

Deciphering Scripts

LING 106

January 28, 2009

1. DISTRIBUTIONAL ANALYSIS AND DECIPHERING SCRIPTS

General goal: given a string of symbols, break it into words and determine the meanings of those words.

Specifically:

- Symbols = phonemes; task = understanding a language
- Symbols = glyphs; task = reading a written text
- Symbols = ciphertext; task = breaking codes

2. WHAT IS A CIPHER?

For our purposes, a *cipher* is a secret writing system: a system in which a sender, Alice, can express a message in such a way that its intended recipient, Bob, can read it, but a casual observer (or even a dedicated one) such as Carol cannot.¹

- The message to be enciphered is called the *plaintext*. The result of the enciphering is called the *ciphertext*.
- The goal is to apply an *effective procedure* to the plaintext to produce the ciphertext. An effective procedure **E**, for our purposes,
 - consists of a finite number of operations performed in a definite order;
 - stops, producing an output;
 - is reversible via a procedure **E**⁻¹.
- An ineffective procedure: “Replace each word in the ciphertext with the Greek letter beta.” Thus, the message **meet me tonight at the library** becomes **β β β β β β**. It’s certainly hard for Carol to figure out the original message; unfortunately, it’s also going to be really hard for Bob. (Reversibility is important!)

The catch: if it’s possible for Bob to calculate **E**⁻¹(ciphertext) once he gets the ciphertext, it’s also possible for Carol to calculate it once she gets the ciphertext. Thus an additional need: unless you already know **E** and **E**⁻¹, it should be hard to figure them out.

¹ Note that one of the best ways to keep someone from reading a message is to disguise the fact that it’s a message at all.

3. SOME EXAMPLES OF CIPHERS

Rearrangement of the plaintext, e.g.:

- E (plaintext): reverse each word in *plaintext*.
- E^{-1} (ciphertext): reverse each word in *ciphertext*.

Example: **meet me tonight at the library** →
 TEEM EM THGINOT TA EHT YRARBIL

This will fool Carol for about two seconds.

Caesar shifting:

- E : replace each letter of the text with the letter n ahead of it in the alphabet, for some predetermined n . (Wrap around, so that the letter after Z is A.)
- E^{-1} : replace each letter of the text with the letter n behind it in the alphabet.

(Note: because English has 26 letters, the case where $n = 26/2 = 13$ is a special case, because $E = E^{-1}$. It's often called *ROT13*.)

Example: if $n = 3$, then **a** becomes **D**, **b** becomes **E**, ..., **z** becomes **C**.
meet me tonight at the library → **PHHW PH WRQLJKW DW WKH OLEUDUB**

This will fool Carol for about five seconds. If she doesn't know the n , she only has to try 25 different possibilities and see which one produces a sensible message.

Exercise: decipher the following Caesar Shift ciphertext, $n = 17$.
KYZJ NRJ GIVKKP VRJP KF UVTZGYVI

Caesar shifting is a special case of a “monoalphabetic substitution cipher”: it's a one-to-one and onto function from letters to letters.

- More generally, instead of shifting, we can assign the letters arbitrarily:

Plaintext: ABCDEFGHIJKLMN**OP**QRSTUVWXYZ
Ciphertext: UONISLJZTKMXBQWFCVYEARHDPG

- Now:

E : replace each letter with the one below it in the above pair of lines.
 E^{-1} : replace each letter with the one above it in the above pair of lines.

meet me tonight at the library → **bsse bs ewqtjze ue ezs xtovuvp**

Exercise: decipher the following ciphertext using the above encryption.
EZTY HUY U LUTVXP OWVTQJ BSYUJS

There are also *polyalphabetic ciphers*: instead of mapping one letter to one letter, they map, e.g., a pair of letters to a pair of letters, or a pair of letters to a single number. For instance:

		Second Letter						
		A	B	C	D	E	F	...
First Letter	A	AA	AC	AF	AJ	AO	AU	
	B	AB	AE	AI	AN	AT	BA	
	C	AD	AH	AM	AS	AZ	BH	
	D	AG	AL	AR	AY	BG	BP	...
	E	AK	AQ	AX	BF	BO	BY	
	F	AP	AW	BE	BN	BX	CI	
	:				:			

		Second Letter						
		A	B	C	D	E	F	...
First Letter	A	1	3	6	10	15	21	
	B	2	5	9	14	20	27	
	C	4	8	13	19	26	34	
	D	7	12	18	25	33	42	...
	E	11	17	24	32	41	51	
	F	16	23	31	40	50	61	
	:				:			

For example, to encode the word **face**, we look up FA and see that it translates to 16 (on the first table) or AP (on the second), and CE translates to 26 or AZ; so **face** → **16 26** or **apaz**.

meet me tonight at the library →

me et me to ni gh ta tt he li br ar yx
FK LJ FK TR JF DU HI WT CS HQ HG GO ZR

→ **FKLJ FK TRJFDUH IW TCS HQHGGOZR**

Note that in both cases, **me** maps to **FK** (even though the two **e**s in the first word seem to map to two different letters); this is a one-to-one and onto function from pairs of letters to pairs of letters.

Extraordinarily Optional Exercise:
 decipher the following ciphertext using the above encryption.

N EAKKER ZQSQ FUO XDTWA K FUOZXFJO TCIIFXG CSJ NTOE

(Note: please do not attempt this exercise unless you really, really want to.)

Once again, as long as Bob has the same table as Alice, he can reverse the encoding fairly easily.

But how long will Carol be fooled by these ciphers?

- For the “simple substitution cipher”: If she tries to use brute force, instead of 26 options there are now $26! = 403,291,461,126,605,635,584,000,000 \approx 4.033 \times 10^{26}$ options. If Carol has a computer that will try a million possibilities every second, it’ll only take a little less than thirteen trillion years to try them all. (Then of course she’ll need to reach through each possibility to see whether it’s a sensible message, something that was much easier when there were 25 possibilities.)

- For the polyalphabetic substitution cipher: If she tries to use brute force, she now has $(26^2)! \approx 1.8837 \times 10^{1621}$ options. You're welcome to work out how long this will take, though you'll want phrases like "heat death of the universe".

So these ciphers are pretty secure, right?

4. WRONG: WHY THESE CIPHERS CAN BE BROKEN

What makes these ciphers breakable is that (a) the ciphertext mirrors certain key facts about the plaintext, and (b) language is ergodic.

- *The ciphertext mirrors facts about the plaintext*

If word breaks are included, simple substitution ciphers can often be broken by looking for certain common words or uncommon patterns. One-letter words are likely to be *a* or *I*; a frequently-repeated three-letter word is likely to be *the*. Only one English word has the unusual letter pattern **ABCABDCEDB**, so if that word appears in the text, Carol immediately knows five of the twenty-six mappings in **E**.

But even if word breaks aren't included, the other thing preserved in a monoalphabetic cipher is letter frequency. For instance, the frequency of the letters in **meet me tonight at the library** and its enciphered counterpart are:

meet me tonight	bsse bs ewqtjze
at the library	ue ezs xtovuvp
5: t	5: E
4: e	4: S
2: ahimr	2: BTUVZ
1: bglnoy	1: JOPQWX

This in and of itself is only of limited usefulness, but combine it with...

- *Language is ergodic*

That is to say, if you take a text of sufficient length, the frequency of letters within that text will roughly match the frequency of letters in the language.

