok farok ororok lalok sprok izok enemok .</td>
</tr>
<tr>
<td>1b. at-voon bichat dat .</td>
<td>7b. wat jjat bichat wat dat vat eneat .</td>
</tr>
<tr>
<td>2a. ok-drubel ok-voon anok plok sprok .</td>
<td>8a. lalok brok anok plok nok .</td>
</tr>
<tr>
<td>2b. at-drubel at-voon pippat rrat dat .</td>
<td>8b. iat lat pippat rrat nnat .</td>
</tr>
<tr>
<td>3a. erok sprok izok hihok ghirok .</td>
<td>9a. wiwok nok izok kantok ok-yurp .</td>
</tr>
<tr>
<td>3b. totat dat arrat vat hilat .</td>
<td>9b. totat nnat quat oloat at-yurp .</td>
</tr>
<tr>
<td>4a. ok-voon anok drok brok jok .</td>
<td>10a. lalok mok nok yorok ghirok clok .</td>
</tr>
<tr>
<td>4b. at-voon krat pippat sat lat .</td>
<td>10b. wat nnat gat mat bat hilat .</td>
</tr>
<tr>
<td>5a. wiwok farok izok stok .</td>
<td>11a. lalok nok crrrok hihok yorok zanzanok .</td>
</tr>
<tr>
<td>5b. totat jjat quat cat .</td>
<td>11b. wat nnat arrat mat zanzanat .</td>
</tr>
<tr>
<td>6a. lalok sprok izok jok stok .</td>
<td>12a. lalok rarok nok izok hihok mok .</td>
</tr>
<tr>
<td>6b. wat dat krat quat cat .</td>
<td>12b. wat nnat forat arrat vat gat .</td>
</tr>
<tr>
<td>Centauri/Arcturan</td>
<td>Arcturan</td>
</tr>
<tr>
<td>-------------------</td>
<td>----------</td>
</tr>
<tr>
<td>Your assignment, translate this to Arcturan:</td>
<td>farok crrrok hihok yorok clok kantok ok-yurp</td>
</tr>
<tr>
<td>1a. ok-voon ororok sprok .</td>
<td>7a. lalok farok ororok lalok sprok izok enemok .</td>
</tr>
<tr>
<td>1b. at-voon bichat dat .</td>
<td>7b. wat jjat bichat wat dat vat eneat .</td>
</tr>
<tr>
<td>2a. ok-drubel ok-voon anok plok sprok .</td>
<td>8a. lalok brok anok plok nok .</td>
</tr>
<tr>
<td>2b. at-drubel at-voon pippat rrat dat .</td>
<td>8b. iat lat pippat rrat nnat .</td>
</tr>
<tr>
<td>3a. erok sprok izok hihok ghirok .</td>
<td>9a. wiwok nok izok kantok ok-yurp .</td>
</tr>
<tr>
<td>3b. totat dat arrat vat hilat .</td>
<td>9b. totat nnat quat oloat at-yurp .</td>
</tr>
<tr>
<td>4a. ok-voon anok drok brok jok .</td>
<td>10a. lalok mok nok yorok ghirok clok .</td>
</tr>
<tr>
<td>4b. at-voon krat pippat sat lat .</td>
<td>10b. wat nnat gat mat bat hilat .</td>
</tr>
<tr>
<td>5a. wiwok farok izok stok .</td>
<td>11a. lalok nok crrrok hihok yorok zanzanok .</td>
</tr>
<tr>
<td>5b. totat jjatquat cat .</td>
<td>11b. wat nnat arrat mat zanzanat .</td>
</tr>
<tr>
<td>6a. lalok sprok izok jok stok .</td>
<td>12a. lalok rarok nok izok hihok mok .</td>
</tr>
<tr>
<td>6b. wat dat krat quart cat .</td>
<td>12b. wat nnat forat arrat vat gat .</td>
</tr>
</tbody>
</table>
**Centauri/Arcturan**

*Your assignment, translate this to Arcturan:*

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

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

- Phrase-based MT
- Syntax-based MT
- Semantics-based MT
- Many ways to mix and match
Learn Translation Knowledge from Non-Parallel Text?

Parallel text

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

English text

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

Albanian text
Learn Translation Knowledge from Non-Parallel Text?

