~~Restricted~~

WAR DEPARTMENT
OFFICE OF THE CHIEF SIGNAL OFFICER
WASHINGTON

MILITARY CRYPTANALYSIS

Part I

MONOALPHABETIC SUBSTITUTION SYSTEMS

By

WILLIAM F. FRIEDMAN

Principal Cryptanalyst
Chief of Signal Intelligence Section
War Plans and Training Division

PREPARED UNDER THE DIRECTION OF THE
CHIEF SIGNAL OFFICER

UNITED STATES
GOVERNMENT PRINTING OFFICE
WASHINGTON : 1938

MILITARY CRYPTANALYSIS, PART I. MONOALPHABETIC SUBSTITUTION SYSTEMS

SECTION I

INTRODUCTORY REMARKS

1. Scope of this text.—*a*. It is assumed that the student has studied the two preceding texts written by the same author and forming part of this series, viz, *Elementary Military Cryptography*, and *Advanced Military Cryptography*. These texts deal exclusively with *cryptography* as defined therein; that is, with the various types of ciphers and codes, their principles of construction, and their employment in cryptographing and decryptographing messages. Particular emphasis was placed upon such means and methods as are practicable for military usage. It is also assumed that the student has firmly in mind the technically precise, special nomenclature employed in those texts, for the terms and definitions therein will all be used in the present text, with essentially the same significances. If this is not the case, it is recommended that the student review his preceding work, in order to regain a familiarity with the specific meanings assigned to the terms used therein. There will be no opportunity herein to repeat this information and unless he understands clearly the significance of the terms employed, his progress will be retarded.

b. This text constitutes the first of a series of texts on *cryptanalysis*. Although most of the information contained herein is applicable to cryptograms of various types and sources, special emphasis will be laid upon the principles and methods of solving military cryptograms. Except for an introductory discussion of fundamental principles underlying the science of cryptanalytics, this first text in the series will deal solely with the principles and methods for the analysis of monoalphabetic substitution ciphers. Even with this limitation it will be possible to discuss only a few of the many variations of this one type; but with a firm grasp upon the general principles no difficulties should be experienced with any variations that may be encountered.

c. This and some of the succeeding texts will deal only with elementary types of cipher systems not because they may be encountered in military operations but because their study is essential to an understanding of the principles underlying the solution of the modern, very much more complex types of ciphers and codes that are likely to be employed by the larger governments today in the conduct of their military affairs in time of war.

d. All of this series of texts will deal only with the solution of visible secret writing. At some future date, texts dealing with the solution of invisible secret writing, and with secret signalling systems, may be prepared.

2. Mental equipment necessary for cryptanalytic work.—*a*. Captain Parker Hitt, in the first United States Army manual[1] dealing with cryptography, opens the first chapter of his valuable treatise with the following sentence:

> Success in dealing with unknown ciphers is measured by these four things in the order named: perseverance, careful methods of analysis, intuition, luck.

[1] Hitt, Capt. Parker, *Manual for the Solution of Military Ciphers*. Army Service Schools Press, Fort Leavenworth, Kansas, 1916. 2d Edition, 1918. (Both out of print.)

These words are as true today as they were then. There is no royal road to success in the solution of cryptograms. Hitt goes on to say:

Cipher work will have little permanent attraction for one who expects results at once, without labor, for there is a vast amount of purely routine labor in the preparation of frequency tables, the rearrangement of ciphers for examination, and the trial and fitting of letter to letter before the message begins to appear.

The present author deems it advisable to add that the kind of work involved in solving cryptograms is not at all similar to that involved in solving "cross-word puzzles", for example. The wide vogue the latter have had and continue to have is due to the appeal they make to the quite common instinct for mysteries of one sort or another; but in solving a cross-word puzzle there is usually no necessity for performing any preliminary labor, and palpable results become evident after the first minute or two of attention. This successful start spurs the cross-word "addict" on to complete the solution, which rarely requires more than an hour's time. Furthermore, cross-word puzzles are all alike in basic principle and once understood, there is no more to learn. Skill comes largely from the embellishment of one's vocabulary, though, to be sure, constant practice and exercise of the imagination contribute to the ease and rapidity with which solutions are generally reached. In solving cryptograms, however, many principles must be learned, for there are many different systems of varying degrees of complexity. Even some of the simpler varieties require the preparation of tabulations of one sort or another, which many people find irksome; moreover, it is only toward the very close of the solution that results in the form of intelligible text become evident. Often, indeed, the student will not even known whether he is on the right track until he has performed a large amount of preliminary "spade work" involving many hours of labor. Thus, without at least a willingness to pursue a fair amount of theoretical study, and a *more than average amount of patience and perseverance*, little skill and experience can be gained in the rather difficult art of cryptanalysis. General Givierge, the author of an excellent treatise on cryptanalysis, remarks in this connection: [2]

The cryptanalyst's attitude must be that of William the Silent: No need to hope in order to undertake, nor to succeed in order to persevere.

b. As regards Hitt's reference to careful methods of analysis, before one can be said to be a cryptanalyst worthy of the name it is necessary that one should have firstly a sound knowledge of the basic principles of cryptanalysis, and secondly, a long, varied, and active *practical* experience in the successful application of those principles. It is not sufficient to have read treatises on this subject. One month's actual practice in solution is worth a whole year's mere reading of theoretical principles. An exceedingly important element of success in solving the more intricate ciphers is the possession of the rather unusual mental faculty designated in general terms as the power of inductive and deductive reasoning. Probably this is an inherited rather than an acquired faculty; the best sort of training for its emergence, if latent in the individual, and for its development is the study of the natural sciences, such as chemistry, physics, biology, geology, and the like. Other sciences such as linguistics and philology are also excellent. Aptitude in mathematics is quite important, more especially in the solution of ciphers than of codes.

c. An active imagination, or perhaps what Hitt and other writers call *intuition*, is essential, but mere imagination uncontrolled by a judicious spirit will more often be a hindrance than a help. In practical cryptanalysis the imaginative or intuitive faculties must, in other words, be guided by good judgment, by practical experience, and by as thorough a knowledge of the general situation or extraneous circumstances that led to the sending of the cryptogram as is possible to obtain. In this respect the many cryptograms exchanged between correspondents whose identities and general affairs, commercial, social, or political, are known are far more readily

[2] Givierge, Général Marcel, *Cours de Cryptographie*, Paris, 1925, p. 301.

solved than are isolated cryptograms exchanged between unknown correspondents, dealing with unknown subjects. It is obvious that in the former case there are good data upon which the intuitive powers of the cryptamalyst can be brought to bear, whereas in the latter case no such data are available. Consequently, in the absence of such data, no matter how good the imagination and intuition of the cryptanalyst, these powers are of no particular service to him. Some writers, however, regard the intuitive spirit as valuable from still another viewpoint, as may be noted in the following: [3]

Intuition, like a flash of lightning, lasts only for a second. It generally comes when one is tormented by a difficult decipherment and when one reviews in his mind the fruitless experiments already tried. Suddenly the light breaks through and one finds after a few minutes what previous days of labor were unable to reveal.

This, too, is true, but unfortunately there is no way in which the intuition may be summoned at will, when it is most needed.[4] There are certain authors who regard as indispensable the possession of a somewhat rare, rather mysterious faculty that they designate by the word "flair", or by the expression "cipher brains." Even so excellent an authority as General Givierge,[5] in referring to this mental facility, uses the following words: "Over and above perseverance and this aptitude of mind which some authors consider a special gift, and which they call intuition, or even, in its highest manifestation, clairvoyance, cryptographic studies will continue more and more to demand the qualities of orderliness and memory." Although the present author believes a special aptitude for the work is essential to cryptanalytic success, he is sure there is nothing mysterious about the matter at all. Special aptitude is prerequisite to success in all fields of endeavor. There are, for example, thousands of physicists, hundreds of excellent ones, but only a handful of world-wide fame. Should it be said, then, that a physicist

[3] Lange et Soudart, *Traité de Cryptographie*, Librairie Félix Alcan, Paris, 1925, p. 104.

[4] The following extracts are of interest in this connection:

The fact that the scientific investigator works 50 per cent of his time by non-rational means is, it seems, quite insufficiently recognized. There is without the least doubt an instinct for research, and often the most successful investigators of nature are quite unable to give an account of their reasons for doing such and such an experiment, or for placing side by side two apparently unrelated facts. Again, one of the most salient traits in the character of the successful scientific worker is the capacity for knowing that a point is proved when it would not appear to be proved to an outside intelligence functioning in a purely rational manner; thus the investigator feels that some proposition is true, and proceeds at once to the next set of experiments without waiting and wasting time in the elaboration of the formal proof of the point which heavier minds would need. Questionless such a scientific intuition may and does sometimes lead investigators astray, but it is quite certain that if they did not widely make use of it, they would not get a quarter as far as they do. Experiments confirm each other, and a false step is usually soon discovered. And not only by this partial replacement of reason by intuition does the work of science go on, but also to the born scientific worker—and emphatically they cannot be made—the structure of the method of research is as it were given, he cannot explain it to you, though he may be brought to agree *a postiori* to a formal logical presentation of the way the method works.—Excerpt from Needham, Joseph, *The Sceptical Biologist*, London, 1929, p. 79.

The essence of scientific method, quite simply, is to try to see how data arrange themselves into causal configurations. Scientific problems are solved by collecting data and by "thinking about them all the time." We need to look at strange things until, by the appearance of known configurations, they seem familiar, and to look at familiar things until we see novel configurations which make them appear strange. We must look at events until they become luminous. That is scientific method . . . Insight is the touchstone . . . The application of insight as the touchstone of method enables us to evaluate properly the role of imagination in scientific method. The scientific process is akin to the artistic process: it is a process of selecting out those elements of experience which fit together and recombining them in the mind. Much of this kind of research is simply a ceaseless mulling over, and even the physical scientist has considerable need of an armchair . . . Our view of scientific method as a struggle to obtain insight forces the admission that science is half art . . . Insight is the unknown quantity which has eluded students of scientific method.—Excerpts from an article entitled *Insight and Scientific Method*, by Willard Waller, in *The American Journal of Sociology*, Vol. XL, 1934.

[5] *Op. cit.*, p. 302.

who has achieved very notable success in his field has done so because he is the fortunate posessor of a *mysterious* faculty? That he is fortunate in possessing a special aptitude for his subject is granted, but that there is anything mysterious about it, partaking of the nature of clairvoyance (if, indeed, the latter is a *reality*) is not granted. While the ultimate nature of any mental process seems to be as complete a mystery today as it has ever been, the present author would like to see the superficial veil of mystery removed from a subject that has been shrouded in mystery from even before the Middle Ages down to our own times. (The principal and easily understandable reason for this is that governments have always closely guarded cryptographic secrets and anything so guarded soon becomes "mysterious.") He would, rather, have the student approach the subject as he might approach any other science that can stand on its own merits with other sciences, because cryptanalytics, like other sciences, has a practical importance in human affairs. It presents to the inquiring mind an interest in its own right as a branch of knowledge; it, too, holds forth many difficulties and disappointments, and these are all the more keenly felt when the nature of these difficulties is not understood by those unfamiliar with the special circumstances that very often are the real factors that led to success in other cases. Finally, just as in the other sciences wherein many men labor long and earnestly for the true satisfaction and pleasure that comes from work well-done, so the mental pleasure that the successful cryptanalyst derives from his accomplishments is very often the only reward for much of the drudgery that he must do in his daily work. General Givierge's words in this connection are well worth quoting:[6]

> Some studies will last for years before bearing fruit. In the case of others, cryptanalysts undertaking them never get any result. But, for a cryptanalyst who likes the work, the joy of discoveries effaces the memory of his hours of doubt and impatience.

d. With his usual deft touch, Hitt says of the element of luck, as regards the role it plays in analysis:

> As to luck, there is the old miners' proverb: "Gold is where you find it."

The cryptanalyst is lucky when one of the correspondents whose ciphers he is studying makes a blunder that gives the necessary clue; or when he finds two cryptograms identical in text but in different keys in the same system; or when he finds two cryptograms identical in text but in different systems, and so on. The element of luck is there, to be sure, *but the cryptanalyst must be on the alert* if he is to profit by these lucky "breaks."

e. If the present author were asked to state, in view of the progress in the field since 1916, what elements might be added to the four ingredients Hitt thought essential to cryptanalytic success, he would be inclined to mention the following:

(1) A broad, general education, embodying interests covering as many fields of practical knowledge as possible. This is useful because the cryptanalyst is often called upon to solve messages dealing with the most varied of human activities, and the more he knows about these activities, the easier his task.

(2) Access to a large library of current literature, and wide and direct contacts with sources of collateral information. These often afford clues as to the contents of specific messages. For example, to be able instantly to have at his disposal a newspaper report or a personal report of events described or referred to in a message under investigation goes a long way toward simplifying or facilitating solution. Government cryptanalysts are sometimes fortunately situated in this respect, especially where various agencies work in harmony.

(3) Proper coordination of effort. This includes the organization of cryptanalytic personnel into harmonious, efficient teams of cooperating individuals.

[6] *Op. cit.*, p. 301.

(4) Under mental equipment he would also include the faculty of being able to concentrate on a problem for rather long periods of time, without distraction, nervous irritability, and impatience. The strain under which cryptanalytic studies are necessarily conducted is quite severe and too long-continued application has the effect of draining nervous energy to an unwholesome degree, so that a word or two of caution may not here be out of place. One should continue at work only so long as a peaceful, calm spirit prevails, whether the work is fruitful or not. But just as soon as the mind becomes wearied with the exertion, or just as soon as a feeling of hopelessness or mental fatigue intervenes, it is better to stop completely and turn to other activities, rest, or play. It is essential to remark that systematization and orderliness of work are aids in reducing nervous tension and irritability. On this account it is better to take the time to prepare the data carefully, rewrite the text if necessary, and so on, rather than work with slipshod, incomplete, or improperly arranged material.

(5) A retentive memory is an important asset to cryptanalytic skill, especially in the solution of codes. The ability to remember individual groups, their approximate locations in other messages, the associations they form with other groups, their peculiarities and similarities, saves much wear and tear of the mental machinery, as well as much time in looking up these groups in indexes.

f. It may be advisable to add a word or two at this point to prepare the student to expect slight mental jars and tensions which will almost inevitably come to him in the conscientious study of this and the subsequent texts. The present author is well aware of the complaint of students that authors of texts on cryptanalysis base much of their explanation upon their foreknowledge of the "answer"—which the student does not know while he is attempting to follow the solution with an unbiased mind. They complain too that these authors use such expressions as "obviously", "naturally", "of course", "it is evident that", and so on, when the circumstances seem not at all to warrant their use. There is no question but that this sort of treatment is apt to discourage the student, especially when the point elucidated becomes clear to *him* only after many hours' labor, whereas, according to the book, the author noted the weak spot at the first moment's inspection. The present author can only promise to try to avoid making the steps appear to be much more simple than they really are, and to suppress glaring instances of unjustifiable "jumping at conclusions." At the same time he must indicate that for pedagogical reasons in many cases a message has been consciously "manipulated" so as to allow certain principles to become more obvious in the illustrative examples than they ever are in practical work. During the course of some of the explanations attention will even be directed to cases of unjustified inferences. Furthermore, of the student who is quick in observation and deduction, the author will only ask that he bear in mind that if the elucidation of certain principles seems prolix and occupies more space than necessary, this is occasioned by the author's desire to carry the explanation forward in very short, easily-comprehended, and plainly-described steps, for the benefit of students who are perhaps a bit slower to grasp but who, once they understand, are able to retain and apply principles slowly learned just as well, if not better than the students who learn more quickly.

