Machine Identification of Words

If most of the letters of a simple substitution cryptogram have been determined there may be several possibilities for those words which contain unidentified letters. For example, we may have the word N E . T , with substitutes found for all the letters except S and W. The correct choice will usually present no problems to the human cryptanalyst, but would be rather difficult for a machine. The following suggestion is proposed as a first faltering step in the direction of doing this by machine.

Assume each word stored in the machine has associated with it a symbol showing the category or field in which it is likely to occur. Most of the words will have a general symbol meaning that they could occur in almost any field. When a cryptogram has been solved as far as possible, it is broken up into word lengths by the method described in the attached 'Machine Separation of Words'. Then the text is scanned for incomplete words and these fragments are compared with the words in the machine's dictionary for possible words these fragments may represent (taking into account the fact that all missing letters must be found among the set of letters still unidentified). If only one such word can be found it is printed out at this point and the new letters so found added in other places in the message if they occur. If two or more words are found which might fit at this point, a frequency count of the category symbols of the other words of the message is taken and compared with the category symbols of the candidate words. The word with the symbol having the highest frequency (except the general symbol) is taken as the text word.

As an example, suppose we have the following message :

T O C O M M A N D E R S E C O N D A R M Y E N E M Y P A T R O L S
H A V E P E N E T R A T E D O U R R I G H T . L A N K T O A
T W O M I L E D E P T H

The word . L A N K , with B, F, J, Q, X and Z still unidentified might be B L A N K (Gen.) or F L A N K (Mil.). On counting the category symbols, the machine finds 4 words from the Military category and 12 words from the General category. It therefore selects F L A N K as the correct word.

This is an extremely simple illustration. The number of categories may have to be increased greatly, the words in the General category may have to be arranged in several ways, provision may have to be made for using pairs of consecutive words or groups of longer length, etc.

Machine Separation of Words

Assume a machine has stored in it a dictionary of the most common words in a particular type of traffic, and a set of instructions as follows : Take the first letter of the message and look through the dictionary to see whether there is a word beginning with this letter. If there is, proceed to the second letter. If there is a word in the dictionary beginning with these two letters proceed to the third letter, etc. Also note at every point whether the letters up to this point represent a complete word. If there is a complete word in the dictionary corresponding to the first r letters and no word beginning with letters 1 to r + 1, print out the word represented by letters 1 to r and repeat the same process beginning with the (r + 1)st letter. If there are words beginning with letters k to k + r (but not a complete word), and no word beginning with letters k to k + r + 1, go back to the last complete word which has not been used and repeat the process.

As an example suppose we have the message A T A N K E R F L Y I N G T H E F I N N I S H F L A G W A S S I G H T E D A T Z E R O S E V E N H U N D R E D . Let the symbol x mean 'no word beginning with these letters is in the dictionary, go back to last complete word which has not been used', . mean 'a word beginning with these letters is in the dictionary, proceed to next letter', o mean 'this is a complete word, proceed to next letter'. The various steps in dividing this message into word lengths would then appear as follows :

1	A	o
2	AT	o
3	ATA	x
4	AT A	o
5	AT AN	o
6	AT ANK	x
7	AT AN K	•
8	AT AN KE	•
9	AT AN KER	x
10	AT A N	•
11	AT A NK	x
12	A T	•
13	A TA	•
14	A TAN	o
15	A TANK	o
16	A TANKE	•
17	A TANKER	o
18	A TANKERF	x
19	A TANKER F	•
20	A TANKER FL	•
21	A TANKER FLY	o
22	A TANKER FLYI	•
23	A TANKER FLYIN	•
24	A TANKER FLYING	o
25	A TANKER FLYINGT	x
26	A TANKER FLYING T	•

The number of steps could be slightly reduced by using some miscellaneous instructions such as requiring that if a complete word at any point is AN, the next word must begin with a vowel. In the example above, steps 8 and 9 would be eliminated by this instruction. However it would probably not be worth while using any instructions except those which refer to words in the dictionary.

There will occasionally be a word in the message which is not in the dictionary. In that case the machine would have to be instructed to go back in turn to each one of the complete words which had not been used and begin again from there. If no sequence of words can be found by regrouping the letters up to this point, the machine would print

out the words up to the inadmissible word and the next letter separately and then proceed as though this were the beginning of a new mwssage. For example, suppose the word 'Finnish' in the message given were not in the dictionary. The machine steps would then appear as on the sheet attached. At step 44 all the complete words have been tried without success. Therefore A TANKER FLYING THE F is printed out and the remainder treated like a new message. The same situation arises at step 48. Here, however, the machine can go back only to step 45 in its search for complete words. The only possibility is IN at step 46. The machine tries IN N, IN NI and IN NIS At this point there is no word in the dictionary beginning with NIS,,and no complete word still untried. Therefore the machine prints out IN N and goes on to I, IS and ISH. Eventually it gets to F, FL, FLA, FLAG, etc. and from here on it is clear sailing.