Can we automatically decipher Albanian into English?
“First NLP Task Ever” (1930s-40s)
Breaking the German Enigma Cipher

German text → over the radio:
DFKWIFKS LWORISJD KSUEIFKR ...

input (intercepted ciphertext): DFKWIFKS LWORISJD KSUEIFKR ...
output (plaintext): VASISTDASHERRCAPITANRICH ...

You can think of this like a multi-class tagging problem.
Each letter of ciphertext can get one of 26 tags (plaintext letters A...Z).
“First NLP Task Ever” (1930s-40s)
Breaking the German Enigma Cipher

Substitution system
N → J

Substitution table changes with every keystroke:
NNN → JTE

Flattens out ciphertext letter distributions.

Secret key = initial rotor ordering and settings

Reversible behavior
NNN → JTE → NNN
Breaking Enigma

You guys are wimps!

Statistical aligner

cipher text

substitution

guessed key

proposed plaintext

Is it German?

German

cipher text

plain text

SHANNON

TURING

KULLBACK
Breaking Enigma

cipher text

substitution

guessed key

proposed plaintext

Is it German?

German

SHANNON
TURING
KULLBACK
Breaking Enigma

- Cipher text
- Substitution
- Proposed plaintext
- Is it German?
- German
- Polished Cipher Bureau & Bletchley Park
- Bombe

Look for Enigma settings that yield plaintext pattern “XYZXYZ...”
Breaking Enigma

- Cipher text
- Substitution
- Guessed key
- Proposed plaintext
- Is it German?
- German

Polish Cipher Bureau & Bletchley Park

BOMBE

Look for Enigma settings that yield plaintext pattern “XYZXYZ…”

Winston Churchill
What’s the Plan for This Talk?

• Break a series of codes
  – Simple letter substitution
  – Phonetic substitution
    • archaeology
    • transliteration
  – Word substitution
  – Foreign language as cipher

• Bonus
  – Two historical ciphers
  – Final thought on translation and cryptography
What’s the Plan for This Talk?

• Break a series of codes
  – Simple letter substitution
  – Phonetic substitution
    • archaeology
    • transliteration
  – Word substitution
  – Foreign language as cipher

• Bonus
  – Two historical ciphers
  – Final thought on translation and cryptography
Letter Substitution Cipher

- Encipherment key:
  
  PLAIN:   ABCDEFGHIJKLMNOPQRSTUVWXYZ  
  CIPHER:  PLOKMIJNUHBYGVTFCDRXESZAQW

- 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 MDDBCYJQL: “TR HYD FKXC, FQ MKX RLQQIQ HYDL MKL DXCTW RDCDLQ JQMNKXTMB PTBMYEQL K FKH CY LQZKTL TC.”
KDCY LQZKTLJKX CY MDBCYJQL: “TR
HYD FKXC, FQ MKX RLQOIQ HYDL
MKL DXCTW RDCDLQ JQMNKXTMB
PTBMYEQL K FKH CY LQZKTL TC.”
KDCY LQZKTLJKX CY MDBCYJQL: “TR
HYD FKXC, FQ MKX RLQQIQ HYDL
MKL DXCTW RDCDLQ JQMNKXTMB
PTBMYEQL K FKH CY LQZKTL TC.”
KDCY LOZKTLJKX CY MDBCYJQL: “TR
HYD FKXC, FQ MKX RLQOIQ HYDL
MKL DXCTW RDCDLQ JQMNKXTMB
PTBMYEQL K FKH CY LOZKTL TC.”
KDCY LQZKTLJKX CY MDBCYJQL: “TR
HYD FKXC, FQ MKX RLQQIQ HYDL
MKL DXCTW RDCDLQ JQMNKXTMB
PTBMYEQL K FKH CY LQZKTL TC.”
KDCY LQZKTLJKX CY MDBCYJQL: “TR
HYD FKXC, FQ MKX RLQQIQ HYDL MKL DXCTW RDCDLQ JQMNKXTMB
PTBMYEQL K FKH CY LQZKTL TC.”