3. Validity of results of cryptanalysis.—Valid, or authentic cryptanalytic solutions cannot and do not represent "opinions" of the cryptanalyst. They are valid only so far as they are wholly objective, and are susceptible of demonstration and proof, employing authentic, objective methods. It should hardly be necessary (but an attitude frequently encountered among laymen makes it advisable) to indicate that the validity of the results achieved by any serious cryptanalytic studies on authentic material rests upon the same sure foundations and are reached by the same general steps as the results achieved by any other scientific studies; viz, observation, hypothesis, deduction and induction, and confirmatory experiment. Implied in the latter is the possibility that two or more qualified investigators, each working independently upon the same material, will achieve identical (or practically identical) results. Occasionally a pseudo-cryptanalyst offers "solutions" which cannot withstand such tests; a second, unbiased, investigator working independently either cannot *consistently* apply the methods alleged to have been applied by the pseudo-cryptanalyst, or else, if he can apply them at all, the results (plain-text translations) are far different in the two cases. The reason for this is that in such cases it is generally found that the "methods" are not clear-cut, straightforward or mathematical in character. Instead, they often involve the making of judgments on matters too tenuous to measure, weigh, or otherwise subject to careful scrutiny. In such cases, the conclusion to which the unprejudiced observer is forced to come is that the alleged "solution" obtained by the first investigator, the pseudo-cryptanalyst, is purely subjective. In nearly all cases where this has happened (and they occur from time to time) there has been uncovered nothing which can in any way be used to impugn the integrity of the pseudo-cryptanalyst. The worst that can be said of him is that he has become a victim of a special or peculiar form of self-delusion, and that his desire to solve the problem, usually in accord with some previously-formed opinion, or notion, has over-balanced, or undermined, his judgment and good sense.[7]

[7] Specific reference can be made to the following typical "case histories":

Donnelly, Ignatius, *The Great Cryptogram*. Chicago, 1888.
Owen, Orville W., *Sir Francis Bacon's Cipher Story*. Detroit, 1895.
Gallup, Elizabeth Wells, *Francis Bacon's Biliteral Cipher*. Detroit, 1900.
Margoliouth, D. S., *The Homer of Aristotle*. Oxford, 1923.
Newbold, William Romaine, *The Cipher of Roger Bacon*. Philadelphia, 1928. (For a scholarly and complete demolition of Professor Newbold's work, see an article entitled *Roger Bacon and the Voynich MS*, by John M. Manly, in *Speculum*, Vol. *VI*, No. 3, July 1931.)
Arensberg, Walter Conrad, *The Cryptography of Shakespeare*. Los Angeles, 1922.
The Shakespearean Mystery. Pittsburgh, 1928.
The Baconian Keys. Pittsburgh, 1928.
Feely, Joseph Martin, *The Shakespearean Cypher*. Rochester, N. Y., 1931.
Deciphering Shakespeare. Rochester, N. Y., 1934.

SECTION II

FUNDAMENTAL PRINCIPLES

4. The four basic operations in cryptanalysis.—*a*. The solution of practically every cryptogram involves four fundamental operations or steps:

(1) The determination of the language employed in the plain-text version.

(2) The determination of the general system of cryptography employed.

(3) The reconstruction of the specific key in the case of a cipher system, or the reconstruction, partial or complete, of the code book, in the case of a code system; or both, in the case of an enciphered code system.

(4) The reconstruction or establishment of the plain text.

b. These operations will be taken up in the order in which they are given above and in which they usually are performed in the solution of cryptograms, although occasionally the second step may precede the first.

5. The determination of the language employed.—*a*. There is not much that need be said with respect to this operation except that the determination of the language employed seldom comes into question in the case of studies made of the cryptograms of an organized enemy. By this is meant that during wartime the enemy is of course known, and it follows, therefore, that the language he employs in his messages will almost certainly be his native or mother tongue. Only occasionally nowadays is this rule broken. Formerly it often happened, or it might have indeed been the general rule, that the language used in diplomatic correspondence was not the mother tongue, but French. In isolated instances during the World War, the Germans used English when their own language could for one reason or another not be employed. For example, for a year or two before the entry of the United States into that war, during the time America was neutral and the German Government maintained its embassy in Washington, some of the messages exchanged between the Foreign Office in Berlin and the Embassy in Washington were cryptographed in English, and a copy of the code used was deposited with the Department of State and our censor. Another instance is found in the case of certain Hindu conspirators who were associated with and partially financed by the German Government in 1915 and 1916; they employed English as the language of their cryptographic messages. Occasionally the cryptograms of enemy agents may be in a language different from that of the enemy. But in general these are, as has been said, isolated instances; as a rule, the language used in cryptograms exchanged between members of large organizations is the mother tongue of the correspondents. Where this is not the case, that is, when cryptograms of unknown origin must be studied, the cryptanalyst looks for any indications on the cryptograms themselves which may lead to a conclusion as to the language employed. Address, signature, and plain-language words in the preamble or in the body of the text all come under careful scrutiny, as well as all extraneous circumstances connected with the manner in which the cryptograms were obtained, the person on whom they were found, or the locale of their origin and destination.

b. In special cases, or under special circumstances a clue to the language employed is found in the nature and composition of the cryptographic text itself. For example, if the letters K and W are entirely absent or appear very rarely in messages, it may indicate that the language is Spanish, for these letters are absent in the alphabet of that language and are used only to spell foreign words or names. The presence of accented letters or letters marked with special signs of one sort or another, peculiar to certain languages, will sometimes indicate the language used. The Japanese Morse telegraph alphabet and the Russian Morse telegraph alphabet contain combinations of dots and dashes which are peculiar to those alphabets and thus the interception of messages containing these special Morse combinations at once indicates the language involved. Finally, there are certain peculiarities of alphabetic languages which, in certain types of cryptograms, *viz*, pure transposition, give clues as to the language used. For example, the frequent digraph C H, in German, leads to the presence, in cryptograms of the type mentioned, of many isolated C's and H's; if this is noted, the cryptogram may be assumed to be in German.

c. In some cases it is perfectly possible to perform certain steps in cryptanalysis *before* the language of the cryptogram has been definitely determined. Frequency studies, for example, may be made and analytic processes performed without this knowledge, and by a cryptanalyst wholly unfamiliar with the language even if it has been identified, or who knows only enough about the language to enable him to recognize valid combinations of letters, syllables, or a few common words in that language. He may, after this, call to his assistance a translator who may not be a cryptanalyst but who can materially aid in making necessary assumptions based upon his special knowledge of the characteristics of the language in question. Thus, cooperation between cryptanalyst and translator results in solution.[1]

6. The determination of the general system.—*a.* Except in the case of the more simple types of cryptograms, the determination of the general system according to which a given cryptogram has been produced is usually a difficult, if not the most difficult, step in its solution. The reason for this is not hard to find.

b. As will become apparent to the student as he proceeds with his study, *in the final analysis, the solution of every cryptogram involving a form of substitution depends upon its reduction to monoalphabetic terms, if it is not originally in those terms.* This is true not only of ordinary substitution ciphers, but also of combined substitution-transposition ciphers, and of enciphered code. If the cryptogram must be reduced to monoalphabetic terms, the manner of its accomplishment is usually indicated by the cryptogram itself, by external or internal phenomena which become apparent to the cryptanalyst as he studies the cryptogram. If this is impossible, or too difficult the cryptanalyst must, by one means or another, discover how to accomplish this reduction, by bringing to bear all the special or collateral information he can get from all the sources at his command. If both these possibilities fail him, there is little left but the long, tedious, and often fruitless process of elimination. In the case of transposition ciphers of the more complex type, the discovery of the basic method is often simply a matter of long and tedious elimination of possibilities. For cryptanalysis has unfortunately not yet attained, and may indeed never attain, the precision found today in qualitative analysis in chemistry, for example, where the analytic process is absolutely clear cut and exact in its dichotomy. A few words in explanation of what is meant may not be amiss. When a chemist seeks to determine the identity of an unknown

[1] The writer has seen in print statements that "during the World War . . . decoded messages in Japanese and Russian without knowing a word of either language." The extent to which such statements are exaggerated will soon become obvious to the student. Of course, there are occasional instances in which a mere clerk with quite limited experience may be able to "solve" a message in an extremely simple system in a language of which he has no knowledge at all; but such a "solution" calls for nothing more arduous than the ability to recognize pronounceable combinations of vowels and consonants—an ability that hardly deserves to be rated as "cryptanalytic" in any real sense. To say that it is possible to solve a cryptogram in a foreign language "without knowing a word of that language" is not quite the same as to say that it is possible to do so with only a slight knowledge of the language; and it may be stated without cavil that the better the cryptanalyst's knowledge of the language, the greater are the chances for his success and, in any case, the easier is his work.

substance, he applies certain specific reagents to the substance and in a specific sequence. The first reagent tells him definitely into which of two primary classes the unknown substance falls. He then applies a second test with another specific reagent, which tells him again quite definitely into which of two secondary classes the unknown substance falls, and so on, until finally he has reduced the unknown substance to its simplest terms and has found out what it is. In striking contrast to this situation, cryptanalysis affords exceedingly few "reagents" or tests that may be applied to determine positively that a given cipher belongs to one or the other of two systems yielding externally similar results. And this is what makes the analysis of an isolated, complex cryptogram so difficult. Note the limiting adjective "isolated" in the foregoing sentence, for it is used advisedly. It is not often that the general system fails to disclose itself or cannot be discovered by painstaking investigation when there is a great volume of text accumulating from a regular traffic between numerous correspondents in a large organization. *Sooner or later* the system becomes known, either because of blunders and carelessness on the part of the personnel entrusted with the cryptographing of the messages, or because the accumulation of text itself makes possible the determination of the general system by cryptanalytic studies. But in the case of a single or even a few isolated cryptograms concerning which little or no information can be gained by the cryptanalyst, he is often unable, without a knowledge of, or a shrewd guess as to the general system employed, to decompose the heterogeneous text of the cryptogram into homogeneous, monoalphabetic text, which is the ultimate and essential step in analysis. The only knowledge that the cryptanalyst can bring to his aid in this most difficult step is that gained by long experience and practice in the analysis of many different types of systems.

c. On account of the complexities surrounding this particular phase of cryptanalysis, and because in any scheme of analysis based upon successive eliminations of alternatives the cryptanalyst can only progress so far as the extent of his own knowledge of *all* the possible alternatives will permit, it is necessary that detailed discussion of the eliminative process be postponed until the student has covered most of the field. For example, the student will perhaps want to know at once how he can distinguish between a cryptogram that is in code or enciphered code from one that is in cipher. It is at this stage of his studies impracticable to give him any helpful indications on his question. In return it may be asked of him why he should expect to be able to do this in the early stages of his studies when often the experienced expert cryptanalyst is baffled on the same score!

d. Nevertheless, in lieu of more precise tests not yet discovered, a general guide that may be useful in cryptanalysis will be built up, step by step as the student progresses, in the form of a series of charts comprising what may be designated *An Analytical Key For Cryptanalysis.* (See Par. 50.) It may be of assistance to the student if, as he proceeds, he will carefully study the charts and note the place which the particular cipher he is solving occupies in the general cryptanalytic panorama. These charts admittedly constitute only very brief outlines, and can therefore be of but little direct assistance to him in the analysis of the more complex types of ciphers he may encounter later on. So far as they go, however, they may be found to be quite useful in the study of elementary cryptanalysis. For the experienced cryptanalyst they can serve only as a means of assuring that no possible step or process is inadvertently overlooked in attempts to solve a difficult cipher.

e. Much of the labor involved in cryptanalytic work, as referred to in Par. 2, is connected with this determination of the general system. The preparation of the text, its rewriting in different forms, sometimes being rewritten in a half dozen ways, the recording of letters, the establishment of frequencies of occurrences of letters, comparisons and experiments made with known material of similar character, and so on, constitute much labor that is most often indispensable, but which sometimes turns out to have been wholly unnecessary, or in vain. In a recent treatise [2] it is stated quite boldly that "this work once done, the determination of the system is often relatively easy." This statement can certainly apply only to the simpler types of ciphers; it is entirely misleading as regards the much more frequently encountered complex cryptograms of modern times.

7. The reconstruction of the specific key.—*a.* Nearly all practical cryptographic methods require the use of a specific key to guide, control, or modify the various steps under the general system. Once the latter has been disclosed, discovered, or has otherwise come into the possession of the cryptanalyst, the next step in solution is to determine, if necessary, and if possible, the specific key that was employed to cryptograph the message or messages under examination. This determination may not be in complete detail; it may go only so far as to lead to a knowledge of the number of alphabets involved in a substitution cipher, or the number of columns involved in a transposition cipher, or that a one-part code has been used, in the case of a code system. But it is often desirable to determine the specific key in as complete a form and with as much detail as possible, for this information will very frequently be useful in the solution of subsequent cryptograms exchanged between the same correspondents, since the nature of the specific key in a solved case may be expected to give clues to the specific key in an unsolved case.

b. Frequently, however, the reconstruction of the key is not a prerequisite to, and does not constitute an absolutely necessary preliminary step in, the fourth basic operation, *viz*, the reconstruction or establishment of the plain text. In many cases, indeed, the two processes are carried along simultaneously, the one assisting the other, until in the final stages both have been completed in their entireties. In still other cases the reconstruction of the specific key may succeed instead of precede the reconstruction of the plain text, and is accomplished purely as a matter of academic interest; or the specific key may, in unusual cases, never be reconstructed.

