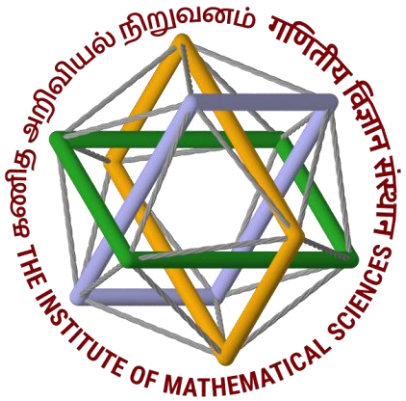
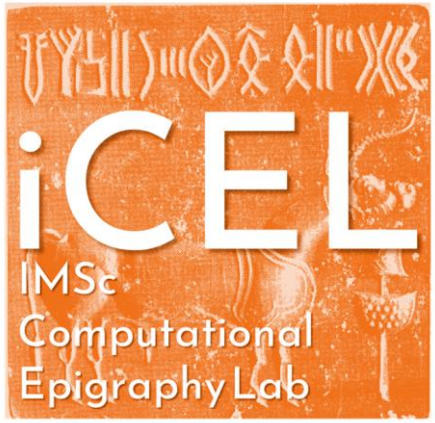


Bits & Scripts



Tutorial

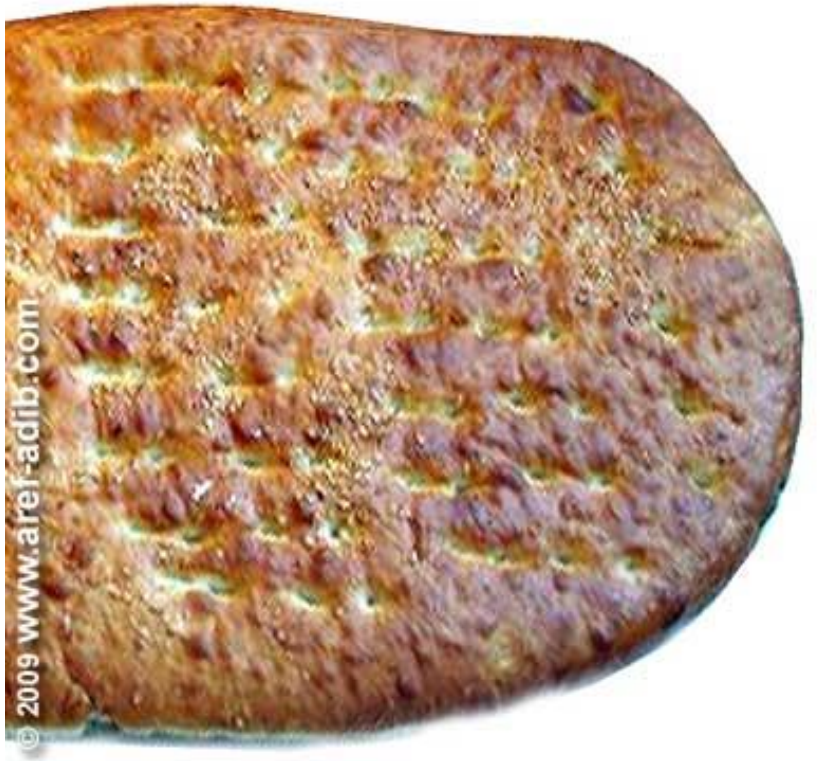
Encountering Writing

Sitabhra Sinha

How can you tell…

This is writing…

…but this is not !

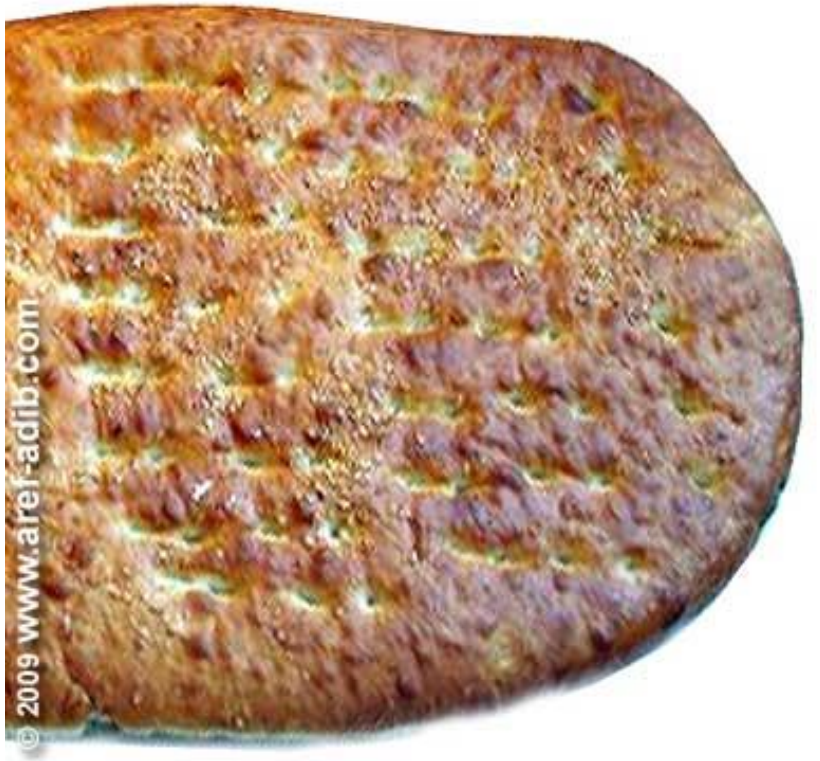


How can you tell…

This is writing…

an Achameneid cuneiform tablet

…but this is not !



a Nan e-Barbari (Persian flat bread)

Writing and Language are not equivalent!

What is writing ? a system of communicating – typically using a language - by means of conventional visible marks

What is language ? A system of **syntactic** communication capable of encoding ideas of arbitrary complexity

Syntax = Compositionality + Recursion

("words/letters") (embedding)

A writing system can be used to write different languages e.g., Roman alphabet

Look for syntactic structure in a symbolic string

The same writing system can be used to write several different languages…

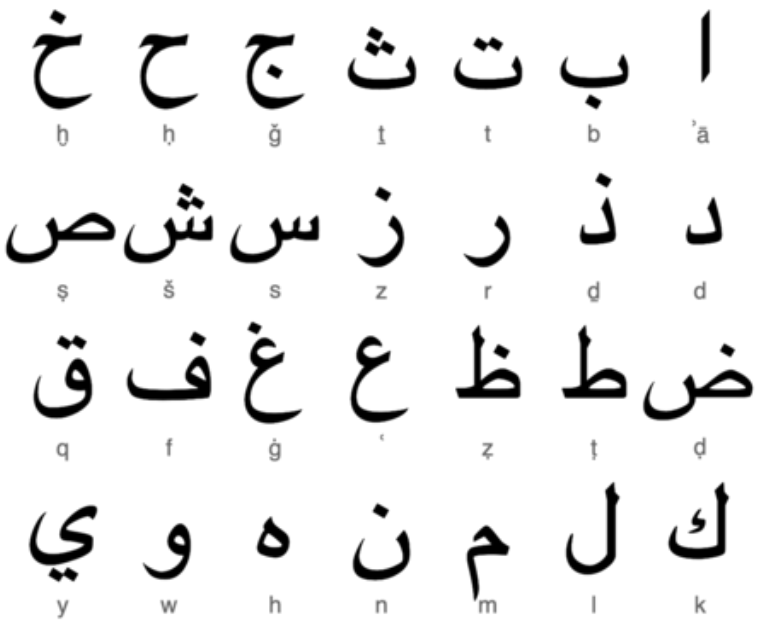
Examples

- **Latin alphabet**

used to write many modern European languages, including English and many modern Asian languages, including Malay, Turkish and Indonesian