didn’t create “ae”
KDCY LQZKTLJXX CY MDBCYJQL: “TR
HYD FKXC, FQ MKX RLQQIQ HYDL
MKL DXCTW RDCDLQ JQMNKXTMB
PTBMYEQL K FKH CY LQZKTL TC.”
don’t like “ao” – back up!
auto repairman to customer: if
KDCY LQZKTLJKX CY MDBCYJQL: “TR
you wait we can freeze your
HYD FKXC, FQ MKX RLQQIQ HYDL
car until future mechanics
MKL DXCTW RDCDLQ JQMNKXTMB
discover a way to repair it
PTBMYEQL K FKH CY LQZKTL TC.”
Letter Substitution Cipher

plaintext samples, unrelated to ciphertext

plaintext $p$ → ciphertext $c$

$P(p)$ → $P(c \mid p)$

"key"
Letter Substitution Cipher

P(c | p)

P(p)

plaintext p

"key"

P(c | p)

ciphertext c
Letter Substitution Cipher

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

plaintext p

"key"
ciphertext c
Letter Substitution Cipher

Find substitution-table values that maximize
\[ P(c) = \sum_p P(p, c) = \sum_p P(p) P(c \mid p) \]

Find plaintext p that maximizes
\[ P(p \mid c) \sim P(p) P(c \mid p) \]
Letter Substitution Cipher

Find substitution-table values that maximize
\[ P(c) = \sum_p P(p, c) = \sum_p P(p) P(c \mid p) \]

Best guess plaintext \( p \) that maximizes
\[ P(p \mid c) \sim P(p) P(c \mid p) \]

Ciphertext \( c \)

Plaintext samples, unrelated to ciphertext

Viterbi

EM

LM

"key"
Decipherment Accuracy vs. Cipher Length

![Decipherment Accuracy vs. Cipher Length Graph](image)
Find substitution-table values that maximize
\[ P(c) = \sum_p P(p, c) = \sum_p P(p)^{0.5} P(c | p) \]

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

[Knight/Yamada 99]

[Ravi & Knight 09b]
Reducing LM Weight During EM

Set EM to maximize

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

instead of

\[ P(c) \approx \sum_p P(p) P(c \mid p) \]

[Ravi & Knight 09b]
Plan for This Talk

• Break a series of codes
  – Simple letter substitution
  – Phonetic substitution
    • archaeology
    • transliteration
  – Word substitution
  – Foreign language as cipher

• Bonus
  – Two historical ciphers
  – Final thought on translation and cryptography
Phonetic Decipherment

ciphertext

primera parte
del ingenioso
hidalgo don ...
Phonetic Decipherment

ciphertext

primera parte
del ingenioso
hidalgo don ...
“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

[Knight & Yamada 99]
Phonetic Decipherment

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

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

EM approach = 93% accurate phonetic decipherment

[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 $P(c \mid p)$ table
  – such that there is a decoding with a good universal $P(p)$ score

• Phoneme & syllable inventory
  – if z, then s
  – all have CV syllables; if VCC, then also VC

• Syllable sonority structure
  – dram, lomp, ? rdam, ? lopm

• Physiological preference constraints
  – tomp, tont, ? tomk, ? tonp

[Knight et al 06]
**Unknown Source Language**

Input: *primera parte del ingenioso don ...*

Output: **NSV.NV.NV** **NVS.NV** **NVS** **VS.NV.SV.V.NV** ... **S** = sonorous consonant phoneme **N** = non-sonorous consonant phoneme **V** = vowel phoneme

P(1) = ? P(CV) = ? P(V | V) = ? P(a | V) = ?
P(2) = ? P(V) = ? P(VV | V) = ? P(a | C) = ?

etc.

etc.

P(CVC) = ? + 7 others

[Knight et al 06]
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!

Parallel data: [Knight & Graehl 97]
Non-parallel data: [Ravi & Knight 09a]
Plan for This Talk

• Break a series of codes
  – Simple letter substitution
  – Phonetic substitution
    • archaeology
    • transliteration
  – Word substitution
  – Foreign language as cipher

• Bonus
  – Two historical ciphers
  – Final thought on translation and cryptography
Word Substitution Cipher

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

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

[Knight et al 06]
Word Substitution Cipher

\[ P(f \mid e) = \begin{bmatrix} 7 \times 7 \text{ subst table} \end{bmatrix} \]

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

TRAIN

\[
P(\text{sentence has } w_1 \mid \text{sentence has } w_2) \quad P(f \mid e) =
\]