8. The reconstruction of the plain text.—*a.* Little need be said at this point on this phase of cryptanalysis. The process usually consists, in the case of substitution ciphers, in the establishment of equivalency between specific letters of the cipher text and the plain text, letter by letter, pair by pair, and so on, depending upon the particular type of substitution system involved. In the case of transposition ciphers, the process consists in rearranging the elements of the cipher text, letter by letter, pair by pair, or occasionally word by word, depending upon the particular type of transposition system involved, until the letters or words have been returned to their original plain-text order. In the case of code, the process consists in determining the meaning of each code group and inserting this meaning in the code text to reestablish the original plain text.

b. The foregoing processes do not, as a rule, begin at the beginning of a message and continue letter by letter, or group by group in sequence up to the very end of the message. The establishment of values of cipher letters in substitution methods, or of the positions to which cipher letters should be transferred to form the plain text in the case of transposition methods, comes at very irregular intervals in the process. At first only one or two values scattered here and there throughout the text may appear; these then form the "skeletons" of words, upon which further work, by a continuation of the reconstruction process, is made possible; in the end the complete or nearly complete [3] text is established.

c. In the case of cryptograms in a foreign language, the translation of the solved messages is a final and necessary step, but is not to be considered as a cryptanalytic process. However, it is commonly the case that the translation process will be carried on simultaneously with the cryptanalytic, and will aid the latter, especially when there are lacunae which may be filled in from the context. (See also Par. 5*c* in this connection.)

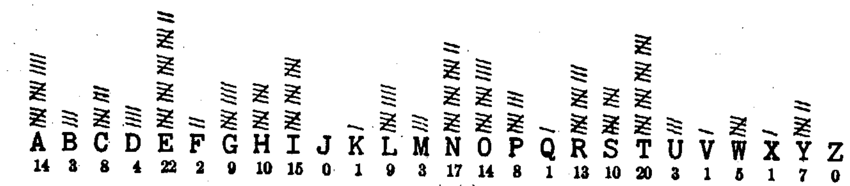
[2] Lange et Soudart, *op. cit.*, p. 106.

[3] Sometimes in the case of code, the meaning of a few code groups may be lacking, because there is insufficient text to establish their meaning.

SECTION III

FREQUENCY DISTRIBUTIONS

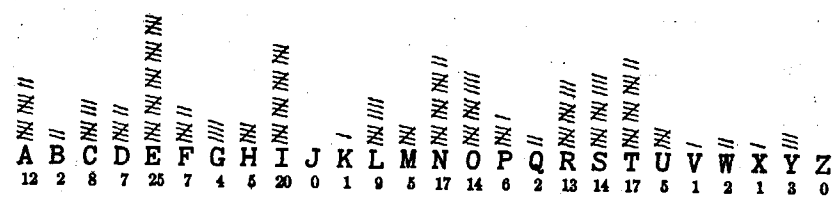
9. The simple or uniliteral frequency distribution.—*a.* It has long been known to cryptographers and typographers that the letters composing the words of any intelligible written text composed in any language which is alphabetic in construction are employed with greatly varying frequencies. For example, if on cross-section paper a simple tabulation, shown in Fig. 1, called a *uniliteral frequency distribution*, is made of the letters composing the words of the preceding sentence, the variation in frequency is strikingly demonstrated. It is seen that whereas certain letters, such as A, E, I, N, O, R, S, and T, are employed very frequently, other letters, such as C, G, P, and W are employed not nearly so frequently, while still other letters, such as F, J, Q, V, and Z are employed either seldom or not at all.



(Total=200 letters)

FIGURE 1.

b. If a similar tabulation is now made of the letters comprising the words of the second sentence in the preceding paragraph, the graph shown in Fig. 2 is obtained. Both sentences have exactly the same number of letters (200).



(Total=200 letters)

FIGURE 2.

c. Although each of these two graphs exhibits great variation in the relative frequencies with which *different* letters are employed in the respective sentences to which they apply, no marked differences are exhibited between the frequencies of the *same* letter in the two graphs. Compare, for example, the frequencies of A, B, C . . . Z in Fig. 1 with those of A, B, C, . . . Z in Fig. 2. Aside from one or two exceptions, as in the case of the letter F, these two graphs agree rather strikingly.

d. This agreement, or *similarity,* would be practically complete if the two texts were much longer, for example, five times as long. In fact, when two texts of similar character, each containing more than 1,000 letters, are compared, it would be found that the respective frequencies of the 26 letters composing the two graphs show only very slight differences. This means, in other words, that in normal text each letter of the alphabet occurs with a rather *constant* or *characteristic frequency* which it tends to approximate, depending upon the length of the text analyzed. The longer the text (within certain limits), the closer will be the approximation.[1]

e. An experiment along these lines will be convincing. A series of 260 official telegrams[2] passing through the War Department Message Center was examined statistically. The messages were divided into five sets, each totaling 10,000 letters, and the five distributions shown in Table 1–A, were obtained.

f. If the five distributions in Table 1–A are summed, the results are as shown in Table 2–A.

TABLE 1–A.—*Absolute frequencies of letters appearing in five sets of Governmental plain-text telegrams, each set containing 10,000 letters, arranged alphabetically*

Message No. 1		Message No. 2		Message No. 3		Message No. 4		Message No. 5	
Letter	Absolute Frequency	Letter	Absolute Frequency	Letter	Absolute Frequency	Letter	Absolute Frequency	Letter	Absolute Frequency
A	738	A	783	A	681	A	740	A	741
B	104	B	103	B	98	B	83	B	99
C	319	C	300	C	288	C	326	C	301
D	387	D	413	D	423	D	451	D	448
E	1, 367	E	1, 294	E	1, 292	E	1, 270	E	1, 275
F	253	F	287	F	308	F	287	F	281
G	166	G	175	G	161	G	167	G	150
H	310	H	351	H	335	H	349	H	349
I	742	I	750	I	787	I	700	I	697
J	18	J	17	J	10	J	21	J	16
K	36	K	38	K	22	K	21	K	31
L	365	L	393	L	333	L	386	L	344
M	242	M	240	M	238	M	249	M	268
N	786	N	794	N	815	N	800	N	780
O	685	O	770	O	791	O	756	O	762
P	241	P	272	P	317	P	245	P	260
Q	40	Q	22	Q	45	Q	38	Q	30
R	760	R	745	R	762	R	735	R	786
S	658	S	583	S	585	S	628	S	604
T	936	T	879	T	894	T	958	T	928
U	270	U	233	U	312	U	247	U	238
V	163	V	173	V	142	V	133	V	155
W	166	W	163	W	136	W	133	W	182
X	43	X	50	X	44	X	53	X	41
Y	191	Y	155	Y	179	Y	213	Y	229
Z	14	Z	17	Z	2	Z	11	Z	5
Total	10, 000		10, 000		10, 000		10, 000		10, 000

[1] See footnote 5, page 16.

[2] These comprised messages from several departments in addition to the War Department and were all of an administrative character.

TABLE 2–A.—*Absolute frequencies of letters appearing in the combined five sets of messages totaling 50,000 letters, arranged alphabetically*

A	3, 683	G	819	L	1 ,821	Q	175	V	766
B	487	H	1, 694	M	1 ,237	R	3, 788	W	780
C	1, 534	I	3, 676	N	3 ,975	S	3, 058	X	231
D	2, 122	J	82	O	3 ,764	T	4, 595	Y	967
E	6, 498	K	148	P	1 ,335	U	1, 300	Z	49
F	1, 416								

g. The frequencies noted in subparagraph *f,* when reduced to the basis of 1,000 letters and then used as a basis for constructing a simple chart that will exhibit the variations in frequency in a striking manner, yield the following graph which is hereafter designated as the *normal,* or *standard uniliteral frequency distribution* for English telegraphic plain text:

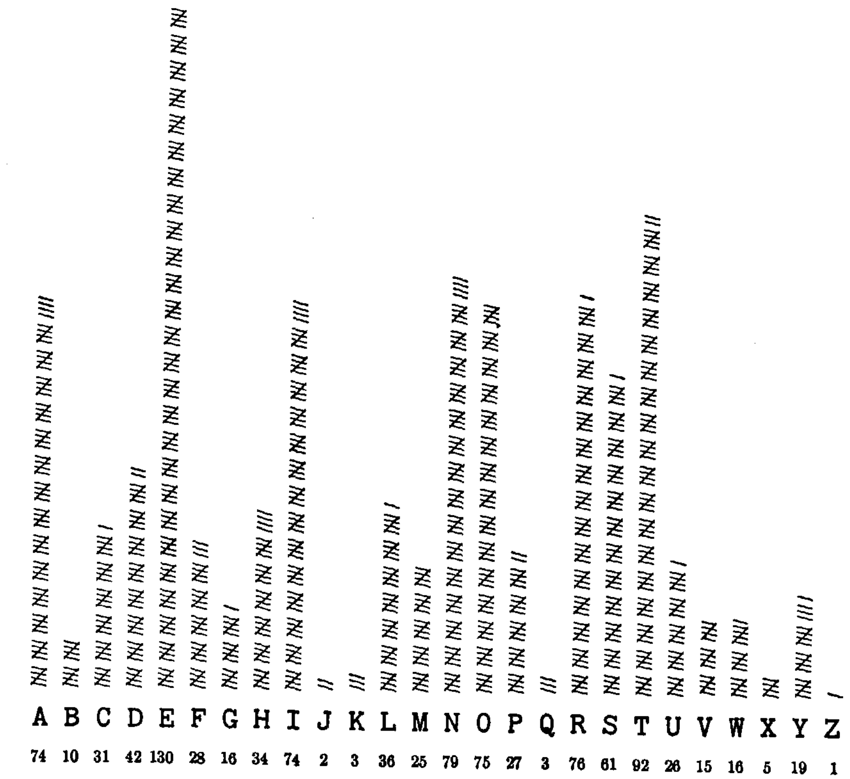


FIGURE 3.

10. Important features of the normal uniliteral frequency distribution.—*a.* When the graph shown in Fig. 3 is studied in detail, the following features are apparent:

(1) It is quite irregular in appearance. This is because the letters are used with greatly varying frequencies, as discussed in the preceding paragraph. This irregular appearance is often described by saying that the graph shows marked *crests and troughs,* that is, points of high frequency and low frequency.

(2) The relative positions in which the crests and troughs fall within the graph, that is, the *spatial relations* of the crests and troughs, are rather definitely fixed and are determined by circumstances which have been explained in a preceding text.[3]

(3) The relative heights and depths of the crests and troughs within the graph, that is, the *linear extensions* of the lines marking the respective frequencies, are also rather definitely fixed, as would be found if an equal volume of similar text were analyzed.

(4) The most prominent crests are marked by the vowels A, E, I, O, and the consonants N, R, S, T; the most prominent troughs are marked by the consonants J, K, Q, X, and Z.

(5) The important data are summarized in tabular form in Table 3.

TABLE 3

	Frequency	Percent of total	Percent of total in round numbers
6 Vowels: A E I O U Y	398	39.8	40
20 Consonants:			
5 High Frequency (D N R S T)	350	35.0	35
10 Medium Frequency (B C F G H L M P V W)	238	23.8	24
5 Low Frequency (J K Q X Z)	14	1.4	1
Total	1,000	100.0	100

(6) The frequencies of the letters of the alphabet are as follows:

A	74	G	16	L	36	Q	3	V	15
B	10	H	34	M	25	R	76	W	16
C	31	I	74	N	79	S	61	X	5
D	42	J	2	O	75	T	92	Y	19
E	130	K	3	P	27	U	26	Z	1
F	28								

(7) The relative order of frequency of the letters is as follows:

E	130	I	74	C	31	Y	19	X	5
T	92	S	61	F	28	G	16	Q	3
N	79	D	42	P	27	W	16	K	3
R	76	L	36	U	26	V	15	J	2
O	75	H	34	M	25	B	10	Z	1
A	74								

(8) The four vowels A, E, I, O (combined frequency 353) and the four consonants N, R, S, T (combined frequency 308) form 661 out of every 1,000 letters of plain text; in other words, *less than $\frac{1}{3}$ of the alphabet is employed in writing $\frac{2}{3}$ of normal plain text.*

[3] Section VII, *Elementary Military Cryptography.*

b. The data given in Fig. 3 and Table 3 represent the relative frequencies found in a large volume of English telegraphic text of a governmental, administrative character. These frequencies will vary somewhat with the nature of the text analyzed. For example, if an equal number of telegrams dealing solely with *commercial* transactions in the *leather industry* were studied statistically, the frequencies would be slightly different because of the repeated occurrence of words peculiar to that industry. Again, if an equal number of telegrams dealing solely with *military* messages of a *tactical* character were studied statistically, the frequencies would differ slightly from those found above for general governmental messages of an administrative character.

c. If ordinary English literary text (such as may be found in any book, newspaper, or printed document) were analyzed, the frequencies of certain letters would be changed to an appreciable degree. This is because in telegraphic text words which are not strictly essential for intelligibility (such as the definite and indefinite articles, certain prepositions, conjunctions and pronouns) are omitted. In addition, certain essential words, such as "stop", "period", "comma", and the like, which are usually indicated in written or printed matter by symbols not easy to transmit telegraphically and which must, therefore, be spelled out in telegrams, occur very frequently. Furthermore, telegraphic text often employs longer and more uncommon words than does ordinary newspaper or book text.

d. As a matter of fact, other tables compiled in the Office of the Chief Signal Officer gave slightly different results, depending upon the source of the text. For example, three tables based upon 75,000, 100,000, and 136,257 letters taken from various sources (telegrams, newspapers, magazine articles, books of fiction) gave as the relative order of frequency for the first 10 letters the following:

For 75,000 letters	E T R N I O A S D L
For 100,000 letters	E T R I N O A S D L
For 136,257 letters	E T R N A O I S L D

TABLE 4.—*Frequency table for 10,000 letters of literary English, as compiled by Hitt*

ALPHABETICALLY ARRANGED

A	778	G	174	L	372	Q	8	V	112
B	141	H	595	M	288	R	651	W	176
C	296	I	667	N	686	S	622	X	27
D	402	J	51	O	807	T	855	Y	196
E	1,277	K	74	P	223	U	308	Z	17
F	197								

ARRANGED ACCORDING TO FREQUENCY

E	1,277	R	651	U	308	Y	196	K	74
T	855	S	622	C	296	W	176	J	51
O	807	H	595	M	288	G	174	X	27
A	778	D	402	P	223	B	141	Z	17
N	686	L	372	F	197	V	112	Q	8
I	667								

Hitt also compiled data for telegraphic text (but does not state what kind of messages) and gives the following table:

TABLE 5.—*Frequency table for 10,000 letters of telegraphic English, as compiled by Hitt*

ALPHABETICALLY ARRANGED

A	813	G	201	L	392	Q	38	V	136
B	149	H	386	M	273	R	677	W	166
C	306	I	711	N	718	S	656	X	51
D	417	J	42	O	844	T	634	Y	208
E	1,319	K	88	P	243	U	321	Z	6
F	205								

ARRANGED ACCORDING TO FREQUENCY

E	1,319	S	656	U	321	F	205	K	88
O	844	T	634	C	306	G	201	X	51
A	813	D	417	M	273	W	166	J	42
N	718	L	392	P	243	B	149	Q	38
I	711	H	386	Y	208	V	136	Z	6
R	677								

e. Frequency data applicable purely to English military text were compiled by Hitt,[4] from a study of 10,000 letters taken from orders and reports. The frequencies found by him are given in Tables 4 and 5.