The final version is A TANKER FLYING THE F IN NIS H FLAG WAS SIGHTED AT ZEROS EVEN HUNDRED.

Note that the machine would print out ZEROS EVEN instead of ZERO SEVEN, since it will always take the longest possible word. The machine might be instructed to scan the finished product for all complete words which have not been used. These may be broken up in different ways and scores for pairs of consecutive words calculated. If, for example, the final version contained the three consecutive words WILL BEAT ZERO ... the machine would already have noted that BE is a complete word in the dictionary, and on breaking BEAT at BE, the machine would also note that AT is another complete word in the dictionary. It would then compare WILL BEAT ZERO with WILL BE AT ZERO. WILL BE will score higher than WILL BEAT and AT ZERO will score higher than BEAT ZERO. Therefore the text will be changed to WILL BE AT ZERO ...

REF ID:A60419

1	A	o
2	AT	o
3	ATA	x
4	AT A	o
5	AT AN	o
6	AT ANK	x
7	AT AN K	.
8	AT AN KE	.
9	AT AN KER	x
10	AT A N'	.
11	AT A NK	x
12	A T	.
13	A TA	.
14	A TAN	o
15	A TANK	o
16	A TANKE	.
17	A TANKER	.
18	A TANKERF	x
19	A TANKER F	.
20	A TANKER FL	.
21	A TANKER FLY	o
22	A TANKER FLYI	.
23	A TANKER FLYIN	.
24	A TANKER FLYING	o
25	A TANKER FLYINGT	x
26	A TANKER FLYING T	.
27	A TANKER FLYING TH	.
28	A TANKER FLYING THE	o
29	A TANKER FLYING THEF	x
30	A TANKER FLYING THE F	.
31	A TANKER FLYING THE FI	.
32	A TANKER FLYING THE FIN	.
33	A TANKER FLYING THE FINN	x
34	A TANKER FLY I	.
35	A TANKER FLY IN	o
36	A TANKER FLY ING	x
37	A TANKER FLY IN G	.
38	A TANKER FLY IN GT	x
39	A TANK E	.
40	A TANK ER	.
41	A TANK ERF	x
42	A TAN K	.
43	A TAN KE	.
44	A TAN KER	x
45	A TANKER FLYING THE F I	.
46	A TANKER FLY ING THE F IN	o
47	A TANKER FLYI NG THE F INN	.
48	A TANKER FLY ING THE F INNI	x
49	A TANKER FLYING THE F IN M	.
50	A TANKER FLYING THE F IN MI	.
51	A TANKER FLYING THE F IN NIS	x

52	F	IN	N	I						.
53	F	IN	N	IS						o
54	F	IN	N	ISH						X
55	F	IN	N	IS	H					.
56	F	IN	N	IS	HF					X
57	F	IN	N	IS	H	F				.
58	F	IN	N	IS	H	FL				.
59	F	IN	N	IS	H	FLA				,
60	F	IN	N	IS	H	FLAG				o
61	F	IN	N	IS	H	FLAGW				X
62	F	IN	N	IS	H	FLAG	W			.
63	F	IN	N	IS	H	FLAG	WA			.
64	F	IN	N	IS	H	FLAG	WAS			o
65	F	IN	N	IS	H	FLAG	WASS			X
66	F	IN	N	IS	H	FLAG	WAS	S		.
67	F	IN	N	IS	H	FLAG	WAS	SI		.
68	F	IN	N	IS	H	FLAG	WAS	SIG		.
69	F	IN	N	IS	H	FLAG	WAS	SIGH		.
70	F	IN	N	IS	H	FLAG	WAS	SIGHT		o
71	F	IN	N	IS	H	FLAG	WAS	SIGHTE		.
72	F	IN	N	IS	H	FLAG	WAS	SIGHTED		o
73	F	IN	N	IS	H	FLAG	WAS	SIGHTEDA		X
74	F	IN	N	IS	H	FLAG	WAS	SIGHTED	A	.

MEMORANDUM

SUBJECT: Suggested Means for Increased Automation in Cryptanalysis.