- **Arabic script**

used to write texts in Arabic, Persian (Farsi and Dari), Malay (Jawi), Cham (Akhar Srak), Uyghur, Kurdish, Punjabi (Shahmukhi), Sindhi, Balti, Balochi, Pashto, Lurish, Urdu, Kashmiri, Rohingya, Somali, among others

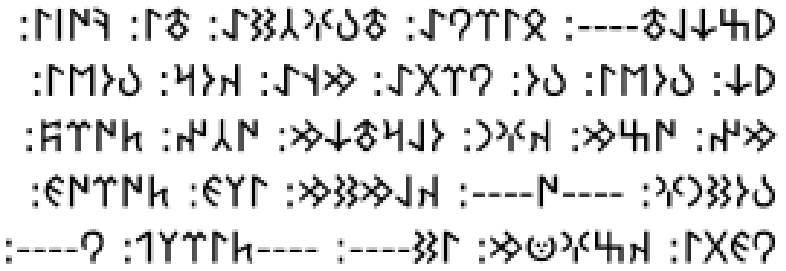


… and the same language can be written using different writing systems

Example

Turkic Languages have been written using

- **Orkhon script**
  (8th-10th century CE)
  written from right to left



- **Ottoman Turkish alphabet**
  (10th-20th century)
  A form of the Perso-Arabic script
  written from right to left

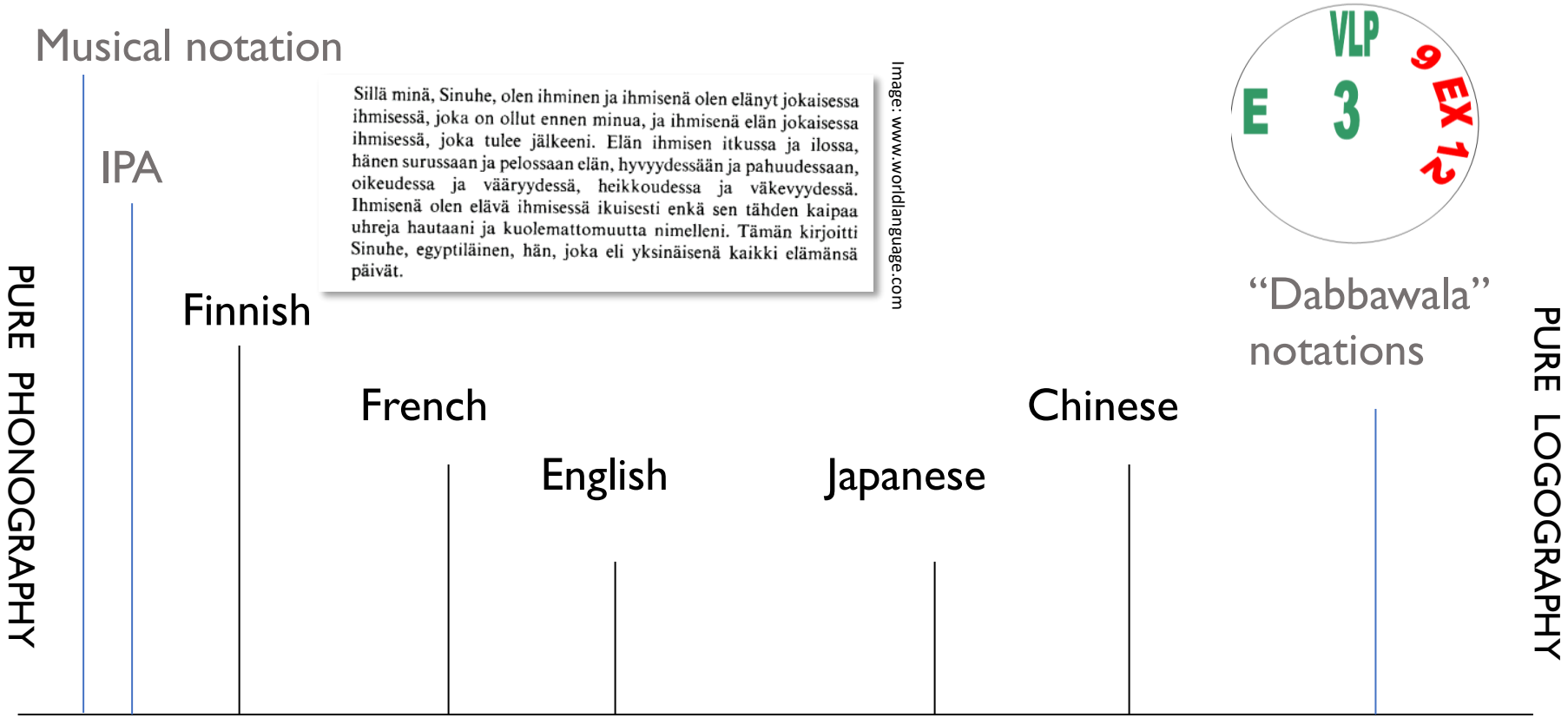
- **Latin Turkish alphabet**
  (1928 onwards)
  written from left to right

Aa Bb Cc Çç Dd Ee Ff Gg Ğğ Hh
Iı İi Jj Kk Ll Mm Nn Oo Öö Pp Rr
Ss Şş Tt Uu Üü Vv Yy Zz

The spectrum of writing

All known writing systems involve a mix of

- Phonetic elements: signs with sound values
- Logographic elements: semantic signs



(Robinson 2002)

Classifying writing systems based on number of characters

Alphabetic (~25 signs)

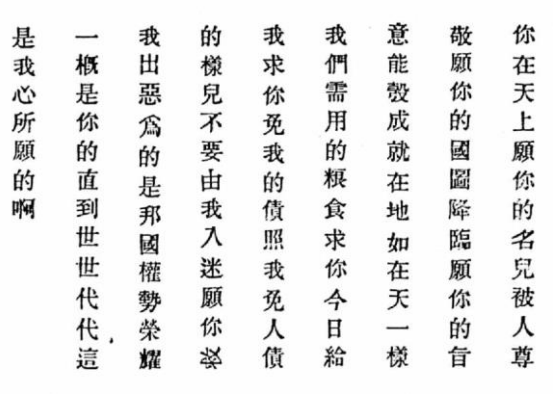
E.g., Latin

Syllabic (~100 signs)

あ a	い i	う u	え e	お o
か ka	き ki	く ku	け ke	こ ko
さ sa	し shi	す su	せ se	そ so
た ta	ち chi	つ tsu	て te	と to
な na	に ni	ぬ nu	ね ne	の no
は ha	ひ hi	ふ fu	へ he	ほ ho
ま ma	み mi	む mu	め me	も mo
や ya		ゆ yu		よ yo
ら ra	り ri	る ru	れ re	ろ ro
わ wa				を (w)o
ん n				

E.g., Japanese Kana

Ideographic (>50000 signs)



E.g., Chinese

Logo-syllabic (~900 signs)