11. **Constancy of the standard or normal, uniliteral frequency distribution.**—*a.* The relative frequencies disclosed by the statistical study of large volumes of text may be considered to be the standard or *normal* frequencies of the letters of written English. Counts made of smaller volumes of text will tend to approximate these normal frequencies, and, within certain limits,[5] the smaller the volume, the lower will be the degree of approximation to the normal, until, in the case of a very short message, the normal proportions may not obtain at all. It is advisable that the student fix this fact firmly in mind, for the sooner he realizes the true nature of any data relative to the frequency of occurrence of letters in text, the less often will his labors toward the solution of specific ciphers be thwarted and retarded by too strict an adherence to these generalized principles of frequency. He should constantly bear in mind that such data are merely statistical generalizations, that they will be found to hold strictly true only in large volumes of text, and that they may not even be approximated in short messages.

b. Nevertheless the normal frequency distribution or the "normal expectancy" for any alphabetic language is, in the last analysis, the best guide to, and the usual basis for, the solution of cryptograms of a certain type. It is useful, therefore, to reduce the normal, uniliteral frequency distribution to a basis that more or less closely approximates the volume of text which the cryptanalyst most often encounters in individual cryptograms. As regards length of messages, counting only the letters in the body, and excluding address and signature, a study of the

[4] *Op. cit.*, pp. 6–7.

[5] It is useless to go beyond a certain limit in establishing the normal-frequency distribution for a given language. As a striking instance of this fact, witness the frequency study made by an indefatigable German, Kaeding, who in 1898 made a count of the letters in about 11,000,000 words, totaling about 62,000,000 letters in German text. When reduced to a percentage basis, and when the relative order of frequency was determined, the results he obtained differed very little from the results obtained by Kasiski, a German cryptographer, from a count of only 1,060 letters. See Kaeding, *Haeufigkeitswoerterbuch*, Steglitz, 1898; Kasiski, *Die Geheimschriften und die Dechiffrir-Kunst*, Berlin, 1863.

260 telegrams referred to in paragraph 9 shows that the arithmetical average is 217 letters; the statistical mean, or weighted average,[6] however, is 191 letters. These two results are, however, close enough together to warrant the statement that the *average* length of telegrams is approximately 200 letters. The frequencies given in Par. 9*f* have therefore been reduced to a basis of 200 letters, and the following uniliteral frequency distribution may be taken as showing the most typical distribution to be expected in 200 letters of telegraphic English text:

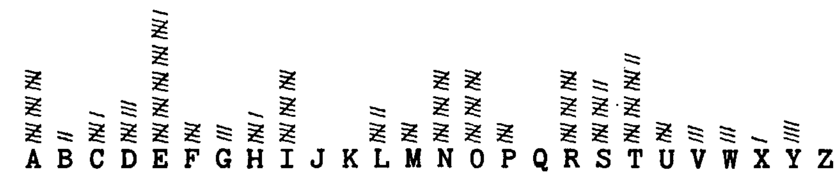


FIGURE 4.

c. The student should take careful note of the appearance of the distribution[7] shown in Fig. 4, for it will be of much assistance to him in the early stages of his study. The manner of setting down the tallies should be followed by him in making his own distributions, indicating every fifth occurrence of a letter by an oblique tally. This procedure almost automatically shows the total number of occurrences for each letter, and yet does not destroy the graphical appearance of the distribution, especially if care is taken to use approximately the same amount of space for each set of five tallies. Cross-section paper is very useful for this purpose.

d. The word "uniliteral" in the designation "uniliteral frequency distribution" means "single letter", and it is to be inferred that other types of frequency distributions may be encountered. For example, a distribution of pairs of letters, constituting a biliteral frequency distribution, is very often used in the study of certain cryptograms in which it is desired that pairs made by combining successive letters be listed. A biliteral distribution of A B C D E F would take these pairs: AB, BC, CD, DE, EF. The distribution could be made in the form of a large square divided up into 676 cells. When distributions beyond biliteral are required (triliteral, quadraliteral, etc.) they can only be made by listing them in some order, for example, alphabetically based on the 1st, 2d, 3d, . . . letter.

[6] The arithmetical average is obtained by adding each different length and dividing by the number of different-length messages; the mean is obtained by multiplying each different length by the number of messages of that length, adding all products, and dividing by the total number of messages.

[7] The use of the terms "distribution" and "frequency distribution", instead of "table" and "frequency table", respectively, is considered advisable from the point of view of consistency with the usual statistical nomenclature. When data are given in tabular form, with frequencies indicated by numbers, then they may properly be said to be set out in the form of a *table*. When, however, the same data are distributed in a chart which partakes of the nature of a graph, with the data indicated by horizontal or vertical linear extensions, or by a curve connecting points corresponding to quantities, then it is more proper to call such a graphic representation of the data a *distribution*.

Section IV

FUNDAMENTAL USES OF THE UNILITERAL FREQUENCY DISTRIBUTION

12. The four facts which can be determined from a study of the uniliteral frequency distribution for a cryptogram. *a.* The following four facts (to be explained subsequently) can usually be determined from an inspection of the uniliteral frequency distribution for a given cipher message of average length, composed of letters:

(1) Whether the cipher belongs to the substitution or the transposition class;

(2) If to the former, whether it is monoalphabetic or polyalphabetic in character;

(3) If monoalphabetic, whether the cipher alphabet is a standard cipher alphabet or a mixed cipher alphabet;

(4) If standard, whether it is a direct or reversed standard cipher alphabet.

b. For immediate purposes the first two of the foregoing determinations are quite important and will be discussed in detail in the next two subparagraphs; the other two determinations will be touched upon very briefly, leaving their detailed discussion for subsequent sections of the text.

13. Determining the class to which a cipher belongs.—*a.* The determination of the class to which a cipher belongs is usually a relatively easy matter because of the fundamental difference in the nature of transposition and of substitution as cryptographic processes. In a transposition cipher the original letters of the plain text have merely been rearranged, without any change whatsoever in their identities, that is, in the conventional values they have in the normal alphabet. Hence, the numbers of vowels (A, E, I, O, U, Y), high-frequency consonants (D, N, R, S, T), medium-frequency consonants (B, C, F, G, H, L, M, P, V, W), and low-frequency consonants (J, K, Q, X, Z) are exactly the same in the cryptogram as they are in the plain-text message. Therefore, the percentages of vowels, high, medium, and low-frequency consonants are the same in the transposed text as in the equivalent plain text. In a substitution cipher, on the other hand, the identities of the original letters of the plain text have been changed, that is, the conventional values they have in the normal alphabet have been altered. Consequently, if a count is made of the various letters present in such a cryptogram, it will be found that the number of vowels, high, medium, and low-frequency consonants will usually be quite different in the cryptogram from what they are in the original plain-text message. Therefore, the percentages of vowels, high, medium, and low-frequency consonants are usually quite different in the substitution text from what they are in the equivalent plain text. From these considerations it follows that if in a specific cryptogram the percentages of vowels, high, medium, and low-frequency consonants are approximately the same as would be expected in normal plain text, the cryptogram *probably* belongs to the transposition class; if these percentages are quite different from those to be expected in normal plain text the cryptogram *probably* belongs to the substitution class.

b. In the preceding subparagraph the word "probably" was emphasized by italicizing it, for there can be no certainty in every case of this determination. *Usually* these percentages in a transposition cipher are close to the normal percentages for plain text; *usually*, in a substitution cipher, they are far different from the normal percentages for plain text. But occasionally a cipher message is encountered which is difficult to classify with a reasonable degree of certainty because the message is too short for the general principles of frequency to manifest themselves. It is clear that if in actual messages there were no variation whatever from the normal vowel and consonant percentages given in Table 3, the determination of the class to which a specific cryptogram belongs would be an extremely simple matter. But unfortunately there is always some variation or deviation from the normal. Intuition suggests that as messages decrease in length there may be a greater and greater departure from the normal proportions of vowels, high, medium, and low-frequency consonants, until in very short messages the normal proportions may not hold at all. Similarly, as messages increase in length there may be a lesser and lesser departure from the normal proportions, until in messages totalling a thousand or more letters there may be no difference at all between the actual and the theoretical proportions. But intuition is not enough, for in dealing with specific messages of the length of those commonly encountered in practical work the question sometimes arises as to exactly how much deviation (from the normal proportions) may be allowed for in a cryptogram which shows a considerable amount of deviation from the normal and which might still belong to the transposition rather than to the substitution class.

c. Statistical studies have been made on this matter and some graphs have been constructed thereon. These are shown in Charts 1–4 in the form of simple curves, the use of which will now be explained. Each chart contains two curves marking the lower and upper limits, respectively, of the theoretical amount of deviation (from the normal percentages) of vowels or consonants which may be allowable in a cipher believed to belong to the transposition class.

d. In Chart 1, curve V_1 marks the lower limit of the theoretical amount of deviation from the normal number of vowels to be expected in a message of given length; curve V_2 marks the upper limit of the same thing. Thus, for example, in a message of 100 letters in plain English there should be between 33 and 47 vowels (A E I O U Y). Likewise, in Chart 2 curves H_1 and H_2 mark the lower and upper limits as regards the high-frequency consonants. In a message of 100 letters there should be between 28 and 42 high-frequency consonants (D N R S T). In Chart 3, curves M_1 and M_2 mark the lower and upper limits as regards the medium-frequency consonants. In a message of 100 letters there should be between 17 and 31 medium-frequency consonants (B C F G H L M P V W). Finally, in Chart 4, curves L_1 and L_2 mark the lower and upper limits as regards the low-frequency consonants. In a message of 100 letters there should be between 0 and 3 low-frequency consonants (J K Q X Z). In using the charts, therefore, one finds the point of intersection of the vertical coordinate corresponding to the length of the message, with the horizontal coordinate corresponding to (1) the number of vowels, (2) the number of high-frequency consonants, (3) the number of medium-frequency consonants, and (4) the number of low-frequency consonants actually counted in the message. If all four points of intersection fall within the area delimited by the respective curves, then the number of vowels, high, medium, and low-frequency consonants corresponds with the number theoretically expected in a normal plain-text message of the same length; since the message under investigation is not plain text, it follows that the cryptogram may certainly be classified as a transposition cipher. On the other hand, if one or more of these points of intersection falls outside the area delimited by the respective curves, it follows that the cryptogram is probably a substitution cipher. The distance that the point of intersection falls outside the area delimited by these curves is a more or less rough measure of the improbability of the cryptogram's being a transposition cipher.

e. Sometimes a cryptogram is encountered which is hard to classify with certainty even with the foregoing aids, because it has been consciously prepared with a view to making the classification difficult. This can be done either by selecting peculiar words (as in "trick cryptograms") or by employing a cipher alphabet in which letters of *approximately similar normal frequencies* have been interchanged. For example, E may be replaced by O, T by R, and so on, thus yielding

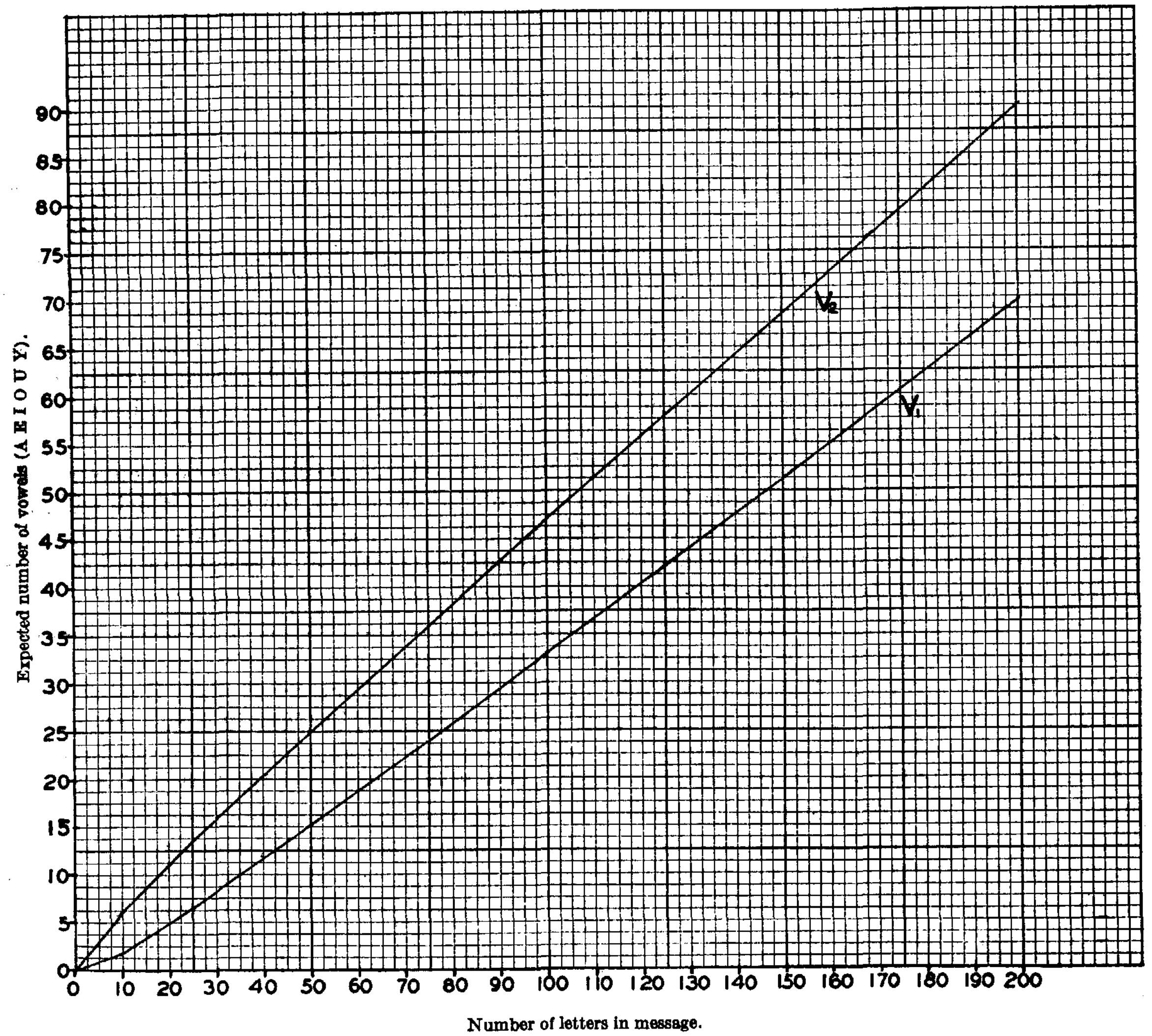


CHART No. 1.—Curves marking the lower and upper limits of the theoretical amount of deviation from the normal number of vowels to be expected in messages of various lengths. (See Par. 13*d*.)

a cryptogram giving external indications of being a transposition cipher but which is really a substitution cipher. If the cryptogram is not too short, a close study will usually disclose what has been done, as well as the futility of so simple a subterfuge.