<table>
<thead>
<tr>
<th>Country</th>
<th>Substitution 1</th>
<th>Probability 1</th>
<th>Substitution 2</th>
<th>Probability 2</th>
<th>Substitution 3</th>
<th>Probability 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>!str!ly!</td>
<td>0.93</td>
<td>!ndwnysy!</td>
<td>0.03</td>
<td>m!lyzy!</td>
<td>0.02</td>
</tr>
<tr>
<td>Britain</td>
<td>bryT!ny!</td>
<td>0.98</td>
<td>!ndwnysy!</td>
<td>0.01</td>
<td>!str!ly!</td>
<td>0.01</td>
</tr>
<tr>
<td>Canada</td>
<td>knd!</td>
<td>0.57</td>
<td>frns!</td>
<td>0.33</td>
<td>m!lyzy!</td>
<td>0.06</td>
</tr>
<tr>
<td>France</td>
<td>frns!</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Indonesia</td>
<td>!ndwnysy!</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Malaysia</td>
<td>m!lyzy!</td>
<td>0.93</td>
<td>lmksyk</td>
<td>0.07</td>
<td>m!lyzy!</td>
<td>0.07</td>
</tr>
<tr>
<td>Mexico</td>
<td>!lmksyk</td>
<td>0.91</td>
<td>m!lyzy!</td>
<td>0.07</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[Knight et al 06]
Word Substitution
(Full Vocabulary, Giga-scale)

(Dou & Knight 2011)
“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!!!”
Time Expressions

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

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

@@Sr
=hdh! !lsh=hr
fy ywm
nys!n
!!sbw@
=hdh=h !!!'y!m
qbl !'y!m
fy !@Sr
mn !lsn&
!!snw!t
b@d ywm
!!!y!m
13 nys!n 1994
lth!ny& !@shr&
thl!th& !y!m
qbl !sbw@yn
qbl !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
8 tshryn !!!wl 1990
7 k!nwn !!!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 !!!'wl 1994

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

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

<table>
<thead>
<tr>
<th>Number</th>
<th>Expression</th>
<th>Year(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>4 Hzyr!n</td>
<td>1967</td>
</tr>
<tr>
<td>12</td>
<td>fy 12 Hzyr!n</td>
<td>1993</td>
</tr>
<tr>
<td>7</td>
<td>5 Hzyr!n</td>
<td>1967</td>
</tr>
<tr>
<td>6</td>
<td>fy 30 Hzyr!n</td>
<td>1989</td>
</tr>
<tr>
<td>6</td>
<td>30 Hzyr!n</td>
<td>1989</td>
</tr>
<tr>
<td>4</td>
<td>fy 30 Hzyr!n</td>
<td>1994</td>
</tr>
<tr>
<td>4</td>
<td>fy 30 Hzyr!n</td>
<td>1993</td>
</tr>
<tr>
<td>3</td>
<td>fy 19 Hzyr!n</td>
<td>1967</td>
</tr>
<tr>
<td>2</td>
<td>ywm 30 Hzyr!n</td>
<td>1989</td>
</tr>
<tr>
<td>2</td>
<td>w 6 Hzyr!n</td>
<td>1994</td>
</tr>
<tr>
<td>2</td>
<td>qbl 5 Hzyr!n</td>
<td>1967</td>
</tr>
<tr>
<td>2</td>
<td>fy 9 Hzyr!n</td>
<td>1967</td>
</tr>
<tr>
<td>2</td>
<td>fy 7 Hzyr!n</td>
<td>1981</td>
</tr>
<tr>
<td>2</td>
<td>fy 6 Hzyr!n</td>
<td>1994</td>
</tr>
<tr>
<td>2</td>
<td>fy 5 Hzyr!n</td>
<td>1967</td>
</tr>
</tbody>
</table>
Time Expressions

