How the Germans Broke a U. S. Code

BY KATHARINE L. SWIFT

~~*Top Secret Daunt*~~

Here is the other side of the coin—the story of the solution by the Germans of one of our diplomatic codes just prior to the entry of the United States into World War II.

Among the documents coming from Germany through TICOM activities at the end of World War II was a series of German technical papers on cryptologic subjects, similar in purpose to those being published in the *Technical Journal.* Two of these papers dealt with the manner in which the Germans had solved our letter code "B7." The first described the method of removing the encipherment from the basic code groups; the second described the "bookbreaking" process, i.e. the way in which the meaning of each basic code group was determined. It is the second paper, published in 1941 and bearing the title, "The Linguistic Interpretation of the American Letter Code B7", by Dr. Hans-Kurt Mueller, (TICOM/D-3c), which is reprinted here. Dr. Mueller describes the initial stage of the bookbreaking process—when the beachhead is established. The bookbreaker proceeds by gradually extending this beachhead.

Progress can be relatively fast in the case of a *one-part*, or alphabetical code, in which the plaintext groups are arranged in alphabetical order accompanied by their code groups in alphabetical or numerical order. Such a code serves for decoding as well as encoding. Below is a brief extract from a typical one-part code:

ABABD	A
ABACF	Abaft
ABAHK	Abandon
ABAJL	.....it
ABALN	Abandoned
ABAMP	.....by
ABAWZ	Abandoning
ABBAD	Abandonment
.....	
.....	
ZYZYZ	Zero

In this case, each code group identified gives clues to the alphabetical limitations of unidentified groups.

In a *two-part*, or randomized code, however, the plaintext groups are arranged in alphabetical order accompanied by their code groups in a nonsystematic order. Such a list can serve only for encoding. For decoding, another list must be provided in which the code groups are arranged in alphabetical or numerical order and are accompanied by their meanings as given in the encoding section. Below is an extract from a typical two-part code:

Encoding Section		*Decoding Section*	
GAJVY	A	ABABD	Obstructed
TOGTY	Abaft	ABACF	Term
FEHIL	Abandon	ABAHK	Zero
BAYLT	.....it	ABAJL	If it has not
ZYZYZ	Abandoned	ABALN	To be sent by
NYSYZ	.....by	ABAMP	Acceding
IFWUZ	Abandoning	ABAWZ	Building
RUMGO	Abandonment	ABBAD	Do not attempt
.....		.....	
.....		.....	
ABAHK	Zero	ZYZYZ	Abandoned[1]

It is obvious that in such a code the identification (or "recovery", as it is called) of any one meaning will give no clue to the alphabetical position of the meaning of any other code group. Without this valuable aid, the identification of meanings is often much less certain and progress is greatly slowed down. Our B7 was such a two-part code.

The materials used by the Germans were basically the same as those used by other bookbreakers then and now, although data processing equipment has replaced the highly tedious copying processes of the Germans in 1940. The modern bookbreaker uses the following materials:

1. *The original messages* (one copy only)
2. *A Message Print* (called Material I and II in this paper)— a copy of the original messages reduced to their basic code groups, with the encipherment removed.
3. *An Index*—an organized listing of all occurrences of every code group, showing the message in which it occurred, its position in that message, and several code groups preceding and following each occurrence. Our modern machine-made index covers the

[1] The above description of one-part and two-part codes was borrowed from *Military Cryptanalytics*, Part II, by Lambros D. Callimahos and William F. Friedman, pp 449–450.

function of the German index (apparently an index of occurrences only without context) plus some of the special lists mentioned.

4. *An Inverse Frequency List*—a list of the code groups in order of frequency of use, starting with the most frequently used group.
5. *A Lane Log*—called the "circuit catalog" in this paper, in which its usefulness is well illustrated. The circuit and date of each message are data included also in most of our modern indexes because of their significance.

As code meanings are identified, a *decode* and an *encode* are built up as further working tools.

The methods used by the German bookbreakers are like those of bookbreakers everywhere. Every possible clue is used: frequency of a code group, its position in the message, its relationship to other groups, repetitions, patterns in beginnings and endings, plaintext preambles, known codes used on the same circuit or, as here, in the same message. There are always weaknesses, and it is the purpose of the bookbreaker to find and exploit those weaknesses to establish his beachhead. The most effective method, of course, unsporting though it may be, is discovery of a "crib"—the same text sent in plain text or in a known system. Apart from this, perhaps the most useful single aid is acquaintance with the past cryptologic practice of the senders. This is well illustrated in the following case history.