f. In the majority of cases, in practical work, the determination of the class to which a cipher of average length belongs can be made from a mere inspection of the message, after the cryptanalyst has acquired a familiarity with the normal appearance of transposition and of substitution ciphers. In the former case, his eyes very speedily note many high-frequency letters, such as E, T, N, R, O, and S, with the absence of low-frequency letters, such as J, K, Q, X,

and Z; in the latter case, his eyes just as quickly note the presence of many low-frequency letters, and a corresponding absence of the usual high-frequency letters.

g. Another rather quickly completed test, in the case of the simpler varieties of ciphers, is to look for *repetitions of groups of letters*. As will become apparent very soon, recurrences of syllables, entire words and short phrases constitute a characteristic of all normal plain text. Since a transposition cipher involves a change in the *sequence* of the letters composing a plain-

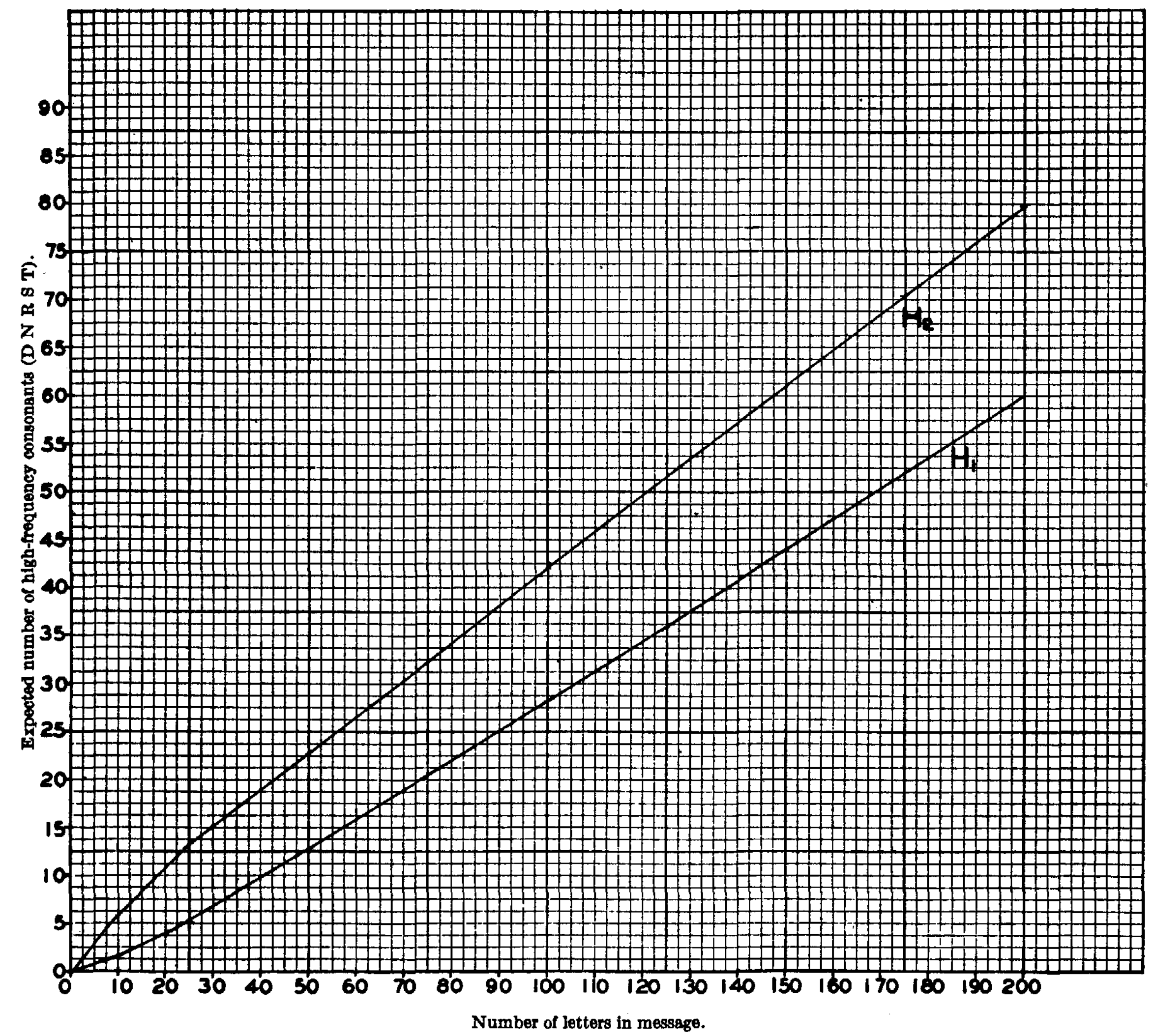


CHART No. 2.—Curves marking the lower and upper limits of the theoretical amount of deviation from the normal number of high-frequency consonants to be expected in messages of various lengths. (See Par. 13*d*.)

text message, such recurrences are broken up so that the cipher text no longer will show repetitions of more or less lengthy sequences of letters. But if a cipher message does show many repetitions and these are of several letters in length, say over four or five, the conclusion is at once warranted that the cryptogram is most probably a substitution and not a transposition cipher. However, for the beginner in cryptanalysis, it will be advisable to make the uniliteral frequency distribution, and note the frequencies of the vowels, the high, medium, and low-frequency consonants. Then, referring to Charts 1 to 4, he should carefully note whether or not the observed frequencies for these categories of letters fall within the limits of the theoretical frequencies for a normal plain-text message of the same length, and be guided accordingly.

h. It is obvious that the foregoing rule applies only to ciphers composed wholly of letters. If a message is composed entirely of figures, or of arbitrary signs and symbols, or of intermixtures

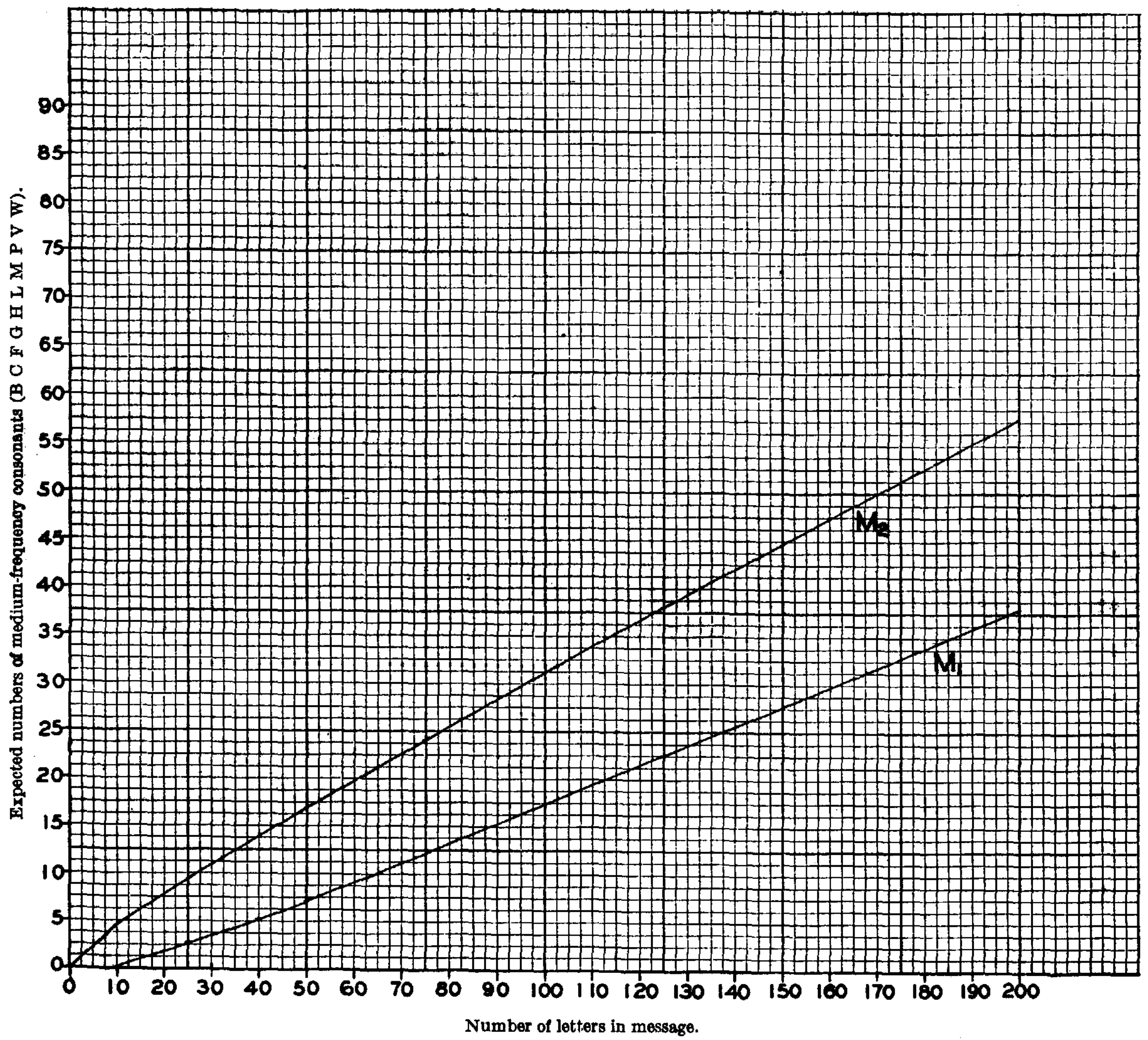


CHART No. 3.—Curves marking the lower and upper limits of the theoretical amount of deviation from the normal number of medium-frequency consonants to be expected in messages of various lengths. (See Par. 13*d*.)

of letters, figures and other symbols, it is immediately apparent that the cryptogram is a substitution cipher.

i. Finally, it should be mentioned that there are certain kinds of cryptograms whose class cannot be determined by the method set forth in subparagraphs *b*, *c*, *d* above. These exceptions will be discussed in a subsequent section of this text.[1]

14. Determining whether a substitution cipher is monoalphabetic or polyalphabetic.—*a*. It will be remembered that a monoalphabetic substitution cipher is one in which a single cipher alphabet is employed throughout the whole message, that is, a given plain-text letter is invariably

[1] Par. 47.

represented throughout the message by one and the same letter in the cipher text. On the other hand, a polyalphabetic substitution cipher is one in which two or more cipher alphabets are employed within the same message; that is, a given plain-text letter may be represented by two or more different letters in the cipher text, according to some rule governing the selection of the equivalent to be used in each case. From this it follows that a single cipher letter may represent two or more different plain-text letters.

b. It is easy to see why and how the appearance of the uniliteral frequency distribution for a substitution cipher may be used to determine whether the cryptogram is monoalphabetic or polyalphabetic in character. The normal distribution presents marked crests and troughs by

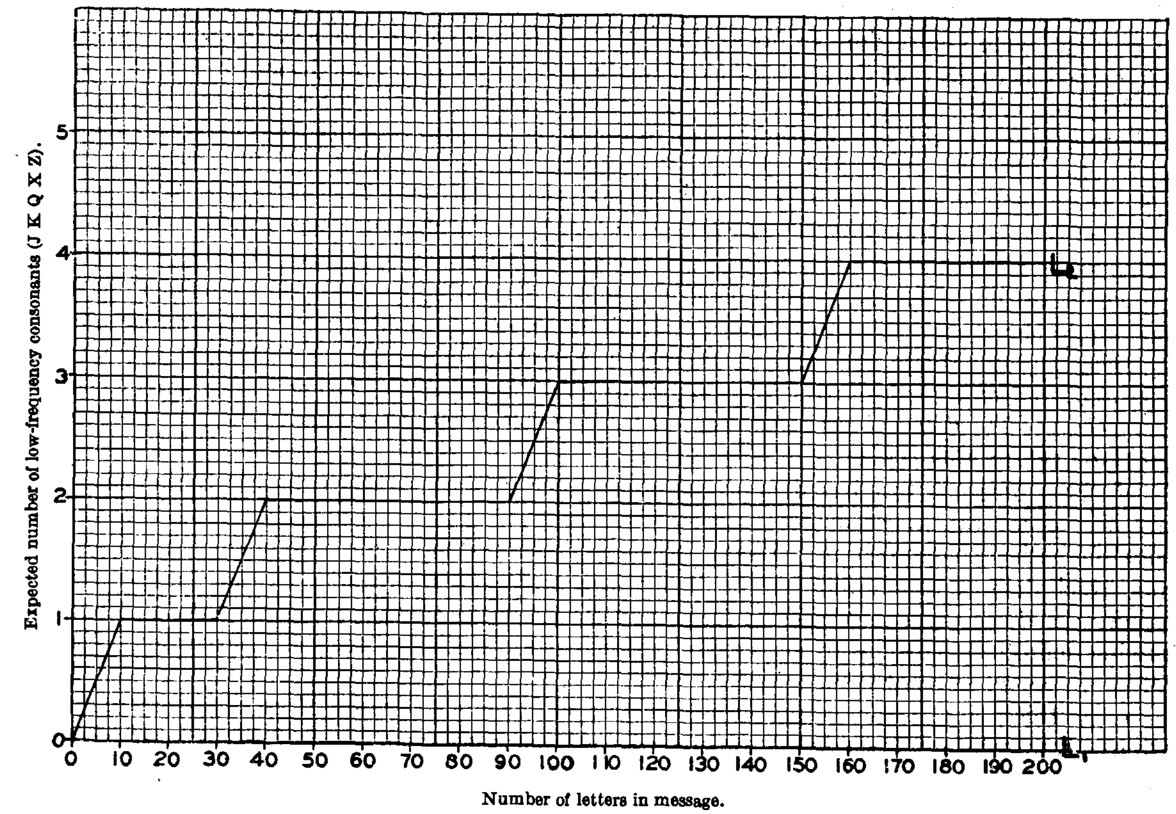


CHART No. 4.—Curves marking the lower and upper limits of the theoretical amount of deviation from the normal number of low-frequency consonants to be expected in messages of various lengths. (See Par. 13*d*.)

virtue of two circumstances. First, the elementary sounds which the symbols represent are used with greatly varying frequencies, it being one of the striking characteristics of every alphabetic language that its elementary sounds are used with greatly varying frequencies.[2] In the second place, except for orthographic aberrations peculiar to certain languages (conspicuously, English and French), each such sound is represented by the same symbol. It follows, therefore, that since in a monoalphabetic substitution cipher each different plain-text letter (=elementary sound) is represented by one and only one cipher letter (=elementary symbol), the uniliteral frequency distribution for such a cipher message must also exhibit the irregular crest and trough appearance of the normal distribution, but with only this important modification—*the absolute*

[2] The student who is interested in this phase of the subject may find the following reference of value: Zipf' G. K., *Selected Studies of the Principle of Relative Frequency in Language*, Cambridge, Mass., 1932.