<table>
<thead>
<tr>
<th>Search query</th>
<th>Documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>January 4, 1967</td>
<td>8040</td>
</tr>
<tr>
<td>February 4, 1967</td>
<td>9270</td>
</tr>
<tr>
<td>March 4, 1967</td>
<td>10700</td>
</tr>
<tr>
<td>April 4, 1967</td>
<td>21800</td>
</tr>
<tr>
<td>May 4, 1967</td>
<td>14000</td>
</tr>
<tr>
<td>June 4, 1967</td>
<td>39300</td>
</tr>
<tr>
<td>July 4, 1967</td>
<td>12600</td>
</tr>
<tr>
<td>August 4, 1967</td>
<td>7970</td>
</tr>
<tr>
<td>September 4, 1967</td>
<td>7390</td>
</tr>
<tr>
<td>October 4, 1967</td>
<td>8800</td>
</tr>
<tr>
<td>November 4, 1967</td>
<td>6560</td>
</tr>
<tr>
<td>December 4, 1967</td>
<td>9770</td>
</tr>
</tbody>
</table>
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 fy 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)

Parallel-train (MOSES: phrase translation, LM=5-gram)
Parallel-train (IBM Model 3: no re-ordering, LM=phrase-path)
Decipherment (LM=phrase-path)

[Ravi & Knight 11]
Exploiting Giga-Scale Non-Parallel Text

Accuracy of learned bilingual dictionary

not translations of each other

English text

Foreign text

deciphering engine

bilingual word-for-word dictionary

Newest Method

[Dou & Knight, 2012]

How much foreign text (running words)

(Dou & Knight subm.)
Plan for This Talk

• Break a series of codes
  – Simple letter substitution
  – Phonetic substitution
    • archaeology
    • transliteration
  – Word substitution
  – Foreign language as cipher

• Bonus
  – Two historical ciphers
  – Final thought on translation and cryptography
Copiale Cipher

[Knight, Megyesi, Schaefer 11]
Copiale Cipher

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

Some scratch-outs, rare

Lines ≈ equal length

Paragraphs and section titles always begin with capitalized Roman letters.

Non-enciphered inscriptions: Copiales 3 and Philipp 1866

Preview text fragments ("catchwords")
Letter Frequencies

digraphs:  
& h  99  
č :  66  
§  49  
: u  48  
z  44  

trigraphs:  
& h &  47  
č : u  23  
η & h  22  
ý u h  18  

tendencies:  
â, ê, í, ô, î followed by 3 and į  
â, ê, í, ô, î preceded by z and π
Clustering of Cipher Letters

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

Scipy software

thanks Jon Graehl
Clustering of Cipher Letters

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

Scipy software

thanks Jon Graehl

underlined letters
unaccented Roman letters
circumflexed vowels
First Decipherment Approach

unaccented Roman letters that cluster:

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

most common letter = 12%  
least common = very small

Decipher against 80 plaintext languages.
First Decipherment Approach

unaccented Roman letters that cluster:

most common letter = 12%
least common = very small

Decipher against 80 plaintext languages.

FAIL

80 languages.
Second Decipherment Approach

Homophonic cipher, e.g.:

A = 8 | l y r
B = û
C = o n
D = η
E = x f η f i z z
F = p
G = ŵ

etc.
Homophonic Cipher

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

FAIL

But, slight numerical preference for German
Cipher Characteristics

<table>
<thead>
<tr>
<th>digraphs:</th>
<th></th>
<th>trigraphs:</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>å, ä</td>
<td>99</td>
<td>å, ä, ü</td>
<td>47</td>
</tr>
<tr>
<td>ç :</td>
<td>66</td>
<td>ç : ü</td>
<td>23</td>
</tr>
<tr>
<td>ã, ä</td>
<td>49</td>
<td>ã, ä, ü</td>
<td>22</td>
</tr>
<tr>
<td>: ü</td>
<td>48</td>
<td>ý, ü</td>
<td>18</td>
</tr>
<tr>
<td>z Ä</td>
<td>44</td>
<td>h, ã</td>
<td>17</td>
</tr>
</tbody>
</table>

**Tendencies:**

- å, ê, ì, ô, ù followed by å and ä
- å, ê, ì, ô, ù preceded by z and Ä

**Should appear adjacent in German text:**

- ?
- ?