E.g., Sumerian cuneiform

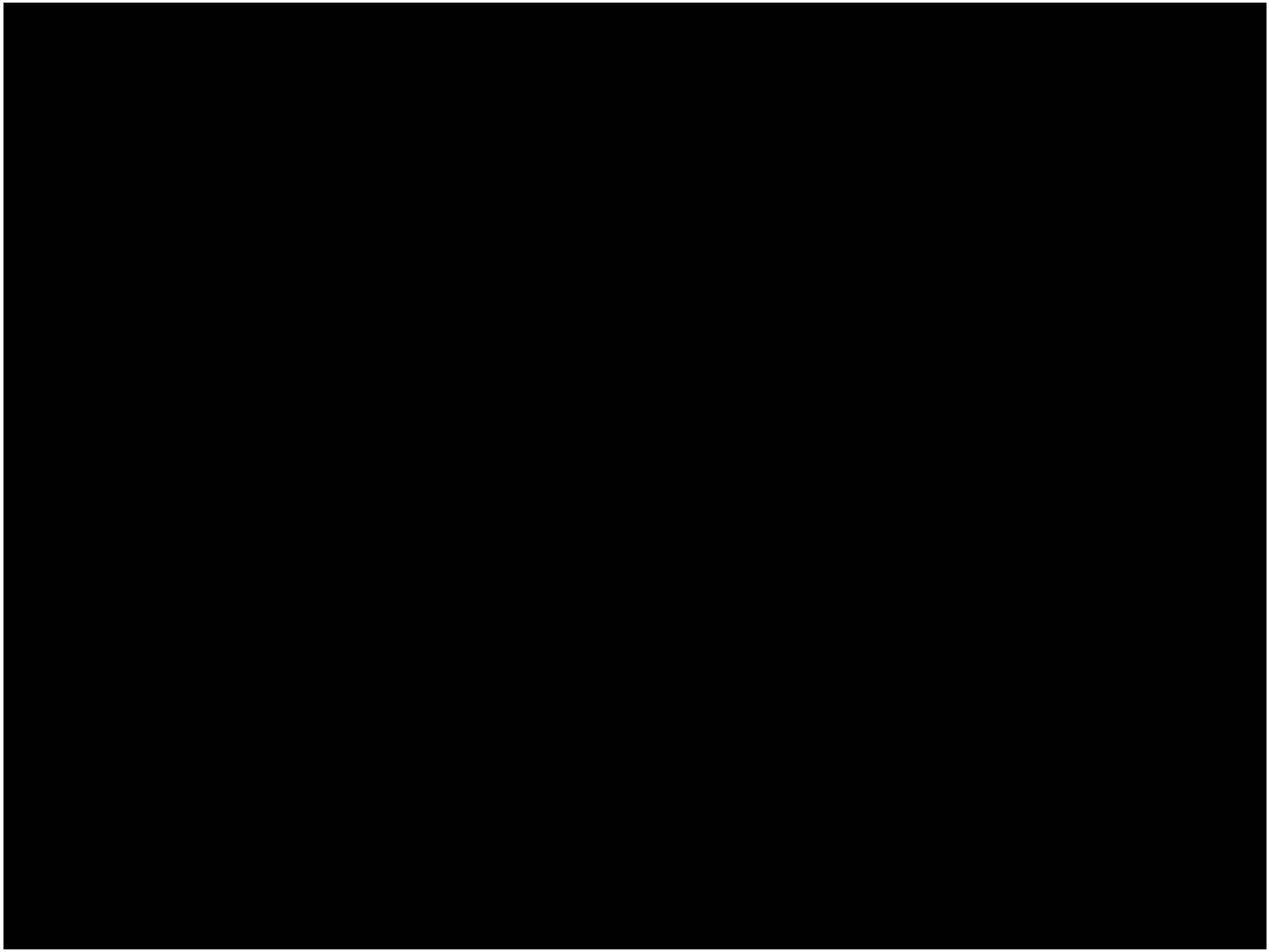
Alphabets themselves further distinguished into

Pure Alphabets: distinct letters for consonants & vowels

Abugida: vowels modify characters for consonants

Abjad: vowels are omitted as they are implied rather than being explicit

Arthur Conan Doyle,
The Adventure of the Dancing Men, in Return of SherlockHolmes (1903)



Granada Television 1984

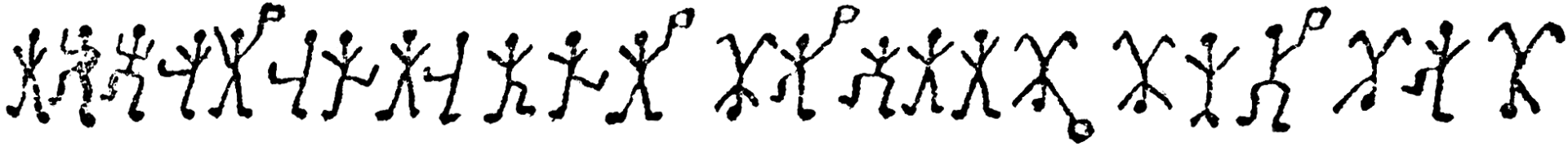
Criminal's (Abe Slaney staying at Elrige's Farm) first message

Criminal's second message

Criminal's third message

Elsie's response

Criminal's final message

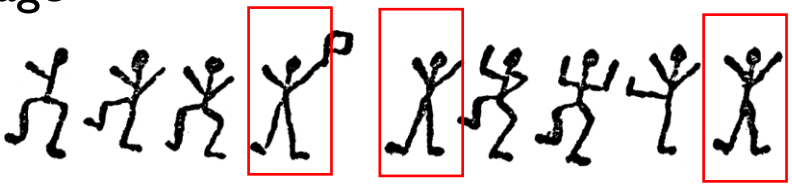


Sherlock Holmes' message to the criminal

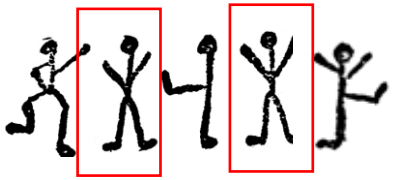
Criminal's (Abe Slaney staying at Elrige's Farm) first message