positions of the crests and troughs will not be the same as in the normal. That is, the letters accompanying the crests and the troughs in the distribution for the cryptogram will be different from those accompanying the crests and the troughs in the normal distribution. But the marked irregularity of the distribution, the presence of accentuated crests and troughs, is in itself an indication that each symbol or cipher letter always represents the same plain-text letter in that cryptogram. Hence the general rule: *A marked crest and trough appearance in the uniliteral frequency distribution for a given cryptogram indicates that a single cipher alphabet is involved and constitutes one of the tests for a monoalphabetic substitution cipher.*

c. On the other hand, suppose that in a cryptogram each cipher letter represents several different plain-text letters. Some of them are of high frequency, others of low frequency. The net result of such a situation, so far as the uniliteral frequency distribution for the cryptogram is concerned, is to prevent the appearance of any marked crests and troughs and to tend to reduce the elements of the distribution to a more or less common level. This imparts a "flattened out" appearance to the distribution. For example, in a certain cryptogram of polyalphabetic construction, $K_c=E_p$, G_p, and J_p; $R_c=A_p$, D_p, and B_p; $X_c=O_p$, L_p, and F_p. The frequencies of K_c, R_c, and X_c will be approximately equal because the summations of the frequencies of the several plain-text letters each of these cipher letters represents at different times will be about equal. If this same phenomenon were true of all the letters of the cryptogram, it is clear that the frequencies of the 26 letters, when shown by means of the ordinary uniliteral frequency distribution, would show no striking differences and the distribution would have the flat appearance of a typical polyalphabetic substitution cipher. Hence, the general rule: *The absence of marked crests and troughs in the uniliteral frequency distribution indicates that two or more cipher alphabets are involved. The flattened-out appearance of the distribution constitutes one of the tests for a polyalphabetic substitution cipher.*

d. The foregoing test based upon the appearance of the frequency distribution constitutes only one of several means of determining whether a substitution cipher is monoalphabetic or polyalphabetic in composition. It can be employed in cases yielding frequency distributions from which definite conclusions can be drawn with more or less certainty by mere ocular examination. In those cases in which the frequency distributions contain insufficient data to permit drawing definite conclusions by such examination, certain statistical tests can be applied. These will be discussed in a subsequent text.

e. At this point, however, one additional test will be given because of its simplicity of application. It may be employed in testing messages up to 200 letters in length, it being assumed that in messages of greater length ocular examination of the frequency distribution offers little or no difficulty. This test concerns the *number of blanks* in the frequency distribution, that is, the number of letters of the alphabet which are entirely absent from the message. It has been found from statistical studies that rather definite "laws" govern the theoretically expected number of blanks in normal plain-text messages and in frequency distributions for cryptograms of different natures and of various sizes. The results of certain of these studies have been embodied in Chart 5.

f. This chart contains two curves. The one labeled *P* applies to the average number of blanks theoretically expected in frequency distributions based upon normal plain-text messages of the indicated lengths. The other curve, labeled *R*, applies to the average number of blanks theoretically expected in frequency distributions based upon perfectly *random* assortments of letters; that is, assortments such as would be found by random selection of letters out of a hat containing thousands of letters, all of the 26 letters of the alphabet being present in equal proportions, each letter being replaced after a record of its selection has been made. Such random assortments correspond to polyalphabetic cipher messages in which the number of cipher alphabets is so large that if uniliteral frequency distributions are made of the letters, the distributions are practically identical with those which are obtained by random selections of letters out of a hat.

g. In using this chart, one finds the point of intersection of the vertical coordinate corresponding to the length of the message, with the horizontal coordinate corresponding to the observed number of blanks in the distribution for the message. If this point of intersection falls closer to curve *P* than it does to curve *R*, the number of blanks in the message approximates or corresponds more closely to the number theoretically expected in a plain-text message than it does to a random (cipher-text) message of the same length; therefore, this is evidence that the cryptogram is monoalphabetic. Conversely, if this point of intersection falls closer to curve *R*

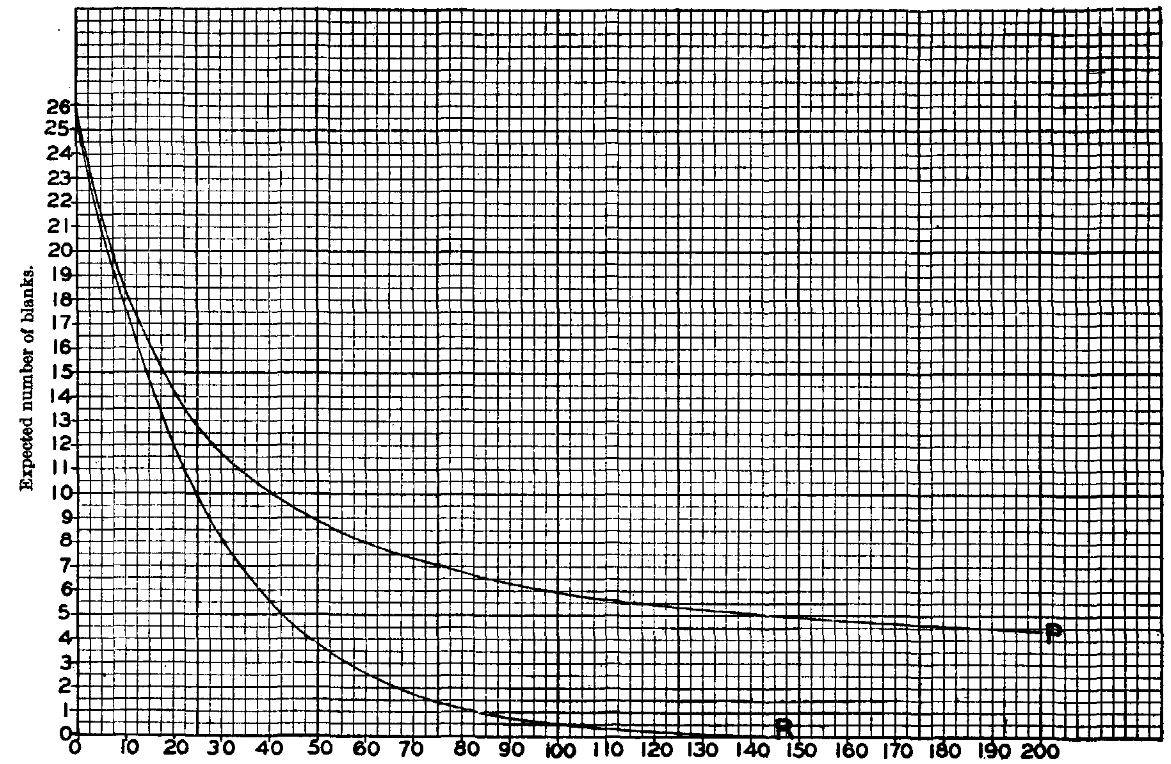


CHART No. 5.—Curves showing the average number of blanks theoretically expected in distributions for plain text (P) and for random text (R) for messages of various lengths. (See Par. 14*f*.)

than to curve *P*, the number of blanks in the message approximates or corresponds more closely to the number theoretically expected in a random text than it does to a plain-text message of the same length; therefore, this is evidence that the cryptogram is polyalphabetic.

h. Practical examples of the use of this chart will be given in some of the illustrative messages to follow.

15. Determining whether the cipher alphabet is a standard, or a mixed cipher alphabet.—*a.* Assuming that the uniliteral frequency distribution for a given cryptogram has been made, and that it shows clearly that the cryptogram is a substitution cipher and is monoalphabetic in character, a consideration of the nature of standard cipher alphabets[3] almost makes it obvious how an inspection of the distribution will disclose whether the cipher alphabet involved is a standard cipher alphabet or a mixed cipher alphabet. If the crests and troughs of the distribu-

[3] See Sec. VIII, *Elementary Military Cryptography.*

tion occupy positions which correspond to the *relative* positions they occupy in the normal frequency distribution, then the cipher alphabet is a standard cipher alphabet. If this is not the case, then it is highly probable that the cryptogram has been prepared by the use of a mixed cipher alphabet.

b. A mechanical test may be applied in doubtful cases arising from lack of material available for study. Just what this test involves, and an illustration of its application will be given in the next section, using specific examples.

16. Determining whether the standard cipher alphabet is direct or reversed.—Assuming that the frequency distribution for a given cryptogram shows clearly that a standard cipher alphabet is involved, the determination as to whether the alphabet is direct or reversed can also be made by inspection, since the difference between the two is merely a matter of the *direction* in which the sequence of crests and troughs progresses—to the right, as in normal reading or writing, or the left. In a direct cipher alphabet the direction in which the crests and troughs of the distribution should be read is the normal direction, from left to right; in a reversed cipher alphabet this direction is reversed, from right to left.

Section V

UNILITERAL SUBSTITUTION WITH STANDARD CIPHER ALPHABETS

17. Principles of solution by construction and analysis of the uniliteral frequency distribution.—*a.* Standard cipher alphabets are of two sorts, direct and reversed. The analysis of monoalphabetic cryptograms prepared by their use follows almost directly from a consideration of the nature of such alphabets. Since the cipher component of a standard cipher alphabet consists either of the normal sequence merely displaced 1, 2, 3, . . . intervals from the normal point of coincidence, or of the normal sequence proceeding in a reversed-normal direction, it is obvious that the uniliteral frequency distribution for a cryptogram prepared by means of such a cipher alphabet employed monoalphabetically will show crests and troughs whose *relative* positions and frequencies will be exactly the same as in the uniliteral frequency distribution for the plain text of that cryptogram. The only thing that has happened is that the whole set of crests and troughs of the distribution has been displaced to the right or left of the position it occupies in the distribution for the plain text; or else the successive elements of the whole set progress in the opposite direction. Hence, it follows that the correct determination of the plain-text value of the letter marking *any* crest or trough of the uniliteral frequency distribution will result at one stroke in the correct determination of the plain-text values of *all* the remaining 25 letters respectively marking the other crests and troughs in that distribution. Thus, having determined the value of a single element of the cipher component of the cipher alphabet, the values of all the remaining letters of the cipher component are automatically solved at one stroke. In more simple language, the correct determination of the value of a single letter of the cipher text automatically gives the values of the other 25 letters of the cipher text. The problem thus resolves itself into a matter of selecting that point of attack which will most quickly or most easily lead to the determination of the value of *one* cipher letter. The single word *identification* will hereafter be used for the phrase "determination of the value of a cipher letter"; to *identify* a cipher letter is to find its plain-text value.

b. It is obvious that the easiest point of attack is to assume that the letter marking the crest of greatest frequency in the frequency distribution for the cryptogram represents E_p. Proceeding from this initial point, the identifications of the remaining cipher letters marking the other crests and troughs are tentatively made on the basis that the letters of the cipher component proceed in accordance with the normal alphabetic sequence, either direct or reversed. If the actual frequency of each letter marking a crest or a trough approximates to a fairly close degree the normal theoretical frequency of the assumed plain-text equivalent, then the initial identification $\theta_c = E_p$ may be *assumed to be correct* and therefore the derived identifications of the other cipher letters may be assumed to be correct. If the original starting point for assignment of plain-text values is not correct, or if the direction of "reading" the successive crests and troughs of the distribution is not correct, then the frequencies of the other 25 cipher letters will not correspond to or even approximate the normal theoretical frequencies of their hypothetical plain-text equivalents on the basis of the initial identification. A new initial point, that is, a different cipher equivalent must then be selected to represent E_p; or else the direction of "reading" the crests and troughs must be reversed. This procedure, that is, the attempt to make the actual frequency relations exhibited by uniliteral frequency distribution for a given cryptogram conform to the theoretical frequency relations of the normal frequency distribution in an effort to solve the cryptogram, is referred to technically as "fitting the actual uniliteral frequency distribution for a cryptogram to the thoretical uniliteral frequency distribution for normal plain text", or, more briefly, as *"fitting the frequency distribution for the cryptogram to the normal frequency distribution"*, or, still more briefly, *"fitting the distribution to the normal."* In statistical work the expression commonly employed in connection with this process of fitting an actual distribution to a theoretical one is "testing the goodness of fit." The goodness of fit may be stated in various ways, mathematical in character.

c. In fitting the actual distribution to the normal, it is necessary to regard the cipher component (that is, the letters A . . . Z marking the successive crests and troughs of the distribution) as partaking of the nature of a wheel or sequence closing in upon itself, so that no matter with what crest or trough one starts, the spatial and frequency relations of the crests and troughs are constant. This manner of regarding the cipher component as being cyclic in nature is valid *because it is obvious that the relative positions and frequencies of the crests and troughs of any uniliteral-frequency distribution must remain the same regardless of what letter is employed as the initial point of the distribution.* Fig. 5 gives a clear picture of what is meant in this connection, as applied to the normal frequency distribution.

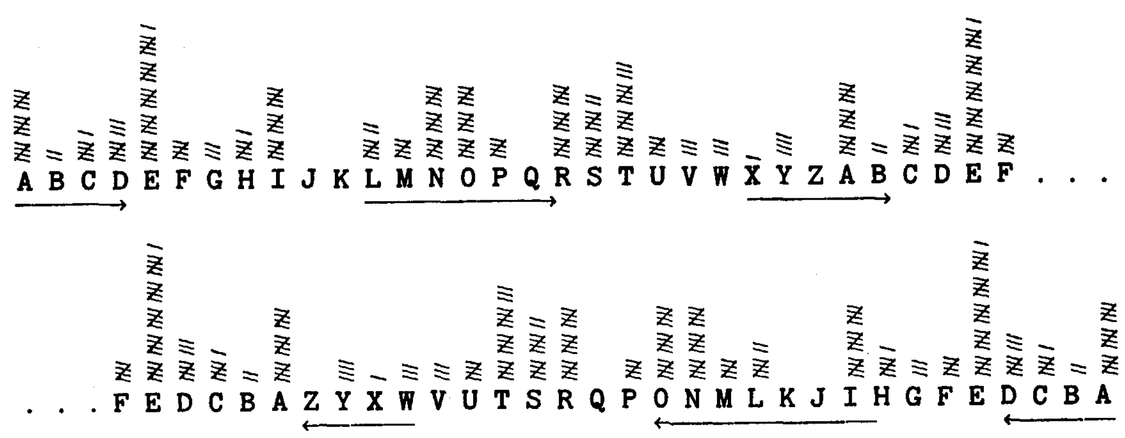


FIGURE 5.

d. In the third sentence of subparagraph *b*, the phrase "assumed to be correct" was advisedly employed in describing the results of the attempt to fit the distribution to the normal, because the final test of the goodness of fit in this connection (that is, of the correctness of the assignment of values to the crests and troughs of the distribution) is whether the *consistent* substitution of the plain-text values of the cipher characters in the cryptogram will yield intelligible plain text. If this is not the case, then no matter how close the approximation between actual and theoretical frequencies is, no matter how well the actual frequency distribution fits the normal, the only possible inferences are that (1) either the closeness of the fit is a pure coincidence in this case, and that another equally good fit may be obtained from the same data, or else (2) the cryptogram involves something more than simple monoalphabetic substitution by

means of a single standard cipher alphabet. For example, suppose a transposition has been applied in addition to the substitution. Then, although an excellent correspondence between the uniliteral frequency distribution and the normal frequency distribution has been obtained, the substitution of the cipher letters by their assumed equivalents will still not yield plain text. However, aside from such cases of double encipherment, instances in which the uniliteral frequency distribution may be easily fitted to the normal frequency distribution and in which at the same time an attempted simple substitution fails to yield intelligible text are rare. It may be said that, in practical operations whenever the uniliteral frequency distribution can be made to fit the normal frequency distribution, substitution of values will result in solution; and, as a corollary, whenever the uniliteral frequency distribution cannot be made to fit the normal frequency distribution, the cryptogram does not represent a case of simple, monoalphabetic substitution by means of a standard alphabet.

