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!)


- "What We Know About the Voynich Manuscript," (S. Reddy and K. Knight), Proc. ACL LaTECH, 2011.


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**: PLOKMIJNUHYGVFCDXESZAQW

- Plaintext: HELLO KITTY . . .  
- Ciphertext: NMYYT BUXXQ . . .

- 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 RLQOIQ 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 LOZKTLJXCY MDRCYJQ: "TR HYD FKXC, FQ MKX RLQQIO HYDL MKL DXCTW RDCDLQ JQMNKXTMB PTBMEQOL 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 LQZKTLJKX CY MDBCYJQL: "TR
O . . 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 ore.a re.a o e f
KDCY LQZKLJXX CY MDBCYJQL: "TR
.o a e a freeze or
HYD FKXC, FQ MKX RLQQIQ HYDL
ar . f re e.a
MKL DXCTW RDCDLQ JQMNKXTMB
. o er a . a. o re a r
PTBMYEQL K FKH CY LOZKTL TC."

frequent cipher letters: Q L K C D T M Y X
frequent English letters: e r a i n o s h
a no re.air.a no no.e if
KDCY LQZKTLJKX CY MDBCYJQL: "TR
.o .a n .e a freeze .o r
HYD FKXC, FQ MKX RLQQIO HYDL
ar ni. fn re .e a i
MKL DXCTW RDCDLQ JQMNKXTMB
.i o.er a .a. no re.air in
PTBMYEQL K FKH CY LQZKTL TC."

frequent cipher letters: Q L K C D T M Y X
frequent English letters: e t o a n i r s h

a to re.air.a to to.e if
KDCY LQZKTLJKX CY MDBCYJQL: "TR
.o .a t .e a freeze .o r
HYD FKXC, FQ MKX RLQQIO HYDL
ar ti. ft re .e a i
MKL DXCTW RDCDLQ JQMNKXTMB
.i o.er a .a. to re.air it
PTBMYEQL K FKH CY LQZKTL TC."

frequent cipher letters: Q L K C D T M Y X
frequent English letters: e t o a n i r s h
a to repair. a to to.e if
KDCY LQZKTLJKX CY MDBCYJQL: "TR
.o .a t .e a freeze .o r
HYD FXXC, FQ MKX RLQQIQ HYDL
ar ti. ft re .e .a i
MKL DXCTW RDCDLQ JQMNKXTMB
.i o. er a .a. to repair it
PTBMYEQL K FKH CY LQZKTL TC."

frequent cipher letters: D Z M X
frequent English letters: e t o a n i r s h

auto repairman to customer: if
KDCY LQZKTLJKX CY MDBCYJQL: "TR
you wait we can freeze your
HYD FXXC, FQ MKX RLQQIQ HYDL
car until future mechanics
MKL DXCTW RDCDLQ JQMNKXTMB
discover a way to repair it
PTBMYEQL K FKH CY LQZKTL TC."
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 $\rightarrow$ 02, 14, 16, 22, 49, 51, 58, 90
    - B $\rightarrow$ 04, 76
    - C $\rightarrow$ 15, 56, 71
      etc
  - flattens out ciphertext distribution
    - “a cab...” becomes “22 56 14 04...”
    - still deterministic in the deciphering direction

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

- Transposition ciphers

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

“First NLP Task Ever” (1930s-40s)
Breaking the German Enigma Cipher

German text \(\rightarrow\) over the radio:
DFKWIFKS LWORISJD KSUEIFKR …
\(\rightarrow\) German text

input (intercepted ciphertext): DFKWIFKSLWORISJDKSUEIFKR …
output (plaintext): VASISTDASHERRCAPITANRICH …
“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

cipher text → substitution

proposed plaintext

Is it German? → German

guessed key
Breaking Enigma

SHANNON TURING KULLBACK

Breaking Enigma

POLISH CIPHER BUREAU & BLETCHLEY PARK

LOOK FOR ENIGMA SETTINGS THAT YIELD PLAINTEXT PATTERN "XYZXYZ..."
Breaking Enigma

- cipher text
- guessed key
- proposed plaintext
- Is it German?
- German

POLISH CIPHER BUREAU & BLETCHLEY PARK

LOOK FOR ENIGMA SETTINGS THAT YIELD PLAINTEXT PATTERN "XYZXYZ..."

WINSTON CHURCHILL

Enigma

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

- 1945: War ends
- 1973: Wartime Enigma decipherment leaked
- 1975: Last surplus Enigma given to developing countries
- 1996: One Turing Enigma treatise declassified
- 2012: Another declassified (but have to go to England)

elegant, powerful, widely-applicable mathematics
One method is to try independently all the possible positions for the middle wheel. We shall want to know the middle wheel couplings which are consequences of these various assumptions. This can be done by setting up inverse rods for the middle wheel. The rods are paired off according to R.H.W. couplings, i.e. M.W. output. This has been done for the couplings $ku$, $fx$, $ep$ which arose in the DANZIGVON crib in Fig 55, assuming the red wheel 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 couplings are possible. Our procedure is rather different according as the U.K.W. does or does not rotate. In the case that the U.K.W. does not rotate it will be sufficient to have 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.
One method is to try independently all the possible positions for the middle wheel. We shall want to know the middle wheel couplings which are consequences of these various assumptions. This can be done by setting up inverse rods for the middle wheel. The rods are paired off according to R.H.W. couplings, i.e. M.W. output. This has been done for the couplings $ku, fx, ep$ which arose in the DANZIGVON crib in Fig 55, assuming the red wheel in the middle. The pairs in each column of this setup give possible M.W. couplings. We have now to find out whether these couplings are possible.