Criminal's second message

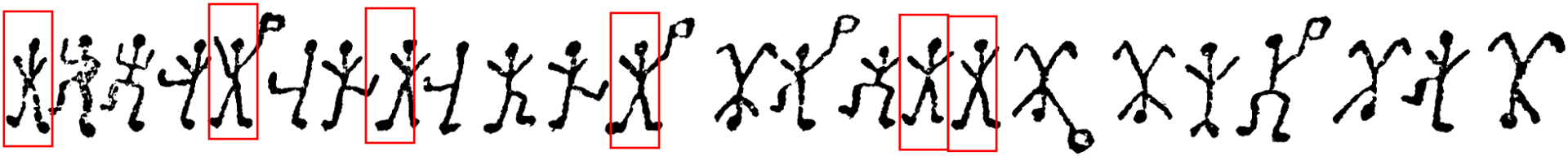
Criminal's third message



Elsie's response

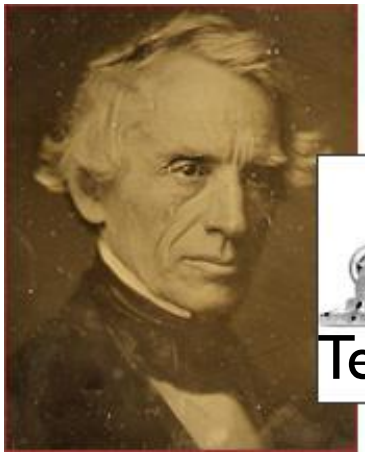


Criminal's final message

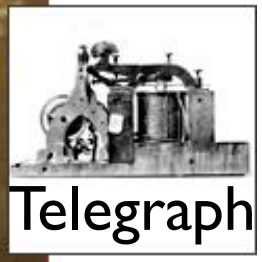


Sherlock Holmes' message to the criminal

Frequency of English Characters



Samuel Morse
1791-1872



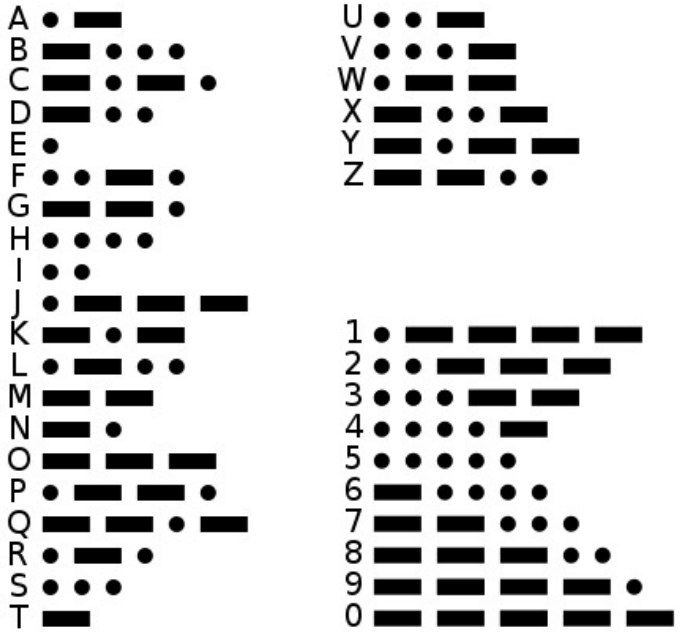
Telegraph

Highest frequency characters **ETAOIN SHRDLU**

To increase the efficiency of encoding, Morse code was originally designed so that the length of each symbol is approximately inverse to the frequency of occurrence of the character that it represents in text of the English language.

But Morse's character frequency counts are based on pieces of text considers the commonest words multiple times

Morse code



If one considers all distinct words so that word frequency does not produce artefacts, e.g., by considering words from a dictionary, the highest frequency characters are

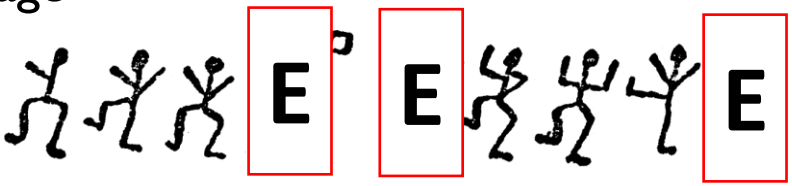
EARIOT NSLCUD

In either case, E is the commonest letter

Criminal's (Abe Slaney staying at Elrige's Farm) first message

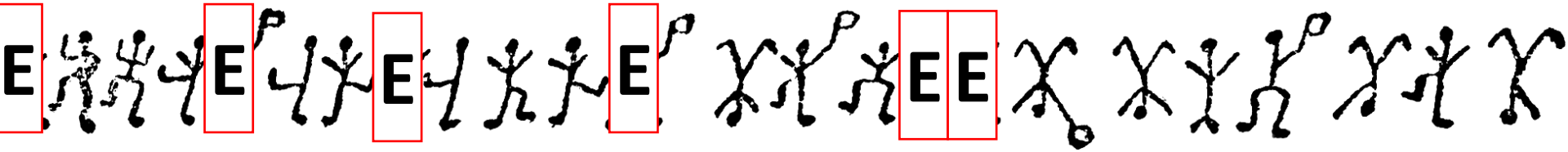
Criminal's second message

Criminal's third message



Elsie's response

Criminal's final message



Sherlock Holmes' message to the criminal

Criminal's (Abe Slaney staying at Elrige's Farm) first message

Criminal's second message

Criminal's third message

Elsie's response

There are no word boundaries – could the flags indicate word endings ?

Criminal's final message

Sherlock Holmes' message to the criminal

Criminal's (Abe Slaney staying at Elrige's Farm) first message

Criminal's second message

Criminal's third message

Elsie's response

No flags – single word!

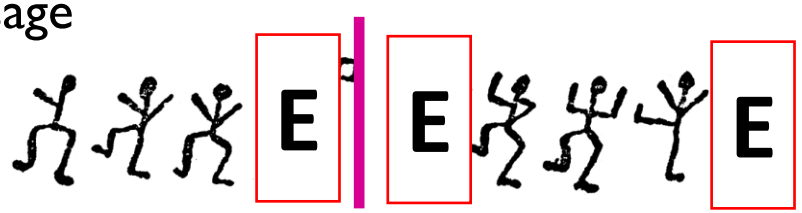
Criminal's final message

Sherlock Holmes' message to the criminal

Criminal's (Abe Slaney staying at Elrige's Farm) first message

Criminal's second message

Criminal's third message



Elsie's response

Possible English words of the form * E * E * are

SEVER
LEVER
NEVER

Criminal's final message

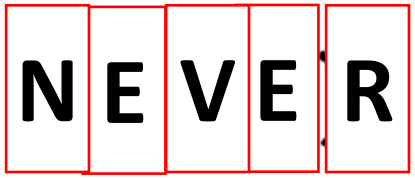
Sherlock Holmes' message to the criminal

Criminal's (Abe Slaney staying at Elrige's Farm) first message

Criminal's second message

Criminal's third message

Elsie's response

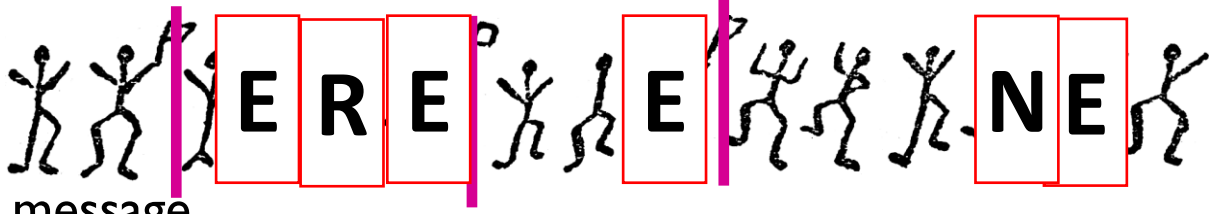


Assuming that this is a conversation between two people, very likely that it is NEVER