1. GENERAL

In normal written alphabetic language there is a spatial relationship between and among contiguous letters, particularly as regards the vowels and consonants, which enables the mind to recognize almost instantaneously upon the receipt of visual impulses from the eyes a feature of such a language which we call pronounceability. It is my belief that the pronounceability phenomenon is what the mind apprehends or recognizes first of all and long before the succession or constellations of letters forming words are recognized as words and the latter become intelligible. What we need to know more about is this phenomenon of pronounceability and how the mind or brain apprehends it. The phenomenon is probably basically electrical in nature and the problem then is to design a machine that will simulate whatever electrical process takes place in the brain, a process which perhaps corresponds to the phenomenon in question. This, I believe, would not be too difficult and I conceive of a complex of components which would be of the nature described and would operate as a system in the manner described below:

2. PROCEDURE

a. Assume we are dealing with a cryptographic version of a text of 100 or more letters of good normal English which has been enciphered monoalphabetically by a random-mixed alphabet. Assume a machine of the digital computer type, into which impulses (in a binary code) corresponding to the succession of the letters or characters of the cipher text are fed, by the usual means (punched cards, perforated paper or magnetic tape, or the like) as the input.

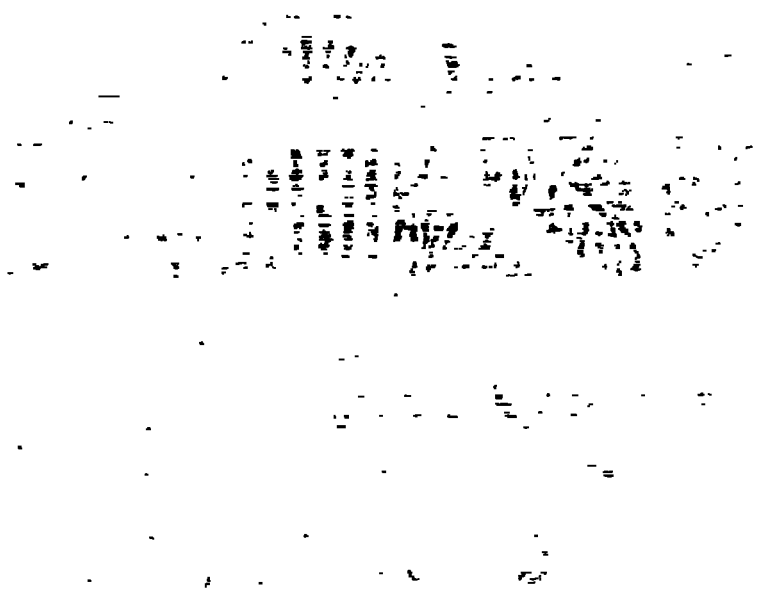
b. The first step for the machine, on acceptance of the input, would be to program its functioning, according to the usual methods, to conduct the impulses to and store them in the memory, making at the same time and storing a unilateral frequency count of the characters of the cipher text. The machine then determines the 10 cipher letters most frequently represented and sets them up in a descending order of frequency.

c. The machine's next operation is to set up 10! permutations of equivalents, beginning with the permutation corresponding to the normal English frequency expectancy series E, T, O, A, I,N, R, S, H, D. The substitution equivalents of the first of these 10! permutations are then applied by the machine to the appropriate letters of the cipher text and at the same time impulses which will trigger off signals corresponding to and capable of making sound spectographic representations of the substitutional plain-text equivalents are set up within the machine and temporarily impressed upon a medium capable of being moved laterally (film, magnetic tape, etc.) past a sound-spectograph reading component.

d. The sound-producing component of the machine, actuated by the reading component, "pronounces" the sequence of spectographic representations -- or, at least, it attempts to do so as best it can. The machine, furthermore, is set so that there is a lower threshold of "pronounceability," which unless reached and passed, will cause a "stuttering" or some phenomenon equivalent to a "lingual impediment." When the machine finds the impediment beyond the threshold or critical value of pronounceability -- in other words, when what it tries to enunciate is "unpronounceable" in English -- it throws out the permutation selected and begins to repeat its sequence of operations on the next of the 10! permutations.

e. The machine should be able in this way to eliminate a great many of the 10! permutations at a very rapid rate, retaining only a few which surpass the critical threshold of "pronounceability". These remaining possibilities will have to be examined visually or "listened to" by the operator.

f. Assuming the correct permutation of the 10! high frequency letters has been isolated the analyst will certainly be able to fill in the remaining letters without too much difficulty.