Our procedure is rather different according as the U.K.W. does or does not rotate. In the case that the U.K.W. does not rotate it will be sufficient to have 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.
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 ears for announcing its own moves. Mathematics should preferably be restricted to branches where diagrams are not used. Of the above possible fields the learning of language would be the most impressive, since it is the most human of these activities. This field, as a 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 reasonably close parallel between the problems of the physicist and those of the cryptographer. The system on which a message is encoded tends to the laws of the universe, the intersected secrets to the evidence available, the keys for a day or a season 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.
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 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 symbols 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 does not appear to be any, but 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. For the system on which a message is encoded corresponds, at least in part, to the laws of the universe, the interpretative meaning to the evidence available, the keys for a day or a month to the pertinent 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.
Alan Turing, on Thinking Machines

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

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

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

OUR HERO
Statistical Machine Translation

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 ...

Translated, parallel documents

Parallel Corpus

12 English sentences in English and Spanish.

<table>
<thead>
<tr>
<th>English</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a. Garcia and associates .&lt;br&gt;1b. Garcia y asociados .</td>
<td>7a. the clients and the associates are enemies .&lt;br&gt;7b. los clientes y los asociados son enemigos .</td>
</tr>
<tr>
<td>2a. Carlos Garcia has three associates .&lt;br&gt;2b. Carlos Garcia tiene tres asociados .</td>
<td>8a. the company has three groups .&lt;br&gt;8b. la empresa tiene tres grupos .</td>
</tr>
<tr>
<td>3a. his associates are not strong .&lt;br&gt;3b. sus asociados no son fuertes .</td>
<td>9a. its groups are in Europe .&lt;br&gt;9b. sus grupos estan en Europa .</td>
</tr>
<tr>
<td>4a. Garcia has a company also .&lt;br&gt;4b. Garcia tambien tiene una empresa .</td>
<td>10a. the modern groups sell strong pharmaceuticals .&lt;br&gt;10b. los grupos modernos venden medicinas fuertes .</td>
</tr>
<tr>
<td>5a. its clients are angry .&lt;br&gt;5b. sus clientes estan enfadados .</td>
<td>11a. the groups do not sell zenzanine .&lt;br&gt;11b. los grupos no venden zanzanina .</td>
</tr>
<tr>
<td>6a. the associates are also angry .&lt;br&gt;6b. los asociados tambien estan enfadados .</td>
<td>12a. the small groups are not modern .&lt;br&gt;12b. los grupos pequenos no son modernos .</td>
</tr>
</tbody>
</table>
### Parallel Corpus

12 English sentences in Centauri and Arcturan.

<table>
<thead>
<tr>
<th>Centauri</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 jiat quat cat .</td>
<td>11b. wat nnat arrat mat zanzanat .</td>
</tr>
<tr>
<td>6a. lalok sprok izok jok stok .</td>
<td>12a. lalok raroq 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>

### Centauri/Arcturan

**Your assignment, translate this to Arcturan:**

<table>
<thead>
<tr>
<th>Centauri</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 nnatquat 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 jiatquat cat .</td>
<td>11b. wat nnat arrat mat zanzanat .</td>
</tr>
<tr>
<td>6a. lalok sprok izok jok stok .</td>
<td>12a. lalok raroq nok izok hihok mok .</td>
</tr>
<tr>
<td>6b. wat dat kratquat cat .</td>
<td>12b. what nnat forat arrat vat gat .</td>
</tr>
</tbody>
</table>
### Centauri/Arcturan

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

<table>
<thead>
<tr>
<th>Centaurian Text</th>
<th>Arcturan Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a. ok-foon ororok sprok</td>
<td>7a. lalok farok ororok lalok sprok izok enemok</td>
</tr>
<tr>
<td>1b. at-foon bichat dat</td>
<td>7b. wat jjat bichat wat dat vat eneat</td>
</tr>
<tr>
<td>2a. ok-drubel ok-foon anok plok sprok</td>
<td>8a. lalok brok anok plok nok</td>
</tr>
<tr>
<td>2b. at-drubel at-foon 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 oloa at-yurp</td>
</tr>
<tr>
<td>4a. ok-foon anok drok brok jok</td>
<td>10a. lalok mok nok yorok ghirok clok</td>
</tr>
<tr>
<td>4b. at-foon 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>
### Centauri/Arcturan

**Your assignment, translate this to Arcturan:**

<table>
<thead>
<tr>
<th>Centauri/Arcturan</th>
<th>Farok Crrrok Hihok Yorok Clok Kantok Ok-yurp</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a. ok-voon ororok sprok .</td>
<td>7a. lalok <strong>farok</strong> 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 ratt dat .</td>
<td>8b. iat lat pippat ratt 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 <strong>farok</strong> izok stok .</td>
<td>11a. lalok nok <strong>crrrok</strong> 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 rarak 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>

---

### Centauri/Arcturan

**Your assignment, translate this to Arcturan:**

<table>
<thead>
<tr>
<th>Centauri/Arcturan</th>
<th>Farok Crrrok Hihok Yorok Clok Kantok Ok-yurp</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 ratt dat .</td>
<td>8b. iat lat pippat ratt 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 <strong>farok</strong> izok stok .</td>
<td>11a. lalok nok <strong>crrrok</strong> 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 rarak 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>
### Centauri/Arcturan