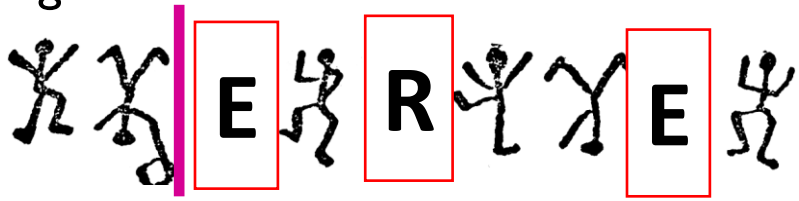
Criminal's final message

Sherlock Holmes' message to the criminal

Criminal's (Abe Slaney staying at Elrige's Farm) first message

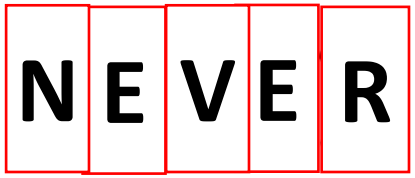


Criminal's second message



Criminal's third message

Elsie's response

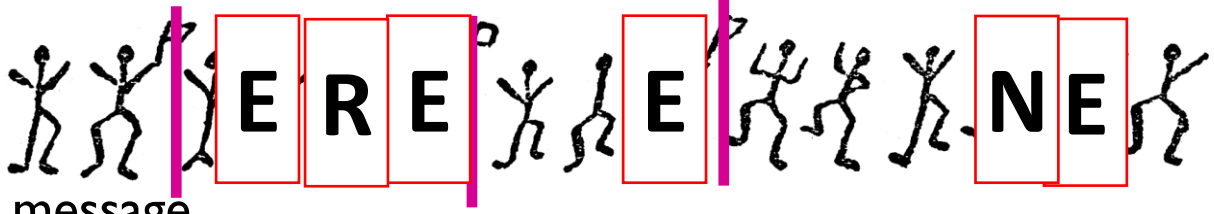


Gives symbols for N, V, R

Criminal's final message

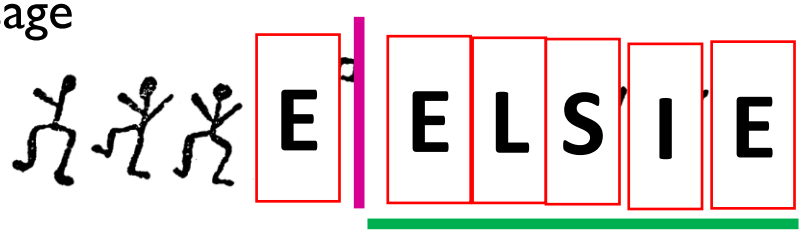
Sherlock Holmes' message to the criminal

Criminal's (Abe Slaney staying at Elrige's Farm) first message



Criminal's second message

Criminal's third message



Elsie's response

E * * * E word occurs twice – is this a name ? ELSIE?

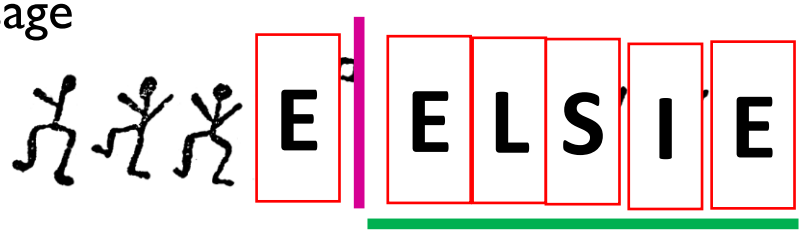
Criminal's final message

Sherlock Holmes' message to the criminal

Criminal's (Abe Slaney staying at Elrige's Farm) first message

Criminal's second message

Criminal's third message



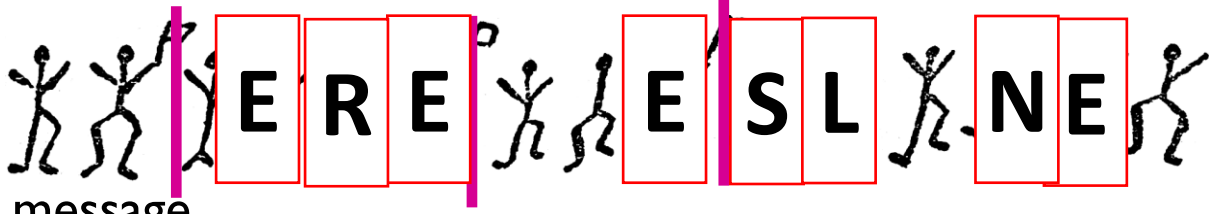
Elsie's response

Gives symbols for L, S, I

Criminal's final message

Sherlock Holmes' message to the criminal

Criminal's (Abe Slaney staying at Elrige's Farm) first message

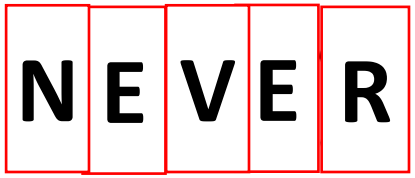


Criminal's second message

Criminal's third message

* * * E may be an instruction.
COME ?

Elsie's response



Criminal's final message

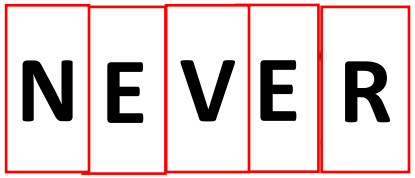
Sherlock Holmes' message to the criminal

Criminal's (Abe Slaney staying at Elrige's Farm) first message

Criminal's second message

Criminal's third message

Elsie's response



Gives symbols for C, O, M

Criminal's final message

Sherlock Holmes' message to the criminal

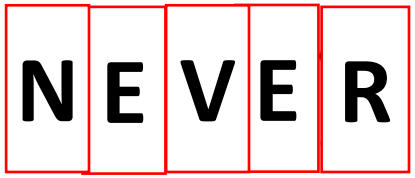
Criminal's (Abe Slaney staying at Elrige's Farm) first message

"AM HERE" ?

Criminal's second message

Criminal's third message

Elsie's response



Criminal's final message

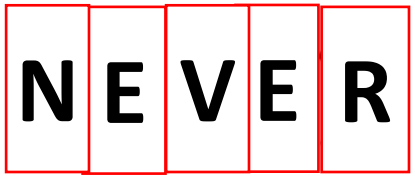
Sherlock Holmes' message to the criminal

Criminal's (Abe Slaney staying at Elrige's Farm) first message

Criminal's second message

Criminal's third message

Elsie's response



Gives symbols for A, H

Criminal's final message

Sherlock Holmes' message to the criminal

Criminal's (Abe Slaney staying at Elrige's Farm) first message

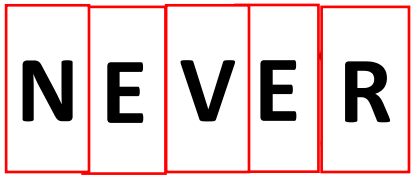
Criminal's second message

A* ELRI*ES
A* = AT ?

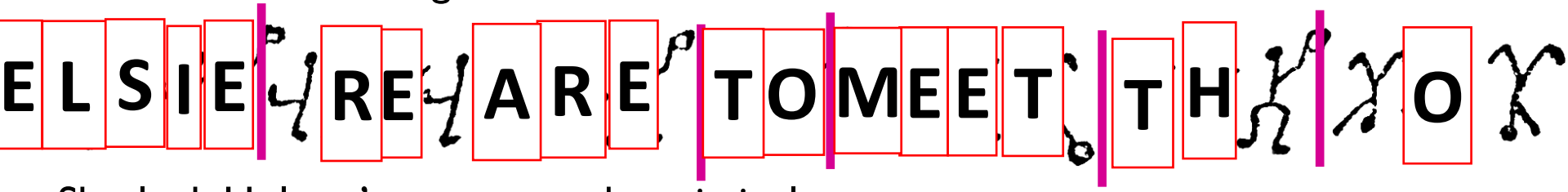
Criminal's third message

Gives T

Elsie's response



Criminal's final message



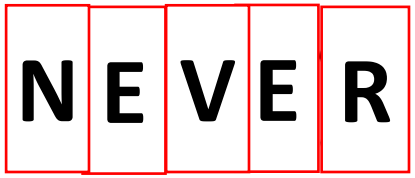
Sherlock Holmes' message to the criminal