The Linguistic Interpretation of the American Letter Code B7

BY DR. HANS-KURT MUELLER

In connection with the article by Dr. Lohan on the solution of the encipherment of the American Code B7, the purely linguistic method will be described by which the meanings were ascertained. This assignment was attacked by three bookbreakers who had a number of assistants available. The day's work was divided into two shifts. On the first shift the bookbreakers worked on the actual meanings of the groups, on the second the groups interpreted were entered in the index and copies of the text. Other work was also done, e. g., taking out groups in special lists (i. e., all passages in which the same group occurred were copied out including 3–4 preceding and 3–4 following groups), the laying out of a "Sach-Code" (a sort of encode arranged by subjects), etc.

The material which the bookbreakers received from the workshop of the cryptanalysts came in three forms: "Original," "Material I" and "Material II." The "Original" contained the original telegrams in enciphered text with the intermediate text (i. e., the deciphered original code groups) written

above. "Material I" and "Material II" contained a copy of the same telegrams in the deciphered code groups only (without the enciphered text).

The bookbreakers worked primarily with "Material I" and "Material II," resorting to the original only in special cases (e. g., in case of garbles) . . . It proved very practical that the bookbreakers used two copies as this avoided friction of a technical nature. (Of course both copies had to be kept up-to-date at all times). The index was also available in several copies (original and two photographic copies). This also proved worthwhile for technical reasons.

Before the bookbreakers began the work of interpretation, other important preliminary work had been done. Not only was the index complete, numerous frequent groups had also been taken out in special lists, beginnings and endings had been copied off, repeats in the texts had been underscored and in the majority of cases entered in separate lists, a circuit catalog had been prepared, and many other things. Accordingly we had at our disposal material which had already been worked over thoroughly from a technical angle. When we opened it up, it began to speak of its own accord, as it were, even though we did not yet understand its language. We soon noticed that it was a rational language, one that must be capable of a systematic interpretation. The numerous repetitions and parallel passages interwoven in manifold ways, the striking peaks, the distribution of frequent and rare groups, all signified that we were dealing with something orderly, organic.

The first task of the bookbreakers was to become familiar with the material, to study the texts over again and again, to memorize striking repetitions, etc. But very soon, about the second day, the first break occurred. It must be stated first that during the preliminary work the cryptanalysts had noted two of the most frequent groups: BIBOT and KUJAN; the suspicion arose early that both meant "PERIOD" or "PARAGRAPH" since their distribution through the text suggested the occurrences of the period in the language. Moreover, their interpretation was facilitated by the fact that in numerous telegrams which had continuations and were marked by the plaintext prefixes "Section 1," "Section 2," "Section 3," etc., the individual sections often ended or began with BIBOT or KUJAN. Two other groups had also attracted attention during the preliminary work, viz., DODEV and GYBUX, which always appeared as an associated pair and were separated by a variable number of groups. It was natural to assume them to mean respectively "OPEN QUOTES" and "CLOSE QUOTES" or "OPEN PARENTHESIS" and "CLOSE PARENTHESIS," or to be dashes between which something was enclosed.

These interpretations, which the cryptanalysts had made, did not, to be sure, give any of the content of the texts but did give valuable clues as to the formal construction, which meant a big gain.

The actual break-in usually results from message beginnings. Generally telegrams begin with a reference to an earlier message, e. g., "My telegram number 267 of 6 February, 10 o'clock a. m." or the like. In our case the matter of the beginnings was not especially favorable. Frequent groups did occur at the beginning which probably must mean "my telegram," "your telegram," "department's telegram," etc., but the following groups, which must contain the number and date, did not show the striking repetitions one would expect. There were several reasons for this (as came out in the further course of the work). For one thing the Americans have the habit of sending messages on the same circuit now in one system, now in another. Thus in the present system we had only single numbers of a series while intervening numbers were sent in another system. Hence we often lacked adequate material for comparisons so that at first it was impossible to determine various relationships. Moreover, the present code (as came out later) has a group for each of the three digit numbers: "236" need not be written in three groups (200 + 30 + 6) or in two (200 + 36) but is represented by a single group; likewise in this code all the 366 dates are included (e. g., April 6 is one group). Obviously this greatly reduced the chances of repetitions. On the other hand there was one circumstance which facilitated the work: the Americans have the habit of giving clock time after the date (9 p.m., 11 a.m., etc.). Since practically only the hours 9–10 a.m. to 7–8 p.m. come into account for filing messages, we really had to deal with only ten to twelve time groups for hours so that here repetitions were bound to pile up. However, the clock time is the least important part of the dating, hence it is easy to see that beginning groups were not as favorable in breaking into B7 as in the case of other codes.