4.1. How this helps

The letters of English ordered roughly by frequency:

ETAIN OSRHD LCUMF WGPYB VKXJQ Z

Specific frequencies:

Letter	%	Letter	%	Letter	%	Letter	%
A	8.13	H	5.63	O	7.15	V	1.05
B	1.43	I	7.33	P	1.90	W	2.09
C	2.93	J	0.17	Q	0.08	X	0.19
D	4.31	K	0.66	R	5.99	Y	1.69
E	12.72	L	3.85	S	6.44	Z	0.07
F	2.15	M	2.51	T	9.76		
G	1.95	N	7.29	U	2.52		

(Note: The above ordering and frequencies are based on personal calculations from an admittedly smallish corpus. This comes pretty close to matching other estimates, though, which also vary depending on size and source of corpus. Perhaps the most familiar ordering of the twelve most common letters is **ETAOIN SHRDLU**.)

Or, ordered by frequency:

%	Letter	%	Letter	%	Letter	%	Letter
12.72	E	5.99	R	2.15	F	0.65	K
9.75	T	5.63	H	2.08	W	0.19	X
8.12	A	4.31	D	1.95	G	0.16	J
7.33	I	3.85	L	1.89	P	0.08	Q
7.29	N	2.92	C	1.68	Y	0.06	Z
7.14	O	2.52	U	1.42	B		
6.44	S	2.51	M	1.04	V		

Consequence: If a plaintext is enciphered with a monoalphabetic cipher function δ , then we expect about 13 percent of the plaintext to be the letter **e**, and thus about 13 percent of the ciphertext should be $\delta(\mathbf{e})$.

Simply replacing the most frequent letter in the ciphertext with **E**, the second most frequent letter with **T**, and so forth will probably get you gibberish, but the longer the ciphertext, the closer it'll get to English, and the easier it'll be to make small fixes.

For instance, suppose you have the first 64 letters of a particular enciphered text, which begins:

U YFSNEVS TY ZUAQETQJ SAVWFS—EVS YFSNEVS WL NWBBAQTYB.

Substituting **e** for the most common letter in the first 64, **t** for the next-most common, and so forth, you get:

U HFEPTDE AH LUMOTAOG EMDIFE—TLE HFEPTDE IN PISSMOAHS.

That's a far cry from intelligible—though the three-letter word after the dash already looks like it might be **the**. But with more letters of ciphertext, things start to improve:

Letters	Decryption
64	U HFEPTDE AH LUMOTAOG EMDIFE—TLE HFEPTDE IN PISSMOAHS.
128	U NLECTSE AN HUMOTAOG EMSILE—THE NLECTSE ID CIRMOANR.
256	S RLEHTNE IR DSWOTIOG EWNAL—TDE RLEHTNE AF HAMMWOIRM.
512	N ODELTRE AO HNWSTASG EWRIDE—THE ODELTRE IF LIUUWSAOU.
1024	O IMECARE TI HOWSATSG EWRNME—AHE IMECARE NF CNLLWSTIL.
2048	O AUECIRE TA HOWSITSG EWRNUE—IHE AUECIRE NF CNMMWSTAM.
4096	N SWECARE IS HNMOAIOG EMRTWE—AHE SWECARE TF CTUUMOISU.
8192	N SWECTRE AS HNUOTAOG EURIWE—THE SWECTRE IF CIMMUOASM.

At this point, it's perhaps possible to interpret the sentence: *A spectre is haunting Europe—the spectre of Communism*. (It'll never get perfect: **w**, representing **o**, happens to be quite common in the ciphertext; in fact, **o** is more common in this text than in English in general. That's why pure frequency analysis keeps guessing that it's representing a more common letter, i.e. **i**, **n**, or **t**.)

Note that the polyalphabetic cipher described above does not preserve letter frequency...

meet me tonight	fk lj fk tr jfd uh
at the library	iw tcs hqh ggozr
5: t	3: FH
4: e	2: GJKRT
2: ahimr	1: CDILOQS UWZ
1: bglnoy	

...but it's possible to use the frequency of bigrams (i.e., pairs of letters) in a manner similar to the one above, as long as one has a sufficiently long text.