18. Theoretical example of solution.—*a.* The foregoing principles will become clearer by noting the cryptographing and solution of a theoretical example. The following message is to be cryptographed.

HOSTILE FORCE ESTIMATED AT ONE REGIMENT INFANTRY AND TWO PLATOONS CAVALRY MOVING SOUTH ON QUINNIMONT PIKE STOP HEAD OF COLUMN NEARING ROAD JUNCTION SEVEN THREE SEVEN COMMA EAST OF GREENACRE SCHOOL FIRED UPON BY OUR PATROLS STOP HAVE DESTROYED BRIDGE OVER INDIAN CREEK.

b. First, solely for purposes of demonstrating certain principles, the uniliteral frequency distribution for this message is presented in Figure 6.

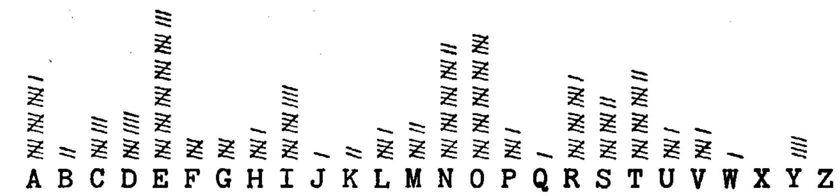


FIGURE 6.

c. Now let the foregoing message be cryptographed monoalphabetically by the following cipher alphabet, yielding the cryptogram and the frequency distribution shown below.

Plain............	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
Cipher............	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F

Plain..............	HOSTI	LEFOR	CEEST	IMATE	DATON	EREGI	MENTI	NFANT	RYAND
Cipher..............	NUYZO	RKLUX	IKKYZ	OSGZK	JGZUT	KXKMO	SKTZO	TLGTZ	XEGTJ
Plain..............	TWOPL	ATOON	SCAVA	LRYMO	VINGS	OUTHO	NQUIN	NIMON	TPIKE
Cipher..............	ZCUVR	GZUUT	YIGBG	RXESU	BOTMY	UAZNU	TWAOT	TOSUT	ZVOQK
Plain..............	STOPH	EADOF	COLUM	NNEAR	INGRO	ADJUN	CTION	SEVEN	THREE
Cipher..............	YZUVN	KGJUL	IURAS	TTKGX	OTMXU	GJPAT	IZOUT	YKBKT	ZNXKK
Plain..............	SEVEN	COMMA	EASTO	FGREE	NACRE	SCHOO	LFIRE	DUPON	BYOUR
Cipher..............	YKBKT	IUSSG	KGYZU	LMXKK	TGIXK	YINUU	RLOXK	JAVUT	HEUAX
Plain..............	PATRO	LSSTO	PHAVE	DESTR	OYEDB	RIDGE	OVERI	NDIAN	CREEK
Cipher..............	VGZXU	RYYZU	VNGBK	JKYZX	UEKJH	XOJMK	UBKXO	TJOGT	IXKKQ

148274—38——3

CRYPTOGRAM

N U Y Z O	R K L U X	I K K Y Z	O S G Z K	J G Z U T	K X K M O
S K T Z O	T L G T Z	X E G T J	Z C U V R	G Z U U T	Y I G B G
R X E S U	B O T M Y	U A Z N U	T W A O T	T O S U T	Z V O Q K
Y Z U V N	K G J U L	I U R A S	T T K G X	O T M X U	G J P A T
I Z O U T	Y K B K T	Z N X K K	Y K B K T	I U S S G	K G Y Z U
L M X K K	T G I X K	Y I N U U	R L O X K	J A V U T	H E U A X
V G Z X U	R Y Y Z U	V N G B K	J K Y Z X	U E K J H	X O J M K
U B K X O	T J O G T	I X K K Q			

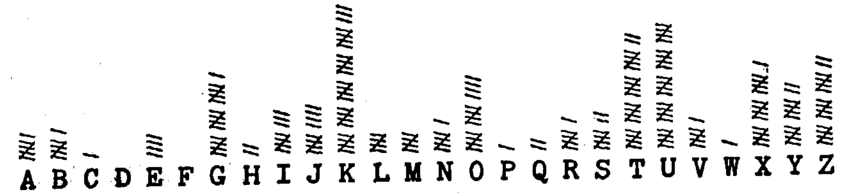


FIGURE 7

d. Let the student now compare Figs. 6 and 7, which have been superimposed in Fig. 8 for convenience in examination. Crests and troughs are present in both distributions; moreover their relative positions and frequencies have not been changed in the slightest particular. Only the absolute position of the sequence as a whole has been displaced six intervals to the right in Fig. 7, as compared with the absolute position of the sequence in Fig. 6.

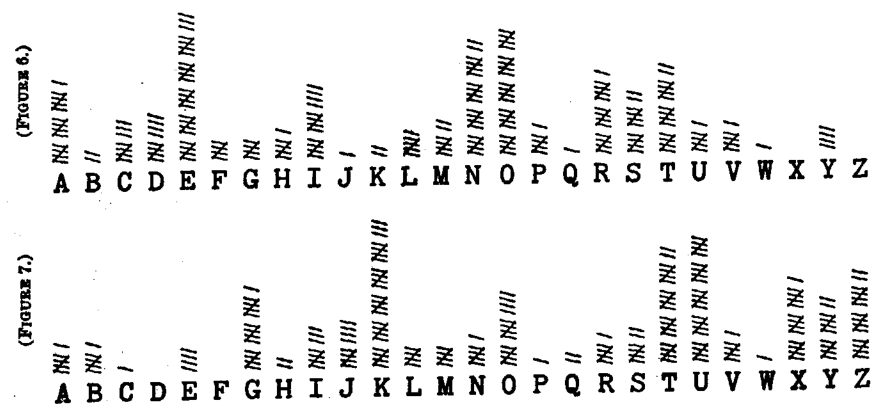


FIGURE 8.

e. If the two distributions are compared in detail the student will clearly understand how easy the solution of the cryptogram would be to one who knew nothing about how it was prepared. For example, the frequency of the highest crest, representing E_p in Fig. 6 is 28; at an interval of four letters before E_p there is another crest representing A_p with frequency 16. Between A and E there is a trough, representing the low-frequency letters B, C, D. On the other side of E, at an interval of four letters, comes another crest, representing I with frequency 14. Between E and I there is another trough, representing the low-frequency letters F, G, H. Compare these crests and troughs with their homologous crests and troughs in Fig. 7. In the latter, the letter K marks the highest crest in the distribution with a frequency of 28; four letters before K there is another crest, frequency 16, and four letters on the other side of K there is another crest, frequency

14. Troughs corresponding to B, C, D and F, G, H are seen at H, I, J and L, M, N in Fig. 7. In fact, the two distributions may be made to coincide exactly, by shifting the frequency distribution for the cryptogram six intervals to the left with respcet to the distribution for the equivalent plain-text message, as shown herewith.

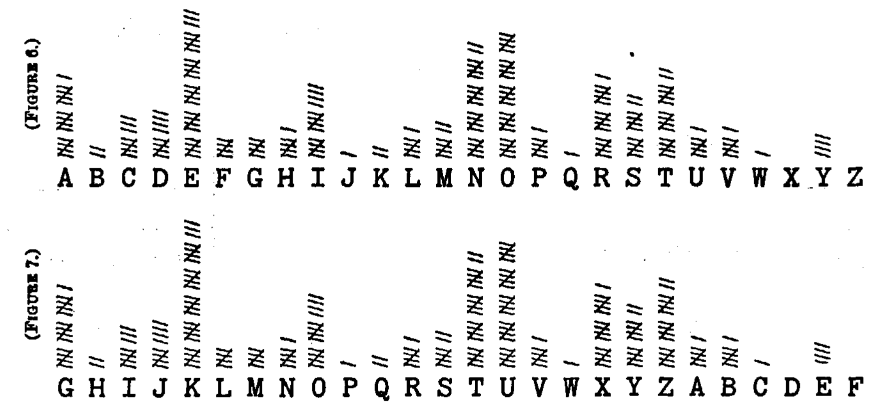


FIGURE 9.

f. Let us suppose now that nothing is known about the cryptographing process, and that only the cryptogram and its uniliteral frequency distribution is at hand. It is clear that simply bearing in mind the spatial relations of the crests and troughs in a normal frequency distribution would enable the cryptanalyst to fit the distribution to the normal in this case. He would naturally first assume that $G_c=A_p$, from which it would follow that if a direct standard alphabet is involved, $H_c=B_p$, $I_c=C_p$, and so on, yielding the following (tentative) deciphering alphabet:

Cipher............	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
Plain.............	U V W X Y Z A B C D E F G H I J K L M N O P Q R S T

g. Now comes the final test: If these assumed values are substituted in the cipher text, the plain text immediately appears. Thus:

N U Y Z O	R K L U X	I K K Y Z	O S G Z K	J G Z U T	etc.
H O S T I	L E F O R	C E E S T	I M A T E	D A T O N	etc.

h. It should be clear, therefore, that the selection of G_c to represent A_p in the cryptographing process has absolutely no effect upon the relative spatial and frequency relations of the crests and troughs of the frequency distribution for the cryptogram. If Q_c had been selected to represent A_p, these relations would still remain the same, the whole series of crests and troughs being merely displaced further to the right of the positions they occupy when $G_c=A_p$.

19. Practical example of solution by the frequency method.—*a. The case of direct standard alphabet ciphers.*—(1) The following cryptogram is to be solved by applying the foregoing principles:

I B M Q O	P B I U O	M B B G A	J C Z O F	M U U Q B	A J C Z O
Z W I L N	Q T T M L	E Q B P U	I Z K P Q	V O Q V N	I V B Z G

(2) From the presence of repetitions and so many low-frequency letters such as B, Q, and Z it is at once suspected that this is a substitution cipher. But to illustrate the steps that must be taken in difficult cases in order to be certain in this respect, a uniliteral frequency distribution

he will note that the relative positions and extensions of the crests and troughs are identical; they merely progress in opposite directions.

20. Solution by completing the plain-component sequence.—*a. The case of direct standard alphabet ciphers.*—(1) The foregoing method of analysis, involving as it does the construction of a uniliteral frequency distribution, was termed a *solution by the frequency method* because it involves the construction of a frequency distribution and its study. There is, however, another method which is much more rapid, almost wholly mechanical, and which, moreover, does not necessitate the construction or study of any frequency distribution whatever. An understanding of the method follows from a consideration of the method of encipherment of a message by the use of a single, direct standard cipher alphabet.

(2) Note the following encipherment:

Message.......... REPEL INVADING CAVALRY

ENCIPHERING ALPHABET

```
Plain............ 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
Cipher........... G H I J K L M N O P Q R S T U V W X Y Z A B C D E F
```

ENCIPHERMENT

```
Plain text....... R E P E L   I N V A D I N G   C A V A L R Y
Cryptogram....... X K V K R   O T B G J O T M   I G B G R X E
```

CRYPTOGRAM

```
X K V K R   O T B G J   O T M I G   B G R X E
```

(3) The enciphering alphabet shown above represents a case wherein the sequence of letters of both components of the cipher alphabet is the normal sequence, with the sequence forming the cipher component merely shifted six intervals in retard (or 20 intervals in advance) of the position it occupies in the normal alphabet. If, therefore, two strips of paper bearing the letters of the normal sequence, equally spaced, are regarded as the two components of the cipher alphabet and are juxtaposed at all of the 25 possible points of coincidence, it is obvious that one of these 25 juxtapositions *must* correspond to the actual juxtaposition shown in the enciphering alphabet directly above.[2] It is equally obvious that if a record were kept of the results obtained by applying the values given at each juxtaposition to the letters of the cryptogram, one of these results would yield the plain text of the cryptogram.

(4) Let the work be systematized and the results set down in an orderly manner for examination. It is obviously unnecessary to juxtapose the two components so that $A_c=A_p$, for on the assumption of a direct standard alphabet, juxtaposing two direct normal components at their normal point of coincidence merely yields plain text. The next possible juxtaposition, therefore, is $A_c=B_p$. Let the juxtaposition of the two sliding strips therefore be $A_c=B_p$, as shown here:

```
Plain............ ABCDEFGHIJKLMNOPQRSTUVWXYZABCDEFGHIJKLMNOPQRSTUVWXYZ
Cipher...........  ABCDEFGHIJKLMNOPQRSTUVWXYZ
```

The values given by this juxtaposition are substituted for the first 20 letters of the cryptogram and the following results are obtained.

```
Cryptogram...................  X K V K R   O T B G J   O T M I G   B G R X E
1st Test—"Plain text"......  Y L W L S   P U C H K   P U N J H   C H S Y F
```

[2] One of the strips should bear the sequence repeated. This permits juxtaposing the two sequences at all 26 possible points of coincidence so as to have a complete cipher alphabet showing at all times.

