MEMO ROUTING SLIP

NEVER USE FOR APPROVALS, DISAPPROVALS, CONCURRENCES, OR SIMILAR ACTIONS

1 NAME OR TITLE Mr Clark	INITIALS		CIRCULATE
ORGANIZATION AND LOCATION [illegible]	DATE		COORDINATION
2 H. Campaigne	H 8 IV		FILE
I would like [illegible]			INFORMATION
3 [illegible] about this [illegible]			NECESSARY ACTION
[illegible]		✓	NOTE AND RETURN
4 Clark			SEE ME
			SIGNATURE

REMARKS

It turned out there was no room for this on the program, which was just as well. I hope eventually to modify this into something worthwhile.

FROM NAME OR TITLE HC

DATE 8 April.

ORGANIZATION AND LOCATION

TELEPHONE

10 May 1955

~~CONFIDENTIAL~~

MEMORANDUM

SUBJECT: Use of Machines in Solving Ciphers - Monoalphabetic
Substitution

There has been recent discussion on the place that automatic computers or data processing machines might play in the solution of ciphers. This memorandum discusses one of the simplest cases, namely, a mono-alphabet substitution cipher. It is hoped that the study of a simple case will be of some help in understanding the role of machines and their limitations.

Description of Cipher Text

The cipher text consists of 100 letters. These are divided into 20 five-letter groups with no apparent indication of any proper word division. The text appears in the form of perforations in a tape and it is assumed that the machine can read this. It is also assumed that the machine will type out the solution on a paper tape, in capital letters, with the proper word separation but no further punctuation.

In order to restrict the operations we start out by making certain assumptions. We do not inquire into the justification for them.

1. The cryptogram is a single alphabet substitution cipher with word separations and punctuation omitted.

2. The plain text is in straight forward English without code words and there has been no effort made to make the cipher into a puzzle by introducing unusual words and phraseologies. There may be unusual words in the normal course of events, however. There may also be words which have not found their way into the dictionary.

It is necessary to provide for the possibility that no solution is obtained. In practice this would usually be because the assumptions are incorrect. The plain text might for example be in Spanish.

Solution

The solution to be outlined below is not well thought out, and it probably would be modified in important details at least. However, at the present stage what is needed is an illustrative solution, and an exact optimum procedure is not of great importance.

It is suggested that the following material, arranged in order of importance, be stored in the machine:

1. The relative frequency of the letters in English.

2. The relative frequency of all digraphs. There are 676 of these. Some of these such as jj are very rare but all may occur. The rarest ones may be said to have a probability of approximately zero.

3. The relative frequency of about 10000 trigraphs. These are some 17000 of these but it would seem unnecessary to list those whose probability is substantially zero. Perhaps 10000 are unnecessarily many.

4. A list of the 10000 most frequent tetragraphs.

5. A list of 10000 most frequent words. These should be grouped by length, one letter words, two letter, three letter, etc. Within each such group the listing may be alphabetical and it would perhaps be advantageous to have alphabetical listings starting with the second, third, etc., letters as well as by the first. Perhaps 10000 is an unnecessarily large number of words (and trigraphs and tetragraphs).

The following are the suggested steps in the solution.

1. Make a frequency count of the letters.

2. Make a skeleton of a tentative solution by putting e's for the most frequent letter, t for the next most frequent, etc., until about one-half of the letters have been replaced. This has perhaps made use of ten different letters which we may assume to be

etaoinshrd

3. Examine the digraphs. The number of these will very from message to message but is usually not far from an average of 25. There are only 100 possibilities and only a few of them are very improbable. Among these are aa,ii,ee,ao and hh. In addition to looking for the individual possibilities we divide the digraphs into classes as follows:

A - vowel, vowel
B - vowel, consonant
C - consonant, vowel
D - consonant, consonant

B and C should predominate. If A and D predominate or if B and C don't predominate sufficiently it is an indication that consonants and vowels should be interchanged. Thus the procedure is to interchange t with e, a, o, etc. and note whether the conditions improve. Similarly n, s, etc. may be tentatively exchanged with i, a, etc. When classes B and C predominate sufficiently it is probable that the assignment as between vowels and consonants are for the most part correct, but within these groups there is ample room for error. a and i are difficult to distinguish.

4. The trigraphs may be assumed to average ten or twelve, with half the letters in place but this number is subject to variations, more so than the number of digraphs. To the extent that they occur they are now examinant for probability. If very improbable ones occur the found letters are interchanged tentatively until the trigraphs become more probable. In these operations it is to be expected that some progress will be made in improving assignments within the vowel group and within the consonant group as well as between the groups. During these operations tentative interchanges between the letters already placed and the more frequent of those not yet placed (perhaps l, u) should also be attempted.

5. The next criterion to attempt is that of the tetragraphs. Those are very few, perhaps five on the average, and this number is subject to very large variations. Moreover, most of them have a very low probability but have to be accepted in any case. The procedure would be to repeat as under the trigraphs.

6. The next step is to attempt to assign a value to the next most frequent letter, the eleventh under our assumptions. We first try the most common of those that are left. This gives rise to a number of digraphs, trigraphs, etc. and these can be expected to determine whether it was a good choice. If it was not the next most frequent one is tried and so forth. If it is difficult to find a satisfactory fit, attempts should be made to interchange with one of those already assigned.

The other letters are added one by one in the same manner.

7. The next logical step is to attempt to break up the tentative text into words. It appears advantageous to begin with the least frequent letters. That is begin with the letter z and try to determine the words in the text of which it is a part. Then proceed to x, etc. One point in favor of this is that there are few words to select from and the search need not be long. Another point is that these infrequent letters are rather likely to have been incorrectly assigned so there is an opportunity for making corrections early in the operation. Once this process gets under way the corrections that are made are of assistance in what follows and it should be possible to identify subsequent words that occur in the list if they are also in the stored list of words.

8. After this process has been carried as far as possible there appears to be nothing more that can be done except to type out the material with the word spacings found. If no solution was found this will be disclosed by an examination by a human.

Evaluation

The method outlined cannot be relied on to give a complete solution. It may be in the first place that the method sometimes fails to arrive at even an approximation. We are not much concerned with this case because it is believed that this only happens because the message is too short or because some of the processes outlined have not been well thought out. We are concerned, however, with the eventuality that the results are correct in the main but the results as printed still are incomplete.

It is thought that such cases would occur and further that in most of these cases it would be possible for a human being to look at the printed answer and deduce the correct answer very quickly. This is because the human being has relatively large span of perception so that he can judge by how words fail to fit together what corrections must be made. He can also detect incorrect word division by the same means. Further, he can locate garbles (caused either by an enciphering error or by faulty transmission). Here also the means is essentially a long perception span.

We may also say that we have not been able to program the machine to make use of the semantics of the text but only the structure on a rather small scale. It does not appear that the matter of semantics vs. structure is the same as the matter of large perception span vs. short span but the effect is similar.

We have only been able to program into the machine a means for making use of the fine structure and what is lacking is the invention of means for making an overall examination of the message and take corrective steps in accordance with the findings.

It may be concluded that in the absence of such means it is an essential part of the solution that a human being should examine the tentative solution as it comes from the machine.

If this must be done the question arises whether it might not be better to have the examination made at an earlier stage in the proceedings. The machine is at its best in the early stages when what is required is counting and other simple operations. It is not doing so well in the stages when the last few letters are being introduced one at a time. It may be that if the partial work were submitted for inspection at a time when 90% of the letters have been inserted that a human being could complete the solution very quickly. This would save quite a little comparison work in connection with these last letters and also the work of word division.

HARRY NYQUIST