5. A FINAL DIGRESSION: HOW TO MAKE BETTER CIPHERS

- *Use shorter texts.* Given that the following message is encrypted with a simple substitution cipher, determine the encryption method **E** and the plaintext:

QTRPLU

Let me know how that goes.

- *Change encryption frequently.* Suppose you're sending an encrypted message each day. If that message is only 100 letters long, each one may be hard to break using frequency analysis; but three months later, someone intercepting the messages will have 9000 letters to work with. But suppose you use the following:

E(text):

- convert the first paragraph of the first editorial in the day's *New York Times* into numbers: A = 0, B = 1, C = 2, ..., Z = 25.
- Pair up the numbers with the letters in your message.
- Caesar shift each letter, using n = the paired-up number.

For example, if your message is **meet me tonight at the library** and the first editorial on the *Times*'s Op-Ed page begins SENATOR MCCAIN'S CAMPAIGN HAS..., you would encode the text as:

<i>plaintext</i>	m	e	e	t	m	e	t	o	n	i	g	h	...
<i>key</i>	s	e	n	a	t	o	r	m	c	c	a	i	...
<i>n</i>	18	4	13	0	19	14	17	12	2	2	0	8	...
<i>ciphertext</i>	e	i	r	t	f	s	k	a	p	k	g	p	...

and transmit the message **eirt fs kapkgpg sv ttt lqhehrq**. This kind of code, called a *Vigenère cipher*, is actually not inherently hard to crack—again, with enough ciphertext. But since your key changes every day, the same message sent the next day might read **flii ds uzruumz lh uhp hisdiee**, and by changing the key daily, anyone intercepting your encoded messages will never get enough identically-encoded text to use to break your code.

Optional Exercise: What were the first five words of the *Times* editorial used to encode the latter message?

- *Use unbreakable encryption methods.* With the advent of computers, it's become easier to use brute force to decipher the kinds of substitution ciphers discussed above, especially when guided by frequency analysis. It's also become possible to encode messages in wholly different ways, using computationally intractable problems.

Of course, by now we're getting farther and farther from tools that bear on linguistic analysis. Back on track....

Linear B

or, deciphering an actual language

Note: quotations, transcriptions, translations, characters, etc. etc., taken from *The Decipherment of Linear B*, John Chadwick. Strongly recommended reading.

6. THE IDEAL: A ROSETTA STONE

In a best-case scenario, one has two identical texts in different writing systems, which can serve as a key. e.g., the Cypriot script, used to write Greek from the 6th to 3rd centuries BCE. Deciphered (1870s) with help from inscriptions in Cypriot and Phoenician, and in the Cypriot and Greek alphabets.

- Cypriot: syllabic, i.e. each symbol represents either a vowel or a consonant-vowel.
- Ill-suited to Greek (but used anyway)
 - k sound = γ, κ, χ (g, k, k^h); p sound = β, π, φ (b, p, p^h); t sound = δ, τ, θ (d, t, t^h)
 - Can't distinguish long and short vowels
 - Consonant clusters represented via "dead" vowels (cf. English to Japanese)
 - etc.

For example:

α	v	θ	ρ	ω	π	o	ζ	=	\ast	F	ϱ	ζ'	μ
a	n	t^h	r	\bar{o}	p	o	s	=	a	to	ro	po	se

- With enough correspondences, possible to equate the alphabets completely.

7. HOW TO SUCCEED IN TRANSLATION (BY REALLY, REALLY TRYING)

- Have a native speaker teach you the language
- Have some fairly small amount of bilingual inscription
- Have some fairly large amount of sufficiently varied inscription (e.g., the problem with Etruscan: lots of text, but it's all funerary inscriptions, and thus contains the same phrases repeatedly)
- Given two methods...
 - Careful analysis
 - Pure guesswork
 - ...choose the former.
- When making informed guesses:
 - Keep one's guesses conservative

- Have another text to check one's theory

8. THE LINEAR B TABLETS

- Sir Arthur Evans, 1900: excavation at Knossos, Crete reveals tablets with three kinds of writing.
 - 2000 – 1650 BCE: Pictorial (symbols for head, hand, star, arrow...)
 - 1750 – 1450 BCE: Linear A
 - 1450 – 1375 BCE: Linear B
- How to succeed in translation?
 - No speakers (of course)
 - No parallel texts

8.1. A few starting facts

- Part I of approach: comparisons to Cypriot writing (for which we know syllabic values).