Make full digraph table for cipher and for German.
Key Observation #1

In Copiale, Ʌ almost always followed by ḱ

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

So guess: Ʌ = C, ḱ = H
One Thing Leads to Another

\[ \begin{align*}
\eta &= \text{CH} & \Rightarrow & \quad \eta \land = \text{CHT} & \Rightarrow & \quad \land = T \ ?
\end{align*} \]

Each step is guesswork.
Must be willing to retract.
Weird task, not knowing German.
No longer care what the book says.
Cluster diagram crucial:

\[ \begin{align*}
\hat{y} &= l & \Rightarrow & \quad i = l , \ \eta = l
\end{align*} \]
Spring Break 2011

German letters

Cipher letters, in groups

Grid
Spring Break 2011

German letters

Cipher letters, in groups

Quite a bit of fooling around →

Grid
German letters

Cipher trigraphs

Trigraph Decoding Guesses

German trigraphs

Quite a bit of fooling around →

Spring Break 2011
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

Actually, those are space bars
gesetz buchs
der hocherleuchtete ♦ e ♦
geheimer theil.
erster abschnitt
geheimer unterricht vor die gesellen.
erster titul.
ceremonien der aufnahme.

wenn die sicherheit der Δ durch den ältern
thürheter besorgt und die Δ vom dirigierenden Λ
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
dirigierenden Λ tisch geführet dieser fragt ihn:

erstlich ob er begehre ♦ 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

First lawbook
of the ☐ e ☐
Secret part.
First section
Secret teachings for apprentices.
First title.
Initiation rite.

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

First, if he desires to become ☐.

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

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

The candidate answers yes.
First lawbook of the Θ e Θ

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

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

First, if he desires to become θ.

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

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

The candidate answers yes.
Voynich Manuscript

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

Latin letter trigram model

\[ \begin{align*}
    a & \rightarrow \{ \text{all Voynich letters} \} \\
    b & \rightarrow \{ \text{all Voynich letters} \} \\
    c & \rightarrow \{ \text{all Voynich letters} \} \\
    \cdots \\
    z & \rightarrow \{ \text{all Voynich letters} \} \\
    _ & \rightarrow _
\end{align*} \]

Input Decipherment

<table>
<thead>
<tr>
<th>Input</th>
<th>Decipherment</th>
</tr>
</thead>
<tbody>
<tr>
<td>VAS92 9FAE AR APAM ZOE ZOR9 QOR92 9 FOR ZOE89 ( \ldots )</td>
<td>quiss squm is onum pom quss hates s qum hatis ( \ldots )</td>
</tr>
</tbody>
</table>

\( \text{VAS92} \_ 9 \text{FAE} \_ \text{AR} \_ \text{APAM} \_ \ldots \)
Letter Clustering

Trigram model over \{a, b, _ \}

\[
\begin{align*}
a & \rightarrow \text{AEIOU} \\
b & \rightarrow \text{BCDFHKNST} \\
_ & \rightarrow _
\end{align*}
\]

Sample tagging with learned model:

\[
\begin{align*}
\text{ab} & _ \text{bba} _ \text{bab} _ \text{b} \\
\text{in} _ \text{the} _ \text{town} _ \\
\text{bbaba} _ \text{a} _ \text{...} \\
\text{where} _ \text{i} _ \text{...}
\end{align*}
\]

\[\text{i n _ t h e _ t o w n _ w h e r e _ i _ w a s} \ldots\]
Letter Clustering

Trigram model over \{a, b, _ \}

\[\text{a} \rightarrow \{\text{all Voynich letters}\} \]

\[\text{b} \rightarrow \{\text{all Voynich letters}\} \]

\[\_ \rightarrow \_ \]

Sample tagging with learned model:

\[\text{V A S 9 2} \_ \text{9 F A E} \_ \text{A R} \_ \text{A P A M} \_ \ldots \]
Letter Clustering

Trigram model over \{a, b, _ \}

a \rightarrow b

\_ \rightarrow \_ 

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 \_ ...

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

English

a \rightarrow \begin{array}{c}
A
E
O
U
\end{array} \begin{array}{c}
\cdots
\end{array}

b \rightarrow \begin{array}{c}
B
C
D
H
L
\end{array} \begin{array}{c}
\cdots
\end{array}

Voynich

a \rightarrow \begin{array}{c}
9
E
R
\end{array} \begin{array}{c}
\cdots
\end{array}

b \rightarrow \begin{array}{c}
8
A
C
O
\end{array} \begin{array}{c}
\cdots
\end{array}

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

P(a)

P(a)
Word Clustering

Bigram model over \{a, b\}

\[ a \rightarrow b \rightarrow a \rightarrow \ldots \]

Sample tagging with learned model:

\[ \begin{align*}
  a & \quad a \quad a \quad a \quad a \\
  VAS92 & \quad 9FAE \quad AR \quad APAM \quad ZOE \\
  a & \quad a \quad a \quad a \quad a \\
  ZOR9 & \quad QRC2 \quad 9 \quad FOR \quad ZOE89 \; \ldots
\end{align*} \]

WAIT, WHAT?
Word Clustering

Voynich words tagged as “a”

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.