MEMORANDUM

SUBJECT: Suggested Means for Increased Automation in Cryptanalysis.

1. GENERAL

In the written form of any polysyllabic language which can be set down by means of the characters or elementary marks of a sequence of symbols called an alphabet, certain spatial relations exist between and among contiguous letters, particularly the vowels and consonants, which enables the mind to recognize almost instantaneously an important feature which is generally characteristic of such writing and language which we call pronounceability. It is my belief that the pronounceability phenomenon is what the mind apprehends or recognizes first of all and long before the succession or constellations of letters forming words are recognized as words and the latter become intelligible. What we need to know more about is this phenomenon of pronounceability and how the mind or brain apprehends it. The phenomenon is probably basically electrical in nature and the problem then is to design a machine that will simulate whatever electrical process takes place in the brain, a process which perhaps corresponds to the phenomenon in question. This, I believe, would not be too difficult and I conceive of a complex of components which would be of the nature described and would operate as a system in the manner described below:

2. PROCEDURE

a. Assume we are dealing with a cryptographic version of a text of 100 or more letters of good normal English which has been enciphered monoalphabetically by a random-mixed alphabet. Assume a machine of the digital computer type, into which impulses (in a binary code) corresponding to the succession of the letters or characters of the cipher text are fed, by the usual means (punched cards, perforated paper or magnetic tape, or the like) as the input.

b. The first step for the machine, on acceptance of the input, would be to program its functioning, according to the usual methods, to conduct the impulses to and store them in the memory, making at the same time and storing a unilateral frequency count of the characters of the cipher text. The machine then determines the 10 cipher letters most frequently represented and sets them up in a descending order of frequency.

c. The machine's next operation is to set up 10!-permutations of equivalents, beginning with the permutation corresponding to the normal English frequency expectancy series E, T, O, A, I,N, R, S, H, D. The substitution equivalents of the first of these 10! permutations are then applied by the machine to the appropriate letters of the cipher text and at the same time impulses which will trigger off signals corresponding to and capable of making sound spectographic representations of the substitutional plain-text equivalents are set up within the machine and temporarily impressed upon a medium capable of being moved laterally (film, magnetic tape, etc.) past a sound-spectograph reading component.

d. The sound-producing component of the machine, actuated by the reading component, "pronounces" the sequence of spectographic representations -- or, at least, it attempts to do so as best it can. The machine, furthermore, is set so that there is a lower threshold of "pronounceability," which unless reached and passed, will cause a "stuttering" or some phenomenon equivalent to a "lingual impediment." When the machine finds the impediment beyond the threshold or critical value of pronounceability -- in other words, when what it tries to enunciate is "unpronounceable" in English -- it throws out the permutation selected and begins to repeat its sequence of operations on the next of the 10! permutations.

e. The machine should be able in this way to eliminate a great many of the 10! permutations at a very rapid rate, retaining only a few which surpass the critical threshold of "pronounceability". These remaining possibilities will have to be examined visually or "listened to" by the operator.

f. Assuming the correct permutation of the 10! high frequency letters has been isolated the analyst will certainly be able to fill in the remaining letters without too much difficulty.

A monoalphabet substitution of a passage from an elementary treatise on cryptography:

SMDMA GDVVM PXRDC IOBPD

DCVQX RIGVF VXDLZ FLEFM

OFYLF IMVPY ESVLE DPVD I

OLEE+G FLLGD ASMYF MTXDM

R*OVFC SCLEP CYIMS MAIQG

D (101 letters total)

* should be Q

+ should be D

REF ID:A60419

S M D M A G D V V M P X R D C I O B P D D C V Q X R I G V F V X D L Z F
L E F M O F Y L F I M V P Y E S V L E D P V D I O L E F G F L L G D A S
M Y F M T X D M R O V F C S C L E P C Y I M S M A I Q G D

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
3	1	5	11	5	9	5	.	6	.	.	8	10	.	4	5	2	3	5	1	.	9	.	4	4	1

LE 4	EF 2	MA 2
GD 3	FL 2	VF 2
SM 3	FM 2	XR 2
DM 2	IO 2	XD 2
DC 2	IM 2	

LEF 2

YF 2 FY 1	DP 1 PD 1	MV 1 VM 1
CS 1 SC 1	FV 1 VF 2	MS 1 SM 3
DM 2 MD 1	FL 2 LF 1	PV 1 VP 1
DV 1 VD 1	FG 1 GF 1	

1. A frequency count of letters and digraphs was made as well as a chart showing the letters preceding and following each of the cipher letters. From these a table of reversed digraphs was made.