**Your assignment, translate this to Arcturan:**

<table>
<thead>
<tr>
<th>Centauri/Arcturan</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 jiat 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 crrok [hihok yorok] zanzanok .</td>
</tr>
<tr>
<td>5b. totat jiatquat cat .</td>
<td>11b. wat nnat arrat mat zanzanat .</td>
</tr>
<tr>
<td>6b. wat dat krat quat 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>Centauri/Arcturan</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 jiat 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 crrok [hihok yorok] zanzanok .</td>
</tr>
<tr>
<td>5b. totat jiatquat cat .</td>
<td>11b. wat nnat arrat mat zanzanat .</td>
</tr>
<tr>
<td>6b. wat dat kratquat 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>1a. ok-voon ororok sprok .</th>
<th>7a. lalok farok ororok lalok sprok izok enemok .</th>
</tr>
</thead>
<tbody>
<tr>
<td>1b. at-voon bichat dat .</td>
<td>7b. wat jiat 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 jok .</td>
<td>10a. lalok mok nok yorok ghirok clok .</td>
</tr>
<tr>
<td>4b. at-voon krat pippat sat lat .</td>
<td>10b. wät nnat gät 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 jiat 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>
## Centauri/Arcturan

Your assignment, translate this to Arcturan:

<table>
<thead>
<tr>
<th>Centauri</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 jiat 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 ghrok 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 crrok hihok yorok zanzanok .</td>
</tr>
<tr>
<td>5b. totat jiatquat 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>

---

## Learn Translation Knowledge from Non-Parallel Text?

- **English/Albanian Parallel text**
  - Translation model

- **English text**
  - Translation model
  - **Albanian text**

Is this what Weaver had in mind? We’ll come back to this idea.
Automatic decipherment

Letter Substitution Cipher

ciphertext c
Letter Substitution Cipher

\[ P(p) \rightarrow \text{plaintext } p \rightarrow P(c | p) \rightarrow \text{ciphertext } c \]

"key"

plaintext samples, unrelated to ciphertext
Letter Substitution Cipher

plaintext samples, unrelated to ciphertext

plaintext \( p \) \[ \rightarrow \] ciphertext \( c \)

\( P(p) \rightarrow \) \( \text{"key"} \)

\( P(c \mid p) \)
Letter Substitution Cipher

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

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

plaintext samples, unrelated to ciphertext

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

"key"

EM

best guess plaintext p

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

Viterbi

LM

ciphertext c

Decipherment Accuracy vs. Cipher Length

% of ciphertext characters wrongly deciphered

length of cipher

2-gram

3-gram
**Letter Substitution Cipher**

plaintext samples, unrelated to ciphertext

Find substitution-table values that maximize

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

[Ravi & Knight 09b]

best guess plaintext p

Find plaintext p that maximizes

\[ P(p \mid c) \sim P(p) P(c \mid p)^3 \]

[Knight/Yamada 99]

**Reducing LM Weight During EM**

Set EM to maximize

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

instead of

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

[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.

<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>

Using 2-gram letter-based LM

[Ravi & Knight 08]
[Shannon 46, 49]
"Communication Theory of Secrecy Systems"

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

\[ \text{Entropy of key, } H(K|C) \]
\[ \text{Entropy of plaintext message, } H(M|C) \]

\[
\text{Unicity distance} = \frac{H(K)}{4.7 - \text{LMCE}} \\
\sim 1.6 \text{ bits/char for a human}
\]

Verifying Shannon’s Prediction of Plaintext Message Uncertainty

<table>
<thead>
<tr>
<th>LM</th>
<th>Analytically Predicted Unicity Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-gram</td>
<td>173</td>
</tr>
<tr>
<td>2-gram</td>
<td>74</td>
</tr>
<tr>
<td>3-gram</td>
<td>50</td>
</tr>
<tr>
<td>Human</td>
<td>28</td>
</tr>
</tbody>
</table>

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

Copiale Cipher

Section headers

Lines = equal length

Paragraphs and section titles always begin with capitalized Roman letters.

Non-enciphered inscriptions: Copiales 3 and Philipp 1866

Some scratch-outs, rare

Preview text fragments ("catchwords")

Letter Frequencies

digraphs: 

c : 66
t : 49
z : 48

trigraphs: 

c : 47
t : 23
C : 22
Y : 18

tendencies: 

â, â, ã, ã followed by 3 and j
â, â, ã, ã preceded by z and π
Clustering of Cipher Letters

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

Scipy software

thanks Jon Graehl
First Decipherment Approach

unaccented Roman letters that cluster:

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

most common letter = 12%
least common = very small

Decipher against 80 plaintext languages.
Second Decipherment Approach

Homophonic cipher, e.g.:

A = 8 1 l y r
B = u
C = o n
D = z
E = x f d ! f i z e
F = p
G = y

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>
<th>tendencies:</th>
</tr>
</thead>
<tbody>
<tr>
<td>ç Ç ◊</td>
<td>99</td>
<td>ç Ç ◊ ^</td>
<td>47</td>
<td>â, ë, ç, â, ö followed by ş and ğ</td>
</tr>
<tr>
<td>ç : ű</td>
<td>66</td>
<td>ç : ű ^</td>
<td>23</td>
<td>â, ô, ç, â, ö preceded by z and ì</td>
</tr>
<tr>
<td>ç Ç ^</td>
<td>49</td>
<td>ç Ç ũ</td>
<td>22</td>
<td>ç, â, ö, ç, ö precised by z and ì</td>
</tr>
<tr>
<td>ç ű</td>
<td>48</td>
<td>ç ű</td>
<td>18</td>
<td>ç, â, ö, ç, ö precised by z and ì</td>
</tr>
<tr>
<td>ç ű</td>
<td>44</td>
<td>ç ç</td>
<td>17</td>
<td>ç, â, ö, ç, ö precised by z and ì</td>
</tr>
</tbody>
</table>

Make full digraph table for cipher and for German

Key Observation #1

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

So guess: ç = C, Ç = H
One Thing Leads to Another

\[ \varphi \mathfrak{h} = \text{CH} \rightarrow \varphi \mathfrak{h} \mathfrak{\lambda} = \text{CHT} \rightarrow \mathfrak{\lambda} = \text{T} \]

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

\[ \ddot{y} = 1 \rightarrow \mathrm{u} = 1, \eta = 1 \]
German letters
Cipher letters, in groups
Grid
Spring Break 2011
Quite a bit of fooling around →
German letters
German trigraphs
Cipher trigraphs
Trigraph Decoding
Guesses
Grid

Quite a bit of fooling around →

Key Observation #2

unaccented Roman letters that cluster: 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

Actually, those are space bars
Copiale Decipherment

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

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

First, if he desires to become ♠.