1970s NSA report recently declassified!
National Security Agency

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

NSA employs more mathematicians and linguists than any other organization.

NSA has more computers than any other organization.

Association for Computational Linguistics

ACL applies statistics to language processing problems.

1950s

1960s

1970s

1980s

1990s

2000s

2010s

2020s

1970s paper on HMM Voynich

1993 paper on Statistical Machine Translation

2011 paper on HMM Voynich

Oh yeah -- we’ve been on Mars since 1962.
Other Unsolved Ciphers

<table>
<thead>
<tr>
<th>Cipher</th>
<th>Author(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voynich Manuscript (1400)</td>
<td>Reddy &amp; Knight 11</td>
</tr>
<tr>
<td>Zodiac Killer (1967)</td>
<td>Ravi &amp; Knight 11</td>
</tr>
<tr>
<td>FBI cipher (1999)</td>
<td>Snyder, Barzilay, Knight 10</td>
</tr>
<tr>
<td>Kryptos (1990)</td>
<td>OBKR</td>
</tr>
</tbody>
</table>

Unsupervised Training for NLP
Plan for This Talk

• Break a series of codes
  – Simple letter substitution
  – Phonetic substitution
    • archaeology
    • transliteration
  – Word substitution
  – Foreign language as cipher

• Bonus
  – Two historical ciphers
  – Final thought on translation and cryptography
## Future Prospects for Translation

<table>
<thead>
<tr>
<th>Cryptography</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manual</td>
<td>Manual encoding</td>
</tr>
<tr>
<td>Mechanical</td>
<td>Mechanical encoding; intuition-based decryption</td>
</tr>
<tr>
<td>Mathematical</td>
<td>Computer decryption, based on information theory</td>
</tr>
<tr>
<td>Higher math, deeper understanding</td>
<td>Public-key systems, based on number theory</td>
</tr>
</tbody>
</table>
## Future Prospects for Translation

<table>
<thead>
<tr>
<th>Cryptography</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Manual</strong></td>
<td>Human translation</td>
</tr>
<tr>
<td>Manual encoding</td>
<td></td>
</tr>
<tr>
<td><strong>Mechanical</strong></td>
<td>1960s Rule-based MT</td>
</tr>
<tr>
<td>Mechanical encoding; intuition-based decryption</td>
<td></td>
</tr>
<tr>
<td><strong>Mathematical</strong></td>
<td>1990s Statistical MT</td>
</tr>
<tr>
<td>Computer decryption, based on information theory</td>
<td></td>
</tr>
<tr>
<td><strong>Higher math, deeper understanding</strong></td>
<td>2020s ??? ??? ???</td>
</tr>
<tr>
<td>Public-key systems, based on number theory</td>
<td></td>
</tr>
</tbody>
</table>

**Timeline:**
- 1920s
- 1950s
- 1980s
- 1960s
- 1990s
- 2020s
thanks
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]
Deterministic Substitution Constraint
Using ILP instead of EM

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

Using 2-gram letter-based LM

<table>
<thead>
<tr>
<th>Cipher Length</th>
<th>EM error</th>
<th>ILP error</th>
</tr>
</thead>
<tbody>
<tr>
<td>52</td>
<td>85 %</td>
<td>21 %</td>
</tr>
<tr>
<td>98</td>
<td>45 %</td>
<td>12 %</td>
</tr>
<tr>
<td>414</td>
<td>10 %</td>
<td>0.5 %</td>
</tr>
</tbody>
</table>

[Ravi & Knight 08]
Shannon analytically predicted uncertainty about key and message.
Graphed it for a human-level language model.