The first break came rather from the conclusions. It struck us that often toward the end of a telegram came a "PERIOD" (BIBOT or KUJAN) followed by only a few groups, sometimes only two. It was also striking that the group after the period was often GUDOR. This occurred almost always in messages which were sent simultaneously to two or more stations. Reference to the fact that a message was being sent to another station was often made at the beginning with the words: "Following telegram sent to the department" not in the code of the message but in another, the so-called B3 (Gray Code) which had been solved years ago. For instance, there was a telegram from Berlin to Moscow introduced by the words in B3 "Following telegram sent to department." From that (supported by acquaintance with American cryptographic practice) we knew that the same telegram with the same wording was sent to Washington as well as to Moscow and that at the end we might expect the words "Repeated to Moscow." This telegram to Moscow was merely a copy of one to Washington; at its close would naturally stand the statement that it had been repeated to Moscow. At the end of this telegram were the three groups: BIBOT ("PERIOD") GUDOR KOMYB. Ac-

cordingly GUDOR must be "REPEATED TO" and KOMYB "MOSCOW" (or GUDOR "REPEATED" and KOMYB "TO MOSCOW," but this alternative dropped out due to deduction from analogies). The following was to be considered: if GUDOR meant "REPEATED TO" it could never stand at the very end of a message; actually it never did.

The first meaning—"REPEATED TO"—and the first city name had thus been obtained.

The next step was to study all messages with GUDOR near the end and determine to which stations they had been sent. In this way a whole series of city names was established (Paris, London, Bucarest, Warsaw, etc.). The telegrams repeated to Ankara yielded exceptionally favorable results. In them there were always three groups after GUDOR at the end. This was striking and immediately suggested that the name Ankara was spelled out. It was known that Code B7 was very old, somewhere about 1920. But Ankara was not made capital of the new Turkish Republic by Kemal Pasha until 1923 and so began to play a political role only in 1923. So it seemed likely that the present code did not have a special group for Ankara and the name had to be spelled out. This was obviously done with three groups and at first it was hard to tell how the word was divided. The following circumstance led to a clear, incontrovertible decision: there was a whole series of messages which were sent to both Washington and Ankara. All had three groups at the end after GUDOR but these were not always the same. The actual picture was:

FYJOV	DAKUC	COCIH
FYJOV	DAKUC	DOLYJ
HUMUW	DAKUC	LAHOF
HUMUW	DAKUC	DOLYJ

In other words the middle group was always the same; the first and last parts could be expressed by different groups, i. e., had several code values. Moreover the groups which occurred as parts 1 and 3 of this combination were very frequent ones which kept recurring both singly and in other combinations, while the middle group was one of the rarer groups of the code and seldom appeared except in this combination. Hence the middle group must represent a rather rare spelling group while the first and third parts must be frequently used short words. What else could they be but the indefinite articles "AN" and "A" (also the letter "A")? Beyond doubt the name broke into the components AN–KAR–A.

This meant a great advance. We not only had two valuable articles (with alternate values)—something that normally comes only much later in the work—but we knew that all substantives and adjectives following "A" began with a consonant and all after "AN" began with a vowel or silent h. Moreover we had the first letter of the alphabet.

Likewise fruitful were the telegrams repeated to Istanbul. Here too GUDOR was always followed by three groups. This time, however, the first and last parts were always represented by the same group while the middle part was variable. It took no effort to establish the division IST–AN–BUL. The group for "BUL" hardly occurred outside this combination, that for "IST" mostly in this combination but also in others (recall that "IST" is a frequent suffix in English used to form nouns of agent, e. g., economist, socialist, Iron-guardist, propagandist). The variable middle portion ("AN") was represented by several frequent groups, and it was no slight satisfaction to establish the fact that two of those we had fixed for Ankara were included among them.

The next task was to find the prepositions, employing the city names for the purpose. Before city names one often finds IN, FROM, TO, etc. Strangely enough, before the groups for Paris, London, Berlin, etc., only a few groups were found repeated, whereas before groups for Ankara, Istanbul and Vichy[1] there were numerous repetitions. That suggested that such expressions as "IN LONDON", "FROM PARIS", "TO BERLIN", etc., were given in the code, i. e., that B7 must contain numerous phrases. The encoder had to place a separate group for the preposition, however, before spelled place names.

Various groups which appeared in the body of the text before ANKARA, ISTANBUL, etc., were suspected of being prepositions, only it was not certain yet which groups represented which prepositions. Some of these prepositions also stood at the beginnings of messages and must therefore be "FOR" or "FROM" since at the beginning there is often a statement of "for whom" the message is intended or "from whom" it comes. In this connection, for instance, the group FIFEL was early suspected of signifying "FOR", an assumption confirmed subsequently.

Very early the suspicion arose that we were dealing with a two-part code. This was soon confirmed.

I have described the first break into the meanings at some length to show by these few illustrations the general principle of the procedure which was applied in the further course of the work. Essentially it is a matter of the simple and elementary principle of deducing the unknowns from the known values.

[1] Mention of Vichy is evidence that this code recovery problem was solved in 1940 or 1941, date of publication of the paper.—K. L. S.