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

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

The candidate answers yes.
Word Substitution Encipherment Key

Word Substitution Keys

Decipherment Key

Encipherment Key
Word Substitution Keys

Numbers/Words Both in Order!

Neither in Order!

Word Substitution

• Interesting for NLP

• Language translation can be viewed as word substitution (and transposition)

• Certainly, that is how IBM models 1-5 view it
**Word Substitution (Small-scale)**

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

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

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

[Knight et al 06]

---

**Word Substitution (Small-scale)**

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

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

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)

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

[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

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

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

---

Foreign Language as a Cipher

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

---

Time Expressions

\[ \text{\@Im} 1990 \\
\text{w/\@ImN} \\
\text{fy @lywm} \\
\text{mn !ls=hr @ljry} \\
\text{!lqn} \\
\text{!y!lm} \\
\text{@ImaN} \\
\text{!lsl@&} \\
\text{17 shbIT 1994} \\
\text{thl!th snwt} \\
\text{dyq@} \\
\text{=dh=hr !lsn&} \\
\text{ywmyn} \\
\text{mn !@Im !lmlIDy} \\
\text{!lsn& !lmbqbl&} \\
\text{fy !lsn&} \\
\text{kl ywm} \\
\text{fy !@Im !lmlIDy} \]

\[ \text{\@Im} 2013 \\
\text{w/\@ImN} \\
\text{fy !lywm} \\
\text{mn !ls=hr @ljry} \\
\text{!lqn} \\
\text{!y!lm} \\
\text{@ImaN} \\
\text{!lsl@&} \\
\text{17 shbIT 1994} \\
\text{thl!th snwt} \\
\text{dyq@} \\
\text{=dh=hr !lsn&} \\
\text{ywmyn} \\
\text{mn !@Im !lmlIDy} \\
\text{!lsn& !lmbqbl&} \\
\text{fy !lsn&} \\
\text{kl ywm} \\
\text{fy !@Im !lmlIDy} \]
Time Expressions

@lm 1990
w!lth!ny&
fy ll!ywm
mn !l!sh=hr l!lj!ry
l!q!rn
!'y!lm
@lm!a

@lm 1992
@lm 1993
yw

@lsbw@ !lm!Dy
fy lldaq&q
@lsn& l!lj!ry&
@lsn&
@lsh=hr !lm!Dy
@lsh=hr l!lj!ry
snwit
sn&
=hdh! @lm
s@&
@l@Sr
@lm 1991

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 ly!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 ly!wl 1926
24 !~Adh!r 1993
22 ly!wl 1957
22 tshryn !!!wl 1948
22 tmwz 1952
21 ly!r 1994
21 k!nwn !!!wl 1988

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

<n> Hzyr!n <n>

<table>
<thead>
<tr>
<th>#</th>
<th>Expression 1</th>
<th>#</th>
<th>Expression 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>4 Hzyr!n 1967</td>
<td>2</td>
<td>fy 30 Hzyr!n 1995</td>
</tr>
<tr>
<td>12</td>
<td>fy 12 Hzyr!n 1993</td>
<td>2</td>
<td>fy 18 Hzyr!n 1994</td>
</tr>
<tr>
<td>7</td>
<td>5 Hzyr!n 1967</td>
<td>2</td>
<td>fy 14 Hzyr!n 1993</td>
</tr>
<tr>
<td>6</td>
<td>fy 30 Hzyr!n 1989</td>
<td>2</td>
<td>fy 14 Hzyr!n 1991</td>
</tr>
<tr>
<td>6</td>
<td>30 Hzyr!n 1989</td>
<td>2</td>
<td>fy 12 Hzyr!n 1990</td>
</tr>
<tr>
<td>4</td>
<td>fy 30 Hzyr!n 1994</td>
<td>2</td>
<td>7 Hzyr!n 1994</td>
</tr>
<tr>
<td>4</td>
<td>fy 30 Hzyr!n 1993</td>
<td>2</td>
<td>6 Hzyr!n 1941</td>
</tr>
<tr>
<td>3</td>
<td>fy 19 Hzyr!n 1967</td>
<td>2</td>
<td>26 Hzyr!n 1994</td>
</tr>
<tr>
<td>2</td>
<td>ywm 30 Hzyr!n 1989</td>
<td>2</td>
<td>21 Hzyr!n 1994</td>
</tr>
<tr>
<td>2</td>
<td>w 6 Hzyr!n 1994</td>
<td>2</td>
<td>1 Hzyr!n 1994</td>
</tr>
<tr>
<td>2</td>
<td>qbl 5 Hzyr!n 1967</td>
<td>2</td>
<td>19 Hzyr!n 1965</td>
</tr>
<tr>
<td>2</td>
<td>fy 9 Hzyr!n 1967</td>
<td>2</td>
<td>18 Hzyr!n 1994</td>
</tr>
<tr>
<td>2</td>
<td>fy 7 Hzyr!n 1981</td>
<td>2</td>
<td>18 Hzyr!n 1940</td>
</tr>
<tr>
<td>2</td>
<td>fy 6 Hzyr!n 1994</td>
<td>2</td>
<td>12 Hzyr!n 1993</td>
</tr>
<tr>
<td>2</td>
<td>fy 5 Hzyr!n 1967</td>
<td>2</td>
<td>11 Hzyr!n 1994</td>
</tr>
</tbody>
</table>
Time Expressions