2. The low frequency letters were marked as consonants. The letters occurring only once or twice account for only 5 of the 101 letters of the message. The letters occurring 3 times were therefore added. This brought the frequency to 11, which was still below the 20% threshold. However if letters occurring 4 times are included, the total frequency would be 23. There are tests which might enable us to decide which of the 4-frequency letters to include and which to exclude, but this might be difficult to build into a machine.

3. A table was made of the letters preceding and following each of the assumed consonants A, R, Q, Z, B and T. All the letters except C, E and Y appeared. Since letters which never contact low-frequency consonants are very likely consonants themselves, these three were added to the assumed consonants. The contact table then appeared as follows :

A	R	Q	Z	B	T	C	E	Y
M	M	M	M	M				
						G	G	
			D	D	D	D	D	
					S	S	S	S
					I	I	I	I
					O	O		
					X	X	X	
					V	V		
L	L	L	L	L	L	L	L	
				F	F	F	F	F
				P	P	P	P	
					Y	Y		
					C			
						E		

M, D, S, I, L, F and P seem to be the best prospects for vowels. The substitutes for A, E, I and O should be found in this group.

4. The reversed digraphs were studied with a view to identifying vowels. Since most reversed digraphs will be of the form c-v or v-c, it should be possible to separate most of the letters which reverse into two classes, these being identified as consonant or vowel by reference to (3). However since word divisions are not shown these indications may be somewhat blurred because the first letter of one word may reverse with the last letter of the previous word. As a start only digraphs which reversed more than once were considered.

S M 3	Y F 2	D M 2	V F 2	F L 2
M S 1	F Y 1	M D 1	F V 1	L F 1

These can be separated into two classes in two ways, (a) M F - D S V Y L and (b) M V Y L - D S F. From (3), V and Y are not very good prospects for vowels; therefore D, S, V, Y and L in (a) probably represent the consonants. This would make M, I, F and P the only good prospects to represent A, E, I and O. However we have the following contacts among these four letters : M P 1, I M 2, F M 2, F I 1. In (b), M, V, Y and L probably represent the consonants. This would leave A, E, I and O to be found among D, S, F, I and P. There are only the following contacts among these five letters : D I 1, D P 1, F I 1, P D 1. Since one of these five letters would be eliminated, the number of these presumed v-v contacts would probably be reduced further. All in all (b) seems like the more probable set-up.

5. An attempt was made to identify the assumed vowels, D, S, F and the assumed consonants, M, V, Y, L . D, the highest frequency letter, which also appeared doubled seemed almost certain to be

E. F and S were assumed to be the two next most frequent vowels, A and O respectively. As for the consonants, M, V, Y, L; V and L which occur doubled seemed like good prospects for T and S respectively. M, the highest frequency consonant is probably N, H or R. From the first four letters of the message, S M D M , M = N seemed by far the best choice. The worksheet now appeared as in Fig. 1.

6. Changes were made in the plain text letter assumptions in order to make the fragments of text recovered look more like English.

O N E N E T T N
S M D M A G D V V M did not seem very likely. If we interchange the values of V and L, V = S, L = T yields O N E N E S S N
S M D M A G D V V M which seems to be an improvement, although there is probably something still wrong. A T T E O N
G F L L G D A S M suggests that G is H, R or L. G = R does not look very good in A R A T T R E
F G F L L G D and G = H does not look very good in O N E N H E S S N
S M D M A G D V V M . But G = L in A L A T T L E
F G F L L G D suggests F = I. The worksheet then looked as in Fig. 2.

7. The solution was completed using these recoveries.

. N E N . L E S S is obviously A N/E N D L E S S .

ONEN ETTN E EE T
SMDMAGDVVMPXRDC1oBPDDCVQXR

TAT ES AS AN A SA NT O
1GVFVXDLZFLEFMoFYLF1MVPYES

TS E TE S A ASS E ON AN
VLEDPVD1oLEFGFLLGDASMYFMTX

EN TA O S ON E
DMRoVFCSCLEPCY1MSMA1QGD

FIG. 1

NEN LESSN E EE S
SMDMAGDVVMPXRDC1oBPDDCVQXR

LS1S ET LT 1N 1 T1 NS
1GVFVXDLZFLEFMoFYLF1MVPYES

ST E SE T 1L1TTLE N IN
VLEDPVD1oLEFGFLLGDASMYFMTX

EN S1 T N N LE
DMRoVFCSCLEPCY1MSMA1QGD

FIG. 2