Criminal's (Abe Slaney staying at Elrige's Farm) first message

Criminal's second message

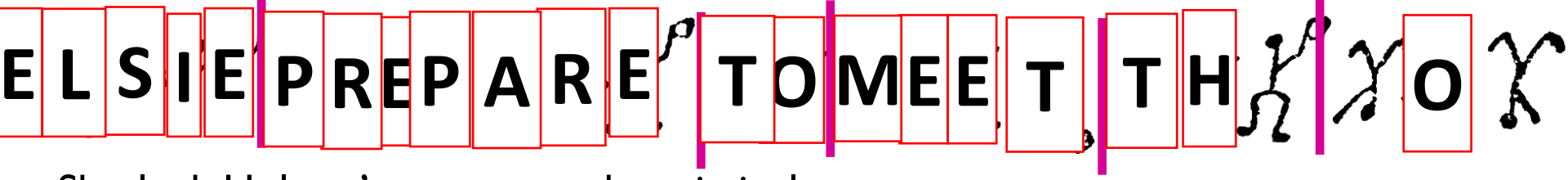
Criminal's third message

Elsie's response



* R E * A R E
PREPARE ?
Gives P

Criminal's final message



Sherlock Holmes' message to the criminal

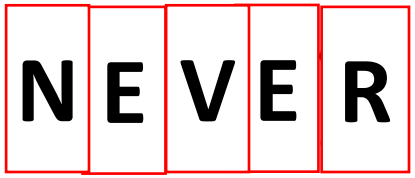
Criminal's (Abe Slaney staying at Elrige's Farm) first message

Criminal's second message

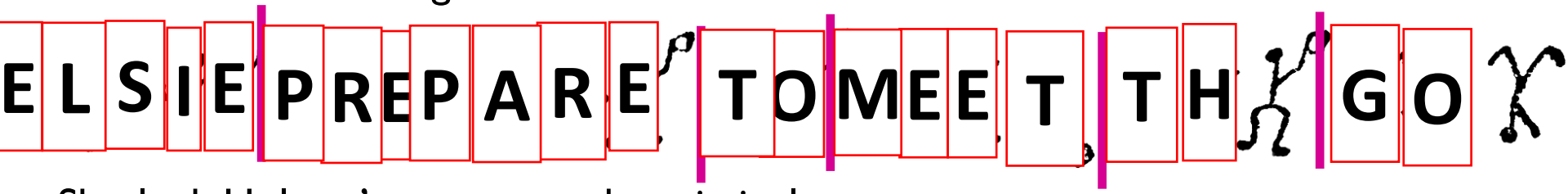
ELRI*ES
ELRIGES ?
Gives G

Criminal's third message

Elsie's response



Criminal's final message



Sherlock Holmes' message to the criminal

Criminal's (Abe Slaney staying at Elrige's Farm) first message

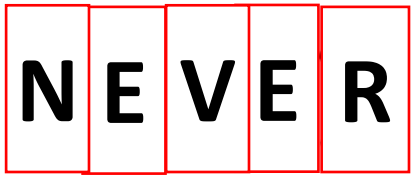
Criminal's second message

SLANE* & TH*

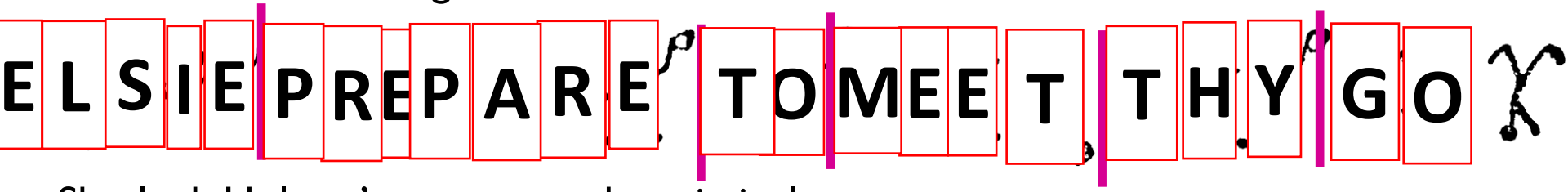
SLANEY & THY
Or
SLANEE & THE?

Criminal's third message

Elsie's response



Criminal's final message



Sherlock Holmes' message to the criminal

Criminal's (Abe Slaney staying at Elrige's Farm) first message

A M H E R E A E S L A N E Y

Criminal's second message

A T E L R I G E S

Criminal's third message

C O M E E L S I E

Elsie's response

N E V E R

GO*

GOD?

Criminal's final message

E L S I E P R E P A R E T O M E E T T H Y G O D

Sherlock Holmes' message to the criminal

Criminal's (Abe Slaney staying at Elrige's Farm) first message

A M H E R E A B E S L A N E Y

Criminal's second message

A T E L R I G E S

A*E is a first name

ABE?

Criminal's third message

C O M E E L S I E

Elsie's response

N E V E R

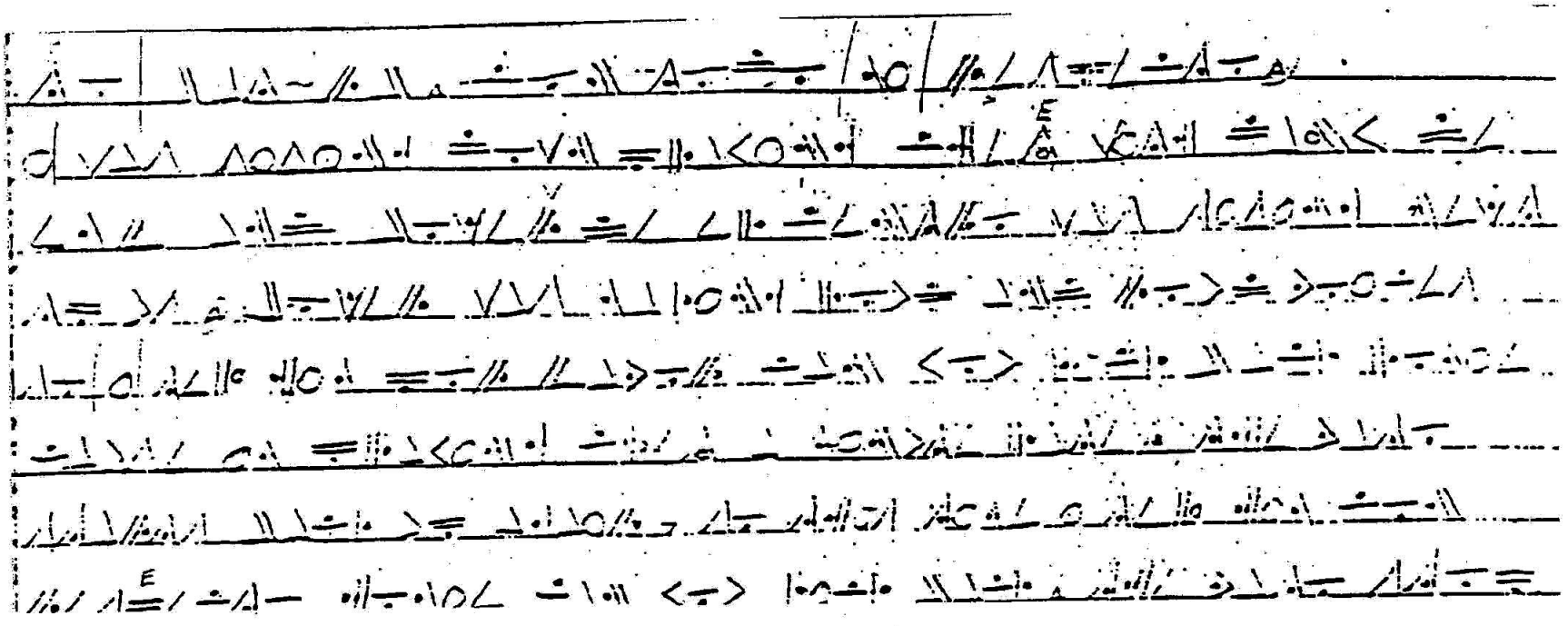
Criminal's final message

E L S I E P R E P A R E T O M E E T T H Y G O D

Sherlock Holmes' message to the criminal

Substitution ciphers in real life

A typical example of coded messages found in prisons in California sent to Stanford's Statistics Department for decryption