Hzyr!n

<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

<table>
<thead>
<tr>
<th>Year</th>
<th>Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>229</td>
<td>fy Hzyr!n 1 lm!Dy</td>
</tr>
<tr>
<td>207</td>
<td>sh Hzyr!n</td>
</tr>
<tr>
<td>75</td>
<td>fy Hzyr!n 1 lm!bl</td>
</tr>
<tr>
<td>61</td>
<td>fy Hzyr!n 1993</td>
</tr>
<tr>
<td>31</td>
<td>fy Hzyr!n 1992</td>
</tr>
<tr>
<td>27</td>
<td>ll!lb@ mn Hzyr!n</td>
</tr>
<tr>
<td>27</td>
<td>fy Hzyr!n 1967</td>
</tr>
<tr>
<td>19</td>
<td>fy 30 Hzyr!n 1 lm!Dy</td>
</tr>
<tr>
<td>18</td>
<td>fy n=h!y&amp; Hzyr!n 1 lm!Dy</td>
</tr>
<tr>
<td>18</td>
<td>fy Hzyr!n 1991</td>
</tr>
<tr>
<td>17</td>
<td>mn Hzyr!n</td>
</tr>
<tr>
<td>17</td>
<td>mndh Hzyr!n 1 lm!Dy</td>
</tr>
<tr>
<td>17</td>
<td>4 Hzyr!n</td>
</tr>
<tr>
<td>16</td>
<td>n=h!y&amp; Hzyr!n 1 lm!Dy</td>
</tr>
<tr>
<td>16</td>
<td>fy Hzyr!n 1990</td>
</tr>
<tr>
<td>15</td>
<td>sh=hr Hzyr!n</td>
</tr>
<tr>
<td>15</td>
<td>fy sh=hr Hzyr!n 1 lm!Dy</td>
</tr>
<tr>
<td>15</td>
<td>fy Hzyr!n 1994</td>
</tr>
<tr>
<td>14</td>
<td>mn 17 Hzyr!n</td>
</tr>
<tr>
<td>14</td>
<td>fy Hzyr!n 1996</td>
</tr>
<tr>
<td>14</td>
<td>fy 30 Hzyr!n</td>
</tr>
<tr>
<td>13</td>
<td>fy n=h!y&amp; Hzyr!n 20 lm!Dy</td>
</tr>
<tr>
<td>13</td>
<td>fy 20 Hzyr!n 1 lm!Dy</td>
</tr>
<tr>
<td>13</td>
<td>4 Hzyr!n 1967</td>
</tr>
<tr>
<td>12</td>
<td>n=h!y&amp; Hzyr!n</td>
</tr>
<tr>
<td>12</td>
<td>ll!lb@ mn Hzyr!n 1967</td>
</tr>
</tbody>
</table>

Deciphering Spanish Time Expressions

MT quality on test set

(Edit distance, lower is better)

[Decipherment (LM=phrase-path)]

[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

How much foreign text (running words)

Deciphering Foreign Language at Giga-Scale

(Dou & Knight 2012)

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
African Languages

African Languages
African Languages

CONTINENTAL UNITED STATES

ALASKA GREENLAND

INDIA AND PAKISTAN

SOUTHEAST ASIA
Africa

almost as big as the Moon

Zero languages spoken
Unsolved ciphers
Voynich Manuscript (VMS)

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

<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1576-1612</td>
<td>Rudolf II purchases VMS</td>
</tr>
<tr>
<td>1608-1622</td>
<td>J. de Tepenecz signs VMS in Bohemian court</td>
</tr>
<tr>
<td>1630s</td>
<td>George Baresch owns VMS and sends letter to Kircher</td>
</tr>
<tr>
<td>1639</td>
<td>GB writes Kircher again</td>
</tr>
<tr>
<td>1665</td>
<td>Marci sends VMS to Kircher with letter</td>
</tr>
<tr>
<td>1665-80</td>
<td>Kircher owns VMS</td>
</tr>
<tr>
<td>1680</td>
<td>Kircher dies</td>
</tr>
</tbody>
</table>

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 Archives |
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 |
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

<table>
<thead>
<tr>
<th>Encoding:</th>
<th>Decoding:</th>
</tr>
</thead>
<tbody>
<tr>
<td>A → CC, OM, ...</td>
<td>... → DO → N</td>
</tr>
<tr>
<td>B → ...</td>
<td>...</td>
</tr>
<tr>
<td>N → HA, MI, DO, NU ...</td>
<td>...</td>
</tr>
<tr>
<td>Z → ...</td>
<td></td>
</tr>
</tbody>
</table>

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:**
  
  \[a \ n \ n \ldots \rightarrow AD \ DB \ BR \ldots \rightarrow ADBR \ldots\]

• **Then, employ anagramming:**
  
  \[a \ n \ n \ldots \rightarrow OM \ DO \ MI \ldots \rightarrow DO \ OM \ MI \ldots \rightarrow DOMI \ldots\]

• 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...
William Newbold, Polymath, PhD UPenn

... 499cc89 ... apparent ciphertext

... 499cc89 ... apparent ciphertext

real ciphertext: DOMI...
Let’s Decipher with Newbold!