This certainly is not intelligible text; obviously, the two components were not in the position indicated in this first test. The cipher component is therefore slid one interval to the right, making $A_c=C_p$, and a second test is made. Thus

```
Plain................ ABCDEFGHIJKLMNOPQRSTUVWXYZABCDEFGHIJKLMNOPQRSTUVWXYZ
Cipher...............   ABCDEFGHIJKLMNOPQRSTUVWXYZ
Cryptogram...................  X K V K R   O T B G J   O T M I G   B G R X E
2d Test—"Plain text"......  Z M X M T   Q V D I L   Q V O K I   D I T Z G
```

Neither does the second test result in disclosing any plain text. But, if the results of the two tests are studied a phenomenon that at first seems quite puzzling comes to light. Thus, suppose the results of the two tests are superimposed in this fashion.

```
Cryptogram.....................  X K V K R   O T B G J   O T M I G   B G R X E
1st Test—"Plain text"....  Y L W L S   P U C H K   P U N J H   C H S Y F
2nd Test—"Plain text"....  Z M X M T   Q V D I L   Q V O K I   D I T Z G
```

(5) Note what has happened. The net result of the two experiments was merely to continue the normal sequence begun by the cipher letters at the heads of the several *columns*. It is obvious that if the normal sequence is completed in each column *the results will be exactly the same as though the whole set of 25 possible tests had actually been performed.* Let the columns therefore be completed, as shown in Fig. 11.

```
 X K V K R O T B G J O T M I G B G R X E
 ---------------------------------------
 Y L W L S P U C H K P U N J H C H S Y F
 Z M X M T Q V D I L Q V O K I D I T Z G
 A N Y N U R W E J M R W P L J E J U A H
 B O Z O V S X F K N S X Q M K F K V B I
 C P A P W T Y G L O T Y R N L G L W C J
 D Q B Q X U Z H M P U Z S O M H M X D K
 E R C R Y V A I N Q V A T P N I N Y E L
 F S D S Z W B J O R W B U Q O J O Z F M
 G T E T A X C K P S X C V R P K P A G N
 H U F U B Y D L Q T Y D W S Q L Q B H O
 I V G V C Z E M R U Z E X T R M R C I P
 J W H W D A F N S V A F Y U S N S D J Q
 K X I X E B G O T W B G Z V T O T E K R
 L Y J Y F C H P U X C H A W U P U F L S
 M Z K Z G D I Q V Y D I B X V Q V G M T
 N A L A H E J R W Z E J C Y W R W H N U
 O B M B I F K S X A F K D Z X S X I O V
 P C N C J G L T Y B G L E A Y T Y J P W
 Q D O D K H M U Z C H M F B Z U Z K Q X
*R E P E L I N V A D I N G C A V A L R Y
 S F Q F M J O W B E J O H D B W B M S Z
 T G R G N K P X C F K P I E C X C N T A
 U H S H O L Q Y D G L Q J F D Y D O U B
 V I T I P M R Z E H M R K G E Z E P V C
 W J U J Q N S A F I N S L H F A F Q W D
```

FIGURE 11.

An examination of the successive horizontal lines of the diagram discloses *one and only one* line of plain text, that marked by the asterisk and reading R E P E L I N V A D I N G C A V A L R Y.

(6) Since each column in Fig. 11 is nothing but a normal sequence, it is obvious that instead of laboriously writing down these columns of letters every time a cryptogram is to be examined, it would be more convenient to prepare a set of strips each bearing the normal sequence doubled (to permit complete coincidence for an entire alphabet at any setting), and have them available for examining any future cryptograms. In using such a set of sliding strips in order to solve a cryptogram prepared by means of a single direct standard cipher alphabet, or to make a test to determine whether a cryptogram has been so prepared, it is only necessary to "set up" the letters of the cryptogram on the strips, that is, align them in a single row across the strips (by sliding the individual strips up or down). The successive horizontal lines, called *generatrices* (singular, *generatrix*), are then examined in a search for intelligible text. If the cryptogram really belongs to this simple type of cipher, one of the generatrices will exhibit intelligible text all the way across; this text will practically invariably be the plain text of the message. This method of analysis may be termed *a solution by completing the plain-component sequence.* Sometimes it is referred to as "running down" the sequence. The principle upon which the method is based constitutes one of the cryptanalyst's most valuable tools.[3]

b. The case of reversed standard alphabets.—(1) The method described under subpar. *a* may also be applied, in slightly modified form, in the case of a cryptogram enciphered by a single reversed standard alphabet. The basic principles are identical in the two cases.

(2) To show this it is necessary to experiment with two sliding components as before, except that in this case one of the components must be a reversed normal sequence, the other, a direct normal sequence.

(3) Let the two components be juxtaposed A to A, as shown below, and then let the resultant values be substituted for the letters of the cryptogram. Thus:

CRYPTOGRAM

PCRCV YTLGD YTAEG LGVPI

Plain................ ABCDEFGHIJKLMNOPQRSTUVWXYZABCDEFGHIJKLMNOPQRSTUVWXYZ
Cipher.............. ZYXWVUTSRQPONMLKJIHGFEDCBA

Cryptogram................	P C R C V	Y T L G D	Y T A E G	L G V P I
1st Test—"Plain text"...	L Y J Y F	C H P U X	C H A W U	P U F L S

(4) This does not yield intelligible text, and therefore the reversed component is slid one space forward and a second test is made. Thus:

Plain................ ABCDEFGHIJKLMNOPQRSTUVWXYZABCDEFGHIJKLMNOPQRSTUVWXYZ
Cipher.............. ZYXWVUTSRQPONMLKJIHGFEDCBA

Cryptogram................	P C R C V	Y T L G D	Y T A E G	L G V P I
2d Test—"Plain text"....	M Z K Z G	D I Q V Y	D I B X V	Q V G M T

(5) Neither does the second test yield intelligible text. But let the results of the two tests be superimposed. Thus:

Cryptogram................	P C R C V	Y T L G D	Y T A E G	L G V P I
1st Test—"Plain text"...	L Y J Y F	C H P U X	C H A W U	P U F L S
2d Test—"Plain text"....	M Z K Z G	D I Q V Y	D I B X V	Q V G M T

[3] It is recommended that the student prepare a set of 25 strips ¼ by ½ by 15 inches, made of well-seasoned wood, and glue alphabet strips to the wood. The alphabet on each strip should be a double or repeated alphabet with all letters equally spaced.

(6) It is seen that the letters of the "plain text" given by the *second* trial are merely the continuants of the normal sequences initiated by the letters of the "plain text" given by the first trial. If these sequences are "run down"—that is, completed within the columns—the results must obviously be the same as though successive tests exactly similar to the first two were applied to the cryptogram, using one reversed normal and one direct normal component. If the cryptogram has really been prepared by means of a single reversed standard alphabet, one of the generatrices of the diagram that results from completing the sequences *must* yield intelligible text.

(7) Let the diagram be made, or better yet, if the student has already at hand the set of sliding strips referred to in the footnote to page 36, let him "set up" the letters given by the *first* trial. Fig. 12 shows the diagram and indicates the plain-text generatrix.

P	C	R	C	V	Y	T	L	G	D	Y	T	A	E	G	L	G	V	P	I
L	Y	J	Y	F	C	H	P	U	X	C	H	A	W	U	P	U	F	L	S
M	Z	K	Z	G	D	I	Q	V	Y	D	I	B	X	V	Q	V	G	M	T
N	A	L	A	H	E	J	R	W	Z	E	J	C	Y	W	R	W	H	N	U
O	B	M	B	I	F	K	S	X	A	F	K	D	Z	X	S	X	I	O	V
P	C	N	C	J	G	L	T	Y	B	G	L	E	A	Y	T	Y	J	P	W
Q	D	O	D	K	H	M	U	Z	C	H	M	F	B	Z	U	Z	K	Q	X
*R	E	P	E	L	I	N	V	A	D	I	N	G	C	A	V	A	L	R	Y
S	F	Q	F	M	J	O	W	B	E	J	O	H	D	B	W	B	M	S	Z
T	G	R	G	N	K	P	X	C	F	K	P	I	E	C	X	C	N	T	A
U	H	S	H	O	L	Q	Y	D	G	L	Q	J	F	D	Y	D	O	U	B
V	I	T	I	P	M	R	Z	E	H	M	R	K	G	E	Z	E	P	V	C
W	J	U	J	Q	N	S	A	F	I	N	S	L	H	F	A	F	Q	W	D
X	K	V	K	R	O	T	B	G	J	O	T	M	I	G	B	G	R	X	E
Y	L	W	L	S	P	U	C	H	K	P	U	N	J	H	C	H	S	Y	F
Z	M	X	M	T	Q	V	D	I	L	Q	V	O	K	I	D	I	T	Z	G
A	N	Y	N	U	R	W	E	J	M	R	W	P	L	J	E	J	U	A	H
B	O	Z	O	V	S	X	F	K	N	S	X	Q	M	K	F	K	V	B	I
C	P	A	P	W	T	Y	G	L	O	T	Y	R	N	L	G	L	W	C	J
D	Q	B	Q	X	U	Z	H	M	P	U	Z	S	O	M	H	M	X	D	K
E	R	C	R	Y	V	A	I	N	Q	V	A	T	P	N	I	N	Y	E	L
F	S	D	S	Z	W	B	J	O	R	W	B	U	Q	O	J	O	Z	F	M
G	T	E	T	A	X	C	K	P	S	X	C	V	R	P	K	P	A	G	N
H	U	F	U	B	Y	D	L	Q	T	Y	D	W	S	Q	L	Q	B	H	O
I	V	G	V	C	Z	E	M	R	U	Z	E	X	T	R	M	R	C	I	P
J	W	H	W	D	A	F	N	S	V	A	F	Y	U	S	N	S	D	J	Q
K	X	I	X	E	B	G	O	T	W	B	G	Z	V	T	O	T	E	K	R

FIGURE 12.

(8) The only difference in procedure between this case and the preceding one (where the cipher alphabet was a direct standard alphabet) is that the letters of the cipher text are first "deciphered" by means of *any* reversed standard alphabet and then the columns are "run down", according to the normal A B C . . . Z sequence. For reasons which will become apparent very soon, the first step in this method is technically termed *converting the cipher letters into their plain-component equivalents;* the second step is the same as before, *viz, completing the plain-component sequence.*

21. Special remarks on the method of solution by completing the plain-component sequence.—*a.* The terms employed to designate the steps in the solution set forth in Par. 20*b*, viz, "converting the cipher letters into their plain-component equivalents" and "completing the plain-component sequence", accurately describe the process. Their meaning will become more clear as the student progresses with the work. It may be said that whenever the plain component of a cipher alphabet is a *known* sequence, no matter how it is composed, the difficulty and time required to solve any cryptogram involving the use of that plain component is practically cut in half. *In some cases this knowledge facilitates, and in other cases is the only thing that makes possible the solution of a very short cryptogram that might otherwise defy solution.* Later on an example will be given to illustrate what is meant in this regard.

b. The student should take note, however, of two qualifying expressions that were employed in a preceding paragraph to describe the results of the application of the method. It was stated that "one of the generatrices will exhibit intelligible text *all the way across;* this text will *practically invariably* be the plain text." Will there ever be a case in which more than one generatrix will yield intelligible text throughout its extent? That obviously depends almost entirely on the number of letters that are aligned to form a generatrix. If a generatrix contains but a very few letters, only five, for example, it may happen as a result of pure chance that there will be two or more generatrices showing what might be "intelligible text." Note in Fig. 11, for example, that there are several cases in which 3-letter and 4-letter English words (ANY, VAIN, GOT, TIP, etc.) appear on generatrices that are not correct, these words being formed by pure chance. But there is not a single case, in this diagram, of a 5-letter or longer word appearing fortuitously, because obviously the longer the word the smaller the probability of its appearance purely by chance; and the probability that two generatrices of 15 letters each will both yield intelligible text along their entire length is exceedingly remote, so remote, in fact, that in practical cryptography such a case may be considered nonexistent.[4]

c. The student should observe that in reality there is no difference whatsoever in principle between the two methods presented in subpars. *a* and *b* of Par. 20. In the former the preliminary step of converting the cipher letters into their plain-component equivalents is apparently not present but in reality it is there. The reason for its apparent absence is that in that case the plain component of the cipher alphabet is identical in all respects with the cipher component, so that the cipher letters require no conversion, or, rather, they are identical with the equivalents that would result if they were converted on the basis $A_c = A_p$. In fact, if the solution process had been arbitrarily initiated by converting the cipher letters into their plain-component equivalents at the setting $A_c = O_p$, for example, and the cipher component slid one interval to the right thereafter, the results of the first and second tests of Par. 20*a* would be as follows:

Cryptogram	X K V K R O T B G J O T M I G B G R X E
1st Test—"Plain text"	L Y J Y F C H P U X C H A W U P U F L S
2nd Test—"Plain text"	M Z K Z G D I Q V Y D I B X V Q V G M T

Thus, the foregoing diagram duplicates in every particular the diagram resulting from the first two tests under Par. 20*b*: a first line of cipher letters, a second line of letters derived from them but showing externally no relationship with the first line, and a third line derived immediately from the second line by continuing the direct normal sequence. This point is brought to attention only for the purpose of showing that a single, broad principle is the basis of the general method of solution by completing the plain-component sequence, and once the student has this firmly in

[4] A person with patience and an inclination toward the curiosities of the science might construct a text of 15 or more letters which would yield two "intelligible" texts on the plain-component completion diagram.

mind he will have no difficulty whatsoever in realizing when the principle is applicable, what a powerful cryptanalytic tool it can be, and what results he may expect from its application in specific instances.

d. In the two foregoing examples of the application of the principle, the plain component was a normal sequence but it should be clear to the student, if he has grasped what has been said in the preceding subparagraph, that this component may be a mixed sequence which, if known (that is, if the sequence of letters comprising the sequence is known to the cryptanalyst), can be handled just as readily as can a plain component that is a normal sequence.

e. It is entirely immaterial at what points the plain and the cipher components are juxtaposed in the preliminary step of converting the cipher letters into their plain-component equivalents. For example, in the case of the reversed alphabet cipher solved in Par. 20*b*, the two components were arbitrarily juxtaposed to give the value A=A, but they might have been juxtaposed at any of the other 25 possible points of coincidence without in any way affecting the final result, *viz*, the production of one plain-text generatrix in the completion diagram.

22. Value of mechanical solution as a short cut.—*a.* It is obvious that the very first step the student should take in his attempts to solve an unknown cryptogram that is obviously a substitution cipher is to try the mechanical method of solution by completing the plain-component sequence, using the normal alphabet, first direct, then reversed. This takes only a very few minutes and is conclusive in its results. It saves the labor and trouble of constructing a frequency distribution in case the cipher is of this simple type. Later on it will be seen how certain variations of this simple type may also be solved by the application of this method. Thus, a very easy short cut to solution is afforded, which even the experienced cryptanalyst never overlooks in his first attack on an unknown cipher.

b. It is important now to note that *if neither of the two foregoing attempts is successful in bringing plain text to light and the cryptogram is quite obviously monoalphabetic in character, the cryptanalyst is warranted in assuming that the cryptogram involves a mixed cipher alphabet.*[5] The steps to be taken in attacking a cipher of the latter type will be discussed in the next section.

[5] There is but one other possibility, already referred to under Par. 17*d*, which involves the case where transposition and monoalphabetic substitution processes have been applied in successive steps. This is unusual, however, and will be discussed in its proper place.

SECTION VI

UNILITERAL SUBSTITUTION WITH MIXED CIPHER ALPHABETS

23. Basic reason for the low degree of cryptographic security afforded by monoalphabetic cryptograms involving standard cipher alphabets.—The student has seen that the solution of monoalphabetic cryptograms involving standard cipher alphabets is a very easy matter. Two methods of analysis were described, one involving the construction of a frequency distribution, the other not requiring this kind of tabulation, being almost mechanical in nature and correspondingly rapid. In the first of these