How to find the unknown function f such f : {code space} → {English alphabet}

Markov chain Monte Carlo

A standard approach to deciphering is to use the statistics of written English to guess at probable choices for f, try these out, and see if the decrypted messages make sense.

The statistics is obtained from a text corpus. The proportion of consecutive text symbols from x to y gives a matrix M(x; y) of 1st order transitions from one symbol to another..
Then the plausibility of the mapping f

$$Pl(f) = \Pi_i M(f(s_i), f(s_{i+1}))$$

where s_i runs over consecutive symbols in the coded message.

Mappings f which have high values of Pl(f) are good candidates for deciphering.
To maximize the plausibility we can run the following Markov Chain Monte Carlo algorithm

- Start with a preliminary guess, say f.
- Compute $\mathrm{Pl}(f)$.
- Change to f_* by making a random transposition of the values f assigns to two symbols.
- Compute $\mathrm{Pl}(f_*)$; if this is larger than $\mathrm{Pl}(f)$, accept f_*.
- If not, flip a $\mathrm{Pl}(f_*)/\mathrm{Pl}(f)$ coin; if it comes up heads, accept f_*.
- If the coin toss comes up tails, stay at f.

“Reps do it”

The space of possible f s is extremely large – can the Monte Carlo guided random walk achieve the “correct” f in reasonable time, or at all?

```
ENTER HAMLET HAM TO BE OR NOT TO BE THAT IS THE QUESTION WHETHER TIS
```

The text is scrambled at random and the Monte Carlo algorithm run

```
 100 ER ENOHDLAE OHDLO UOZEOUNORU O UOZEO HD OITO HEOQSET IUROFHE HENO ITORUZAEN
 200 ES ELOHRNDE OHRNO UOVEOULOSU O UOVEO HR OITO HEOQAET IUSOPHE HELO ITOSUVDEL
 300 ES ELOHANDE OHANO UOVEOULOSU O UOVEO HA OITO HEOQRET IUSOFHE HELO ITOSUVDEL
 400 ES ELOHINME OHINO UOVEOULOSU O UOVEO HI OATO HEOQRET AUSOWHE HELO ATOSUVMEL
 500 ES ELOHINME OHINO UODEOULOSU O UODEO HI OATO HEOQRET AUSOWHE HELO ATOSUDMEL
 600 ES ELOHINME OHINO UODEOULOSU O UODEO HI OATO HEOQRET AUSOWHE HELO ATOSUDMEL
 900 ES ELOHANME OHANO UODEOULOSU O UODEO HA OITO HEOQRET IUSOWHE HELO ITOSUDMEL
1000 IS ILOHANMI OHANO RODIORLOSR O RODIO HA OETO HIOQUIT ERSOWHI HILO ETOSRDMIL
1100 ISTILOHANMITOHANOT ODIO LOS TOT ODIOTHATOEROTHIOQUIRTE SOWHITHILOTEROS DMIL
1200 ISTILOHANMITOHANOT ODIO LOS TOT ODIOTHATOEROTHIOQUIRTE SOWHITHILOTEROS DMIL
1300 ISTILOHARMITOHAROT ODIO LOS TOT ODIOTHATOENOTHIOQUINTE SOWHITHILOTENOS DMIL
1400 ISTILOHAMRITOHAMOT OFIO LOS TOT OFIOTHATOENOTHIOQUINTE SOWHITHILOTENOS FRIL
1600 ESTEL HAMRET HAM TO CE OL SOT TO CE THAT IN THE QUENTIOS WHETHEL TIN SOCREL
1700 ESTEL HAMRET HAM TO BE OL SOT TO BE THAT IN THE QUENTIOS WHETHEL TIN SOBREL
1800 ESTER HAMLET HAM TO BE OR SOT TO BE THAT IN THE QUENTIOS WHETHER TIN SOBLER
1900 ENTER HAMLET HAM TO BE OR NOT TO BE THAT IS THE QUESTION WHETHER TIS NOBLER
2000 ENTER HAMLET HAM TO BE OR NOT TO BE THAT IS THE QUESTION WHETHER TIS NOBLER
```

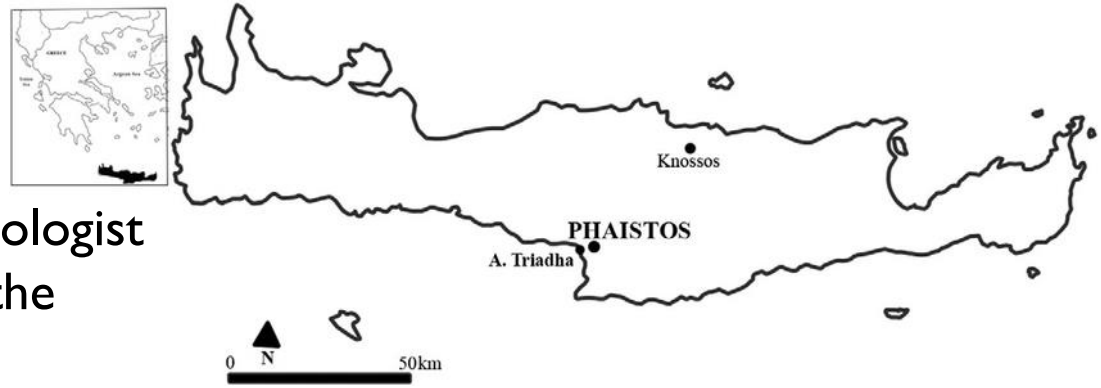
The simple optimization process converges to original text in a few thousand steps

Substitution ciphers in real life

Decipherment of the typical example of coded messages found in prisons in California sent to Stanford's Statistics Department for decryption

to bat-rb. con todo mi respeto. i was sitting down playing chess with danny de emf and boxer de el centro was sitting next to us. boxer was making loud and loud voices so i tell him por favor can you kick back homie cause im playing chess a minute later the vato starts back up again so this time i tell him con respecto homie can you kick back. the vato stop for a minute and he starts up again so i tell him check this out shut the f**k up cause im tired of your voice and if you got a problem with it we can go to celda and handle it. i really felt disrespected thats why i told him. anyways after i tell him that the next thing I know that vato slashes me and leaves. dy the time i figure im hit i try to get away but the c.o. is walking in my direction and he gets me right dy a celda. so i go to the hole. when im in the hole my home boys hit doxer so now "b" is also in the hole. while im in the hole im getting schoold wrong and