**Entropy of key, $H(K|C)$**

**Entropy of plaintext message, $H(M|C)$**

**Unicity distance** = $H(K) / (4.7 - LMCE)$

$\log_2 26! \approx 1.6$ bits/char for a human.
Verifying Shannon’s Prediction of Plaintext Message Uncertainty

ANALYTIC CURVES (Shannon’s)

for Human Language Model

ACTUAL CURVES (ours)

<table>
<thead>
<tr>
<th>1-gram</th>
<th>Analytically Predicted Unicity Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>173</td>
<td></td>
</tr>
<tr>
<td>74</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td></td>
</tr>
</tbody>
</table>

[1-gram 2-gram 3-gram]

(Ravi & Knight 08)
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.
Word Substitution Cipher

P(f | e) = 7 x 7 subst table

P(sentence has w1 | sentence has w2)

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

[Knight et al 06]
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…

رضف رئيس السلطة الفلسطينية محمود عباس مجدا
تصريحات وزير الخارجية الإسرائيلي سيفاون شلوم الذي
قال فيها إنه يتعين على إسرائيل إعادة النظر في استعدادها
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-الأوروبية الأولى إنه يتعين على إسرائيل
احترام شعب الشعب الفلسطيني حتى لو فازت حماس بالانتخابات، وأضاف "إذا نجحت حماس أو فتح سيكون
هذا خيار الشعب الفلسطيني، وعلى الجميع قول هذا
الخيار بكل رحب.

من جانبه شجب رئيس الحكومة الفلسطينية أحمد فريق
الطيب الأحادي الجانب للاشدداب الإسرائيلي من غزة،
وأكد أن إسرائيل تبرد مغادرة هذه الأراضي لتعزيز
سيطرتها على الضفة الغربية.

وقال فريق في كلمته خلال مؤتمر عزته وزارة
الأوقاف في رام الله "سليميون من غزة ولكننا لا نعرف
ما هو شكل هذا الانسحاب وماذا سيتركون، وما هو
مستقبل المعايير والحدود، وكل ذلك غامض لأنه فرار
أحادي الجان.
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 FOREST BECAUSE MAN IS THE MOST DANGEROUS ANIMAL OF ALL TO KILL SOMETHING GIVES ME THE MOST THRILLING EXPERIENCE IT IS EVEN BETTER THAN GETTING YOUR ROCKS OFF WITH A GIRL THE BEST PART OF IT IS THAT WHEN I DIE I WILL BE REBORN IN PARADISE AND THEY HAVE KILLED WILL BECOME MY SLAVES I WILL NOT GIVE YOU MY NAME BECAUSE YOU WILL TRY TO SLOI DOWN OR ATOP MY COLLECTION OF SLAVES FOR MY AFTERLIFE EBEORIETEMETHHPITI"
Deciphering Zodiac 408
Bayesian models

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

<table>
<thead>
<tr>
<th>Language Model</th>
<th>Initial Sample</th>
<th>Decipherment Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-gram</td>
<td>Random</td>
<td>62.3 / 48.5 / 47.4</td>
</tr>
<tr>
<td>5-gram</td>
<td>Random</td>
<td>all wrong!</td>
</tr>
<tr>
<td>&quot;</td>
<td>3-gram solution</td>
<td>42.6</td>
</tr>
<tr>
<td>Word 1-gram</td>
<td>Random</td>
<td>all wrong!</td>
</tr>
<tr>
<td>\textit{Interpolated} 5-gram and word 1-gram</td>
<td>Random</td>
<td>79.2</td>
</tr>
<tr>
<td>&quot;</td>
<td>5-gram solution</td>
<td>3.3 / 2.6</td>
</tr>
</tbody>
</table>

[Chiang et al 10]

[Chiang et al 11]

[Chiang et al 10]

[Chiang et al 11]
Unsolved Zodiac 340

Has no obvious reading order bias:

<table>
<thead>
<tr>
<th>% cipher bigram types that repeat (freq &gt; 1)</th>
<th>Left/Right order</th>
<th>Up/Down order</th>
<th>Diag. North-East</th>
<th>Diag. South-East</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zodiac 408 (solved)</td>
<td>13 %</td>
<td>5</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>Zodiac 340 (unsolved)</td>
<td>7</td>
<td>6</td>
<td>8</td>
<td>5</td>
</tr>
</tbody>
</table>

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