Linear B	Cypriot	Value in Cypriot
┆	┆	<i>ta</i>
┆	┆	<i>lo</i>
┆	┆	<i>to</i>
μ	μ	<i>se</i>
┆	┆	<i>pa</i>
┆	┆	<i>na</i>
∧	∧	<i>ti</i>

- Some facts...
 - *se* ends many words in Cypriot-written Greek: *s* is a common word ending (and *e* is the “dead” vowel)
 - The corresponding Linear B character is rare as a word ending; nor does any other symbol show that kind of distribution.
 - Thus: Linear B is not used to write Greek.

(The latter fact matches archaeological conclusions: culture of Minoan civilization on Crete is wholly different from Mycenaean Greece.)

8.2. *The guesswork approach*

Ways to guess:

- Guess the language involved, e.g.
 - Greek (even though your results aren't very good)
 - Something hard to verify because of imperfect knowledge of the language or its connections (Basque, Etruscan)
- Make up your own language

8.2.1. *Some particularly bad results*

- Bedřich Hrozný (Hittite expert): draw comparisons between Linear B and Cypriot, Egyptian, Hittite, the Indus valley script, Cuneiform, Phoenician... Assume language is related to Hittite.

Result:

**Place of administration Hatahuâ: the palace has consumed all (?)
Place of administration Sahur(i)ta (is) a bad (?) field (?): this
(delivers in) tribute 22 (?) (measures), 6 T-measures of saffron
capsules**

Correct translation turns out to be:

**Thus the priestess and the key-bearers and the Followers and
Westreus (hold) leases: so much wheat 21.6 units**

The judgment of Chadwick:

“It is a sad story which recurs too often in the world of scholarship: an old and respected figure produces in his dotage work unworthy of his maturity, and his friends and pupils have not the courage to tell him so.”

- F.G. Gordon: assign Basque values to the characters, assign to each sign a pictographic value based on some rough resemblance, hope for the best.

Result: poetry.

**...the lord walking on wings the breathless path, the star-smiter,
the foaming gulf of waters, dogfish smiter on the creeping flower;
the lord, smiter of the horse-hide..., the dog climbing the path, the
dog emptying with the foot the water-pitchers, climbing the circling
path, parching the wine-skin...**

- And so on, and so on:
 - Assume the language is Greek, even though you don't know anything about the archaic forms of Greek
 - Assume the language is something Semitic
 - Ignore the tablets and translate the inscription on the rim of a jar (Ventris's judgment: the marks are just decorative doodling)
 - Georgiev: take the language to be archaic Greek with pre-Hellenic influence, and thus translate as Greek those things which work, and if it doesn't work, call it an unknown earlier language.

8.2.2. *Some much better educated guesswork and methodology*

- Ernst Sittig: Compare non-Greek Cypriot inscriptions to Linear B, speculate they are the same language, and correspond symbols based on frequency patterns.

Incorrect assumption that the languages were the same—but it's a method that would have worked if they were.

- Arthur Evans: Find certain “determinatives”, the equivalent of capital letters in English, which mark the next word as “religious” or “royal” or “place name”, etc.

One result: on a tablet with horsehead symbols, there is twice a smaller, maneless head preceded by the same two-sign word each time. Compare the signs to similar Cypriot symbols, the result is **po-lo**, much like the Greek *pōlos* ‘foal’.