Phaistos Disk



Found in 1908 by an Italian archaeologist Louis Pernier in connection with the excavations of Phaistos in Crete.

features 241 tokens, comprising 45 distinct signs, apparently made by pressing hieroglyphic "seals" into the soft clay - spiraling clockwise toward the center

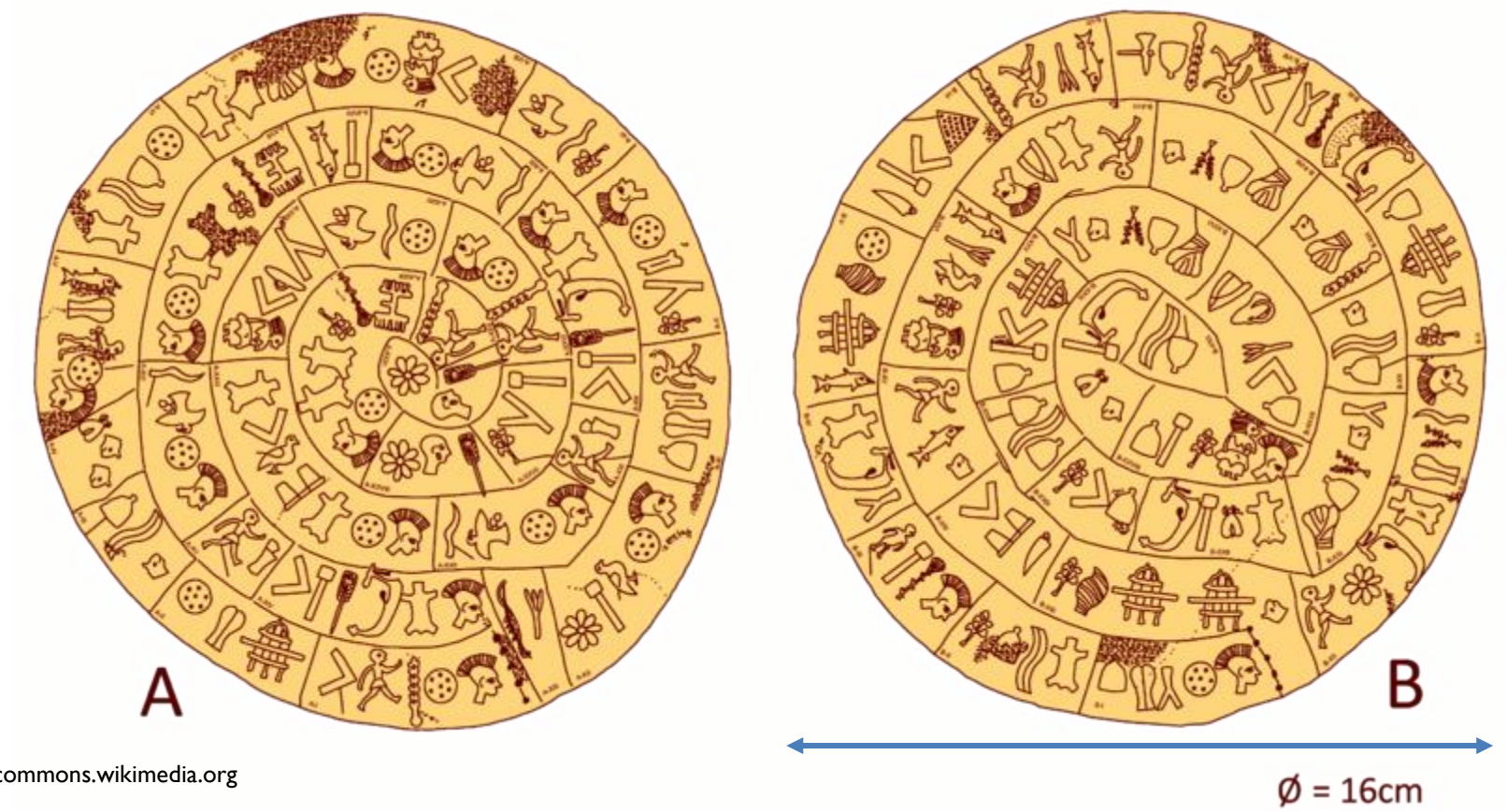


Image:commons.wikimedia.org

Phaistos Disk sign statistics

Sign				Frequency		
No.	Glyph	Font	Name	A	B	A+B
02			PLUMED HEAD	14	5	19
07			HELMET	3	15	18
12			SHIELD	15	2	17
27			HIDE	10	5	15
18			BOOMERANG	6	6	12
01			PEDESTRIAN	6	5	11
23			COLUMN	5	6	11
29			CAT	3	8	11
35			PLANE TREE	5	6	11
25			SHIP	2	5	7
13			CLUB	3	3	6
24			BEEHIVE	1	5	6
26			HORN	5	1	6
33			TUNNY	2	4	6
40			OX BACK	3	3	6
45			WAVY BAND	2	4	6
08			GAUNTLET	1	4	5
22			SLING	0	5	5
31			EAGLE	5	0	5
06			WOMAN	2	2	4
10			ARROW	4	0	4
36			VINE	0	4	4
37			PAPYRUS	2	2	4
38			ROSETTE	3	1	4
39			LILY	1	3	4
19			CARPENTRY PLANE	3	0	3
32			DOVE	2	1	3
34			BEE	1	2	3
03			TATTOOED HEAD	2	0	2
09			TIARA	0	2	2
14			MANACLES	1	1	2
16			SAW	0	2	2
20			DOLIUM	0	2	2
21			COMB	2	0	2
28			BULLS LEG	2	0	2
41			FLUTE	2	0	2
04			CAPTIVE	1	0	1
05			CHILD	0	1	1
11			BOW	1	0	1
15			MATTOCK	0	1	1
17			LID	1	0	1
30			RAM	0	1	1
42			GRATER	0	1	1
43			STRAINER	0	1	1
44			SMALL AXE	1	0	1

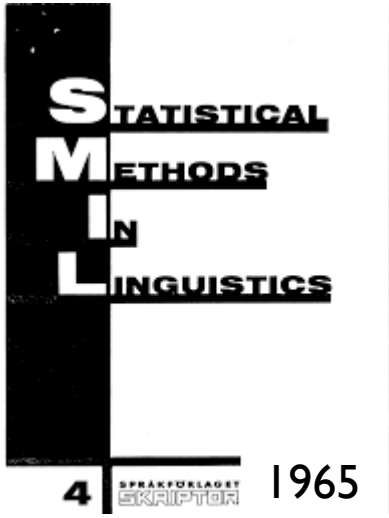
Inferring the signary size

The frequency of occurrence of symbols on the Phaistos Disc.
(Reference number of symbol in Evans's list)

	44																			
	43	41																		
	42	28																		
	30	21				40														
	17	20				45														
	15	16	38	39	36	33				35										
	11	14	34	37	31	26				29										
	5	9	32	10	22	24				23				46						
	4	3	19	6	8	13	25			1	18			27		12	7	2		
t=0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

We make the general assertion that, in a small sample of an alphabetic or syllabic writing system consisting of L_1 characters of M_1 different kinds, the probable number of symbols in the alphabet or syllabary is, subject to various restrictions discussed below, given approximately by $L_1^2/(L_1-M_1) - L_1$.

Alan L Mackay



⇒ If the Phaistos disk is genuine (the writing system follows the same general frequency distribution as other known alphabetic/syllabic systems), about 10 more symbols in the signary