Hcc89 ... → DOMI...

real ciphertext

Doubling

DOMI...

non-deterministic anagramming

DO OM MI...

lookup in 22² table

OM DO MI...

non-deterministic mapping from 11 Latin letters to full 22

a n n...

real ciphertext

non-deterministic anagramming

22x22 table (values guessed)

OM DO MI...

lookup in 22² table

a n n...

non-deterministic mapping from 11 Latin letters to full 22

o n n...
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 40E FCC89 40FCC9 40P9 SCBS9 40BSC9 EFAM OPAE29
2ZC9 40FC9 40FAM 289 40FCC9 SC89 40FCC9 40FCC9 ESC89 E0P9
8ZC9 40PCCC9 8ARSC89 40FC9 40P9

(last paragraph, f103r)
Alphabet: Currier/D’Imperio
Transcription

<table>
<thead>
<tr>
<th>C</th>
<th>S</th>
<th>Z</th>
<th>P</th>
<th>F</th>
<th>B</th>
<th>V</th>
<th>Q</th>
<th>X</th>
<th>W</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>J</td>
<td>A</td>
<td>E</td>
<td>R</td>
<td>O</td>
<td>I</td>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

VMS Letters

<table>
<thead>
<tr>
<th>count</th>
<th>letter</th>
</tr>
</thead>
<tbody>
<tr>
<td>25468</td>
<td>O</td>
</tr>
<tr>
<td>20227</td>
<td>C</td>
</tr>
<tr>
<td>17655</td>
<td>9</td>
</tr>
<tr>
<td>14281</td>
<td>A</td>
</tr>
<tr>
<td>12973</td>
<td>8</td>
</tr>
<tr>
<td>11008</td>
<td>S</td>
</tr>
<tr>
<td>10471</td>
<td>E</td>
</tr>
<tr>
<td>10026</td>
<td>F</td>
</tr>
<tr>
<td>6716</td>
<td>R</td>
</tr>
<tr>
<td>5994</td>
<td>P</td>
</tr>
<tr>
<td>5423</td>
<td>4</td>
</tr>
<tr>
<td>4501</td>
<td>Z</td>
</tr>
<tr>
<td>4076</td>
<td>M</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>count</th>
<th>letter</th>
</tr>
</thead>
<tbody>
<tr>
<td>2886</td>
<td>2</td>
</tr>
<tr>
<td>1752</td>
<td>N</td>
</tr>
<tr>
<td>1413</td>
<td>B</td>
</tr>
<tr>
<td>1046</td>
<td>J</td>
</tr>
<tr>
<td>950</td>
<td>Q</td>
</tr>
<tr>
<td>908</td>
<td>X</td>
</tr>
<tr>
<td>591</td>
<td>T</td>
</tr>
<tr>
<td>524</td>
<td>*</td>
</tr>
<tr>
<td>431</td>
<td>V</td>
</tr>
<tr>
<td>316</td>
<td>I</td>
</tr>
<tr>
<td>217</td>
<td>W</td>
</tr>
<tr>
<td>157</td>
<td>D</td>
</tr>
<tr>
<td>156</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>count</th>
<th>letter</th>
</tr>
</thead>
<tbody>
<tr>
<td>148</td>
<td>U</td>
</tr>
<tr>
<td>96</td>
<td>6</td>
</tr>
<tr>
<td>74</td>
<td>Y</td>
</tr>
<tr>
<td>52</td>
<td>K</td>
</tr>
<tr>
<td>31</td>
<td>G</td>
</tr>
<tr>
<td>17</td>
<td>L</td>
</tr>
<tr>
<td>14</td>
<td>H</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Total
63k character tokens
### VMS Words

<table>
<thead>
<tr>
<th>count</th>
<th>word</th>
<th>count</th>
<th>word</th>
</tr>
</thead>
<tbody>
<tr>
<td>863</td>
<td>8AM</td>
<td>212</td>
<td>0FAM</td>
</tr>
<tr>
<td>537</td>
<td>OE</td>
<td>211</td>
<td>8AN</td>
</tr>
<tr>
<td>501</td>
<td>SC89</td>
<td>191</td>
<td>4OFAE</td>
</tr>
<tr>
<td>469</td>
<td>AM</td>
<td>186</td>
<td>BOE</td>
</tr>
<tr>
<td>426</td>
<td>ZC89</td>
<td>177</td>
<td>4OFCC9</td>
</tr>
<tr>
<td>396</td>
<td>SOE</td>
<td>174</td>
<td>SCC9</td>
</tr>
<tr>
<td>363</td>
<td>OR</td>
<td>172</td>
<td>SOE</td>
</tr>
<tr>
<td>350</td>
<td>AR</td>
<td>155</td>
<td>OCR</td>
</tr>
<tr>
<td>344</td>
<td>SC9</td>
<td>154</td>
<td>OPAM</td>
</tr>
<tr>
<td>318</td>
<td>8AR</td>
<td>153</td>
<td>OPAM</td>
</tr>
<tr>
<td>308</td>
<td>4OFCC9</td>
<td>152</td>
<td>4OFAR</td>
</tr>
<tr>
<td>305</td>
<td>4OFCC9</td>
<td>151</td>
<td>8AM</td>
</tr>
<tr>
<td>283</td>
<td>2C9</td>
<td>150</td>
<td>8AM</td>
</tr>
<tr>
<td>279</td>
<td>4OFAN</td>
<td>147</td>
<td>4F9</td>
</tr>
<tr>
<td>272</td>
<td>4OFCC9</td>
<td>144</td>
<td>4OFAN</td>
</tr>
<tr>
<td>270</td>
<td>89</td>
<td>144</td>
<td>4OFAN</td>
</tr>
<tr>
<td>262</td>
<td>4OFAN</td>
<td>143</td>
<td>OPAE</td>
</tr>
<tr>
<td>260</td>
<td>AE</td>
<td>141</td>
<td>OPA</td>
</tr>
<tr>
<td>253</td>
<td>8AE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>243</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>219</td>
<td>SOR</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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
  - 40fC89 40fC89 ...