Problem: since the language isn't Greek, that's not very helpful.

- Dr. Alice E. Kober: ask basic questions about the grammar, e.g.
 - Does it use different endings to express grammatical forms?
 - Is there a consistent “plural” inflection?
 - Does it distinguish genders?

Kober's results:

- Distinction between MALE and one class of animals vs. FEMALE, another class of animals, swords, etc. Thus, grammatical gender (and marks added to animals to indicate sex)
- “Kober's triplets”: many sets of words of the form
(sign₁+...+sign_n) + X + ending₁/ending₂/()

Evidence of inflection...?

9. MICHAEL VENTRIS

Michael Ventris (1922-1956), British architect, developed a full set of statistical tables of symbols, including frequency overall, initially, finally, etc. (Bennett's numbering will be used henceforth.)

9.1. Locational frequency

Some symbols common at the starts of words: **08** (previously, "royal" determinative), **61** (previously, "religious" determinative), **38**. These had previously been taken to be unpronounced determinatives.

- However: all three could occur word-internally, and **61** was also common word-finally, making it look unlikely that they were determinatives. More likely conclusion: vowel-only symbols in a syllabary.

Recall that a syllabary will have mostly CV symbols and a few V symbols. Thus: word-internal vowels will usually be represented by a CV syllable, but word-initial vowels must be represented by a V syllable. Examples—

alphabetical **a-l(a)-fa-be-ti-ca-l(a)** = V-CV-CV-CV-CV-CV-CV
individual **i-n(i)-di-vi-du-a-l(a)** = V-CV-CV-CV-CV-V-CV

So V *can* appear word-internally, but will appear much more commonly word-initially.

Conclusion: 08, 38, maybe also 61, are V syllables in a syllabary.

78 a common final sign, e.g.

36-14-12-41 70-27-04-27 51-80-04-78
11-02-70-27-04-27-78 77-60-40-11-02-78 61-39-58-70-78...

Additional evidence of its independence:

- Appears often as a final sign in lists
- Seems separable: compare **70-27-04-27** with **11-02-[70-27-04-27]-78**

Speculation: -78 corresponds to *and* (cf Latin suffix *-que*).

(Other prefixes and suffixes similarly identifiable.)

9.2. *Correlating sounds of syllables*

Suppose that roots end in consonants, inflectional suffixes start with vowels, e.g. Latin

MASCULINE	FEMININE
domin+us	domin+a
bon+us	bon+a
serv+us	serv+a

Then these would be written syllabically as **do-mi-nu-s(e)** and **do-mi-na**: that is, they would look like Kober triplets, with the roots the same and the final consonant plus suffix differing.

More importantly: if these were written syllabically as X-Y-P-Q and X-Y-Z, and another word is J-K-L-Q, then we might have:

do	mi	nu	s(e)	do	mi	na	se	r(e)	vu	s(e)
X	Y	P	Q	X	Y	Z	J	K	L	Q

Thus, even without knowing the actual syllabic values, we can guess that P and Z start with the same consonant, and P and L end with the same vowel.

Conclusion: A chart can be made of syllable correspondences.

Vowel:	I = <i>i</i> ?	II = <i>o</i> ?	III = <i>e</i> ?	IV	V = <i>a</i> ?
Pure vowel	61	—	—	—	08
Consonant I	—	—	59	—	57
II	40	10	75	42	54
III	39	—	(39)	—	03
IV	46	36	(46)	—	(57)

...and so on.

9.3. *The final step: guessing at some words*

Suppose that certain common nouns are the names of major nearby towns: Amnisos, Knossos, Tulissos, and variants of those with suffixes are adjectival forms, etc. etc. Then line up the syllables this gives with other syllables, translate certain other words, etc. etc.

Conclusion: Linear B really *is* Greek!

Only it happens to be archaic Greek, with certain final letters omitted—including the *s* that supported the initial conclusion that it wasn't Greek.