- 8 distinct words appear tripled
  - 40fC89 40fC89 40fC89 ...
  - 0fC89 0fC89 0fC89 ...
  - 1fC89 1fC89 1fC89 ...
  - 2fC89 2fC89 2fC89 ...
  - 3fC89 3fC89 3fC89 ...
  - 4fC89 4fC89 4fC89 ...
  - 5fC89 5fC89 5fC89 ...
  - 6fC89 6fC89 6fC89 ...
  - 7fC89 7fC89 7fC89 ...
  - 8fC89 8fC89 8fC89 ...
  - 9fC89 9fC89 9fC89 ...
  - 0fC89 0fC89 0fC89 ...
  - 1fC89 1fC89 1fC89 ...
  - 2fC89 2fC89 2fC89 ...
  - 3fC89 3fC89 3fC89 ...
  - 4fC89 4fC89 4fC89 ...
  - 5fC89 5fC89 5fC89 ...
  - 6fC89 6fC89 6fC89 ...
  - 7fC89 7fC89 7fC89 ...
  - 8fC89 8fC89 8fC89 ...
  - 9fC89 9fC89 9fC89 ...
  - 0fC89 0fC89 0fC89 ...
  - 1fC89 1fC89 1fC89 ...
  - 2fC89 2fC89 2fC89 ...
  - 3fC89 3fC89 3fC89 ...
  - 4fC89 4fC89 4fC89 ...
  - 5fC89 5fC89 5fC89 ...
  - 6fC89 6fC89 6fC89 ...
  - 7fC89 7fC89 7fC89 ...
  - 8fC89 8fC89 8fC89 ...
  - 9fC89 9fC89 9fC89 ...

+ many more!
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 _ ...
```

Sample tagging with learned model:

```
_ a b b a _ b a b _
in the town _
```

```
_ b b a a _ a _
where _ i _ ...
```

```
in the _ town _ where _ i _ was ...
```
Letter Clustering

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

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

Sample tagging with learned model:

\[
\]

Sample tagging with learned model:

\[
\]
Letter Clustering

English

\[ P(\text{letter} \mid \text{tag}) \]

\[ P(\text{tag} \mid \text{letter}) \]

Voynich

\[ P(a) \]

\[ P(a) \]

Bigram model over \{a, b\}

a \rightarrow b \rightarrow a \rightarrow b \rightarrow a \rightarrow b \rightarrow a \rightarrow b \rightarrow a \rightarrow ...

Sample tagging with learned model:

```
VAS92 9FAE AR APAM ZOE
a a a a a a ...
ZOR9 QRC2 9 FOR ZOE89 ...
```

WAIT, WHAT?
Voynich sections, per drawings observed. Captain Currier’s "two languages" (1976).

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.

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


Association for Computational Linguistics

1950s

1960s

1970s paper on HMM Voynich

1970s

1980s

1990s

1993 paper on Statistical Machine Translation

2000s

2010s

2011 paper on HMM Voynich

2020s

ACL applies statistics to language processing problems

CRYPTOLOG

NSA newsletter declassified in 2013.


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

Heavily redacted.
CRYPTOLOG: Voynich

The Voynich Manuscript

The manuscript has been written in a script called ‘Voynich script’ which has puzzled scholars for centuries. It is one of the few manuscripts that have not been deciphered. It is believed to have been written in the 14th or 15th century.

CRYPTOLOG: Machine Translation

Machine translation is the automatic process of translating text from one language into another. It involves the use of computer algorithms and artificial intelligence to understand and generate meaning in text.

Partial Machine Translation: A Final Report

Partial machine translation involves breaking down a text into smaller segments and translating each segment separately. This approach can be useful when dealing with complex or highly technical text.

MACHINE TRANSLATION: What can it do for us?

Machine translation can be used in a variety of applications, such as interpreting, data entry, and text analysis. It can also be used to improve the accessibility of content for people who speak different languages.

As machine translation stands today, we haven’t reached the stage where we can feed a source (foreign-language) test into a computer and produce a text in the target (in our case, English) language as we do for some other objects like music, painting, etc.
CRYPTOLOG: Evaluating Translations

An Objective Approach to SCORING TRANSLATIONS

Author's note: The philosophy underlying the translation grading system described in this paper has been developed and applied by Henry Tetsu 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 "he" reflects this collaboration. I personally take full responsibility for presenting our findings here.

Translation as an intellectual activity has been practiced since antiquity for practical as well as literary reasons. The traditional means of evaluating translation quality has been through subjective judgments across the board, using criteria such as source language-to-English equivalence and fluency. Over the past 2 or 3 years I have developed a way to score translations that provides a more objective measure of quality, which may obviate this problem even though our results are imperfect (total grading any kind of connected impossible). Our first large project, which I will describe in detail, was the Russian PGE. We have sub projects in a number of other PGEs in languages, mainly Indo-European on other families. The results are promising enough in those instances to lend its use in the PGE Handbook.

CRYPTOLOG: Linguists

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

Most linguists specified that they wanted 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 elevate successfully with managers or others for promotion.

Despite almost unanimous complaints about lack of recognition, few specific suggestions on how to rectify the situation were forthcoming. Working conditions, salary, and promotions were often mentioned. It was felt that the Agency reflects the physical as well as the psychological needs of its linguists.

In the past, as the Agency has grown, linguists have had to wait a rather long time before a position became available. Those who have been promoted have often had to leave for other positions in the Agency before a suitable one appears.

PROMOTION. The word inevitably stirs the 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.

PUBLICATIONS. Publication does not equate with reports prepared as a regular part of the job. The Agency is now publishing some of its technical and scientific work, and a number of linguists have been involved in these efforts.

SOME TIPS ON GETTING PROMOTED

Written for the April 1978 edition of WIN (Women in NSA) in response to the article on promotion. The words inevitably stir some kind of reaction in every red-blooded NSA employee: hope, pleasure, challenge; despair, frustration, disappointment; even inertia, resentment, resignation. Despite disparate views on promotion.

PERSONAL SUMMARY

Serving on the Agency Grade 14 for my experience there has simply held impressions and reinforced the critical importance of the content and context of this piece.

TEACHING COMPUTER SCIENCE TO LINGUISTS

by P16

He entered the Agency with foreign language skills and taught courses in the area of linguistics. His students found him to be knowledgeable, engaging, and approachable. He felt that the Agency was so foreign to the students that they found it difficult to approach him with questions or concerns.

PUBLICATIONS. The Agency is now publishing some of its technical and scientific work, and a number of linguists have been involved in these efforts. The word inevitably stirs some kind of reaction in every red-blooded NSA employee: hope, pleasure, challenge; despair, frustration, disappointment; even inertia, resentment, resignation. Despite disparate views on promotion.

PUBLICATIONS. The Agency is now publishing some of its technical and scientific work, and a number of linguists have been involved in these efforts. The word inevitably stirs some kind of reaction in every red-blooded NSA employee: hope, pleasure, challenge; despair, frustration, disappointment; even inertia, resentment, resignation. Despite disparate views on promotion.
Back to Word Clustering

Bigram model over \{a, b\}

\[a \rightarrow b \rightarrow a \]...

Let’s try 10 clusters.

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

10-Class Word Clustering: English

\[\text{MY \ THE \ AND} \]

\[\text{etc \ etc \ etc} \]

\[\text{HAD \ WOULD \ SAID \ TO} \]

\[\text{etc \ etc \ etc} \]
10 Classes: 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

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

(Reedy & Knight, 2011)

# times that word \( w \) occurs in page \( d \)

# pages that contain word \( w \)
Do Content Words Indicate Topics?

Frequency of EFCC89 in Voynich B pages

Are VMS Pages in Order?

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

![Graph showing % of correct guesses for Hungarian, Arabic, WSJ, and Voynich B bigrams and unigrams. Voynich does not have many repeated bigrams.](image)

- Hungarian
- Arabic
- WSJ
- Voynich B
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 PARADICE AND THEY HAVE KILLED WILL BECOME MY SLAVES. I WILL NOT GIVE YOU MY NAME BECAUSE YOU WILL TRY TO SLOW DOWN OR ATOP MY COLLECTION OF SLAVES FOR MY AFTERLIFE EBEORIETEMETHPIT"

Plaintext has many misspellings

Final 18 plaintext characters of 408 are "junk"

Deciphering Zodiac 408
Bayesian models & Gibbs sampling

<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</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>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>

[Ravi & Knight 11]

See also Malte Nuhn's paper at ACL 2013!
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.

Other Unsolved Ciphers

Beale (1885)  
Dorabella (1897)  
Kryptos (1990)  
Taman Shud (1948)  
FBI (1999)  

(2 pp total)
Collected Ciphers

+ many more!

Writing as a code for speech
J. Eric S. Thompson

Yuri Knorozov

Resists notion that the glyphs have a phonetic component.

It’s phonetic.
Archaeological Decipherment

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

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

[Knight & Yamada 99]
"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..."

[Don Quixote]

[Knight & Yamada 99]
Archaeological Decipherment

Phoneme trigram model
\[ P(L \mid ts \_a) = 0.03 \]

Modern Spanish sounds

Phoneme-to-letter model
\[ P(y \mid L) = ? \]

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 \( P(c \mid p) \) 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:** \textit{primera parte del ingenioso hidalgo don ...}

**Output:** \textit{NSV.NV.NV NVS.NV NVS VS.NV.SV.V.NV ...}

\textbf{S} = sonorous consonant phoneme  
\textbf{N} = non-sonorous consonant phoneme  
\textbf{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 \textbf{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

**SPEECH RECOGNITION**
- [Bahl, Jelinek, Mercer 83]
- [Jelinek, Bahl, Mercer 75]
- [Viterbi 67]
- [Jelinek 69]
- [Fano 63]
- [Shannon 51]

**MACHINE TRANSLATION**
- [Brown et al 90]
- [Bahl & Mercer 76]
  - part-of-speech tagging

**STATISTICAL NLP**
- [Kelly 56]
- [Shannon 49]
- [Shannon 46]
- [Shannon 50]

**CRYPTOGRAPHY**
- [Shannon 46]
- [Shannon 49]
- [Shannon 51]

**INFORMATION THEORY**
- [Hartley 28]
- [Fano 63]
- [Shannon 48]

**FINANCE**
- Brown & Mercer [today]

**COMPUTER GAMES**
- Brown & Mercer [today]

**HISTORY**
- Shannon

*...insights into communication theory & cryptography developed simultaneously and “they were so close together you couldn’t separate them” (Shannon)*
## Decipherment and NLP

<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;</td>
</tr>
<tr>
<td></td>
<td>intuition-based decryption</td>
</tr>
<tr>
<td>Mathematical</td>
<td>Computer decryption,</td>
</tr>
<tr>
<td></td>
<td>based on information theory</td>
</tr>
<tr>
<td>Higher math, deeper understanding</td>
<td>Public-key systems,</td>
</tr>
<tr>
<td></td>
<td>based on number theory</td>
</tr>
</tbody>
</table>

- 1920s
- 1950s
- 1960s
- 1990s
- 1980s
- 2020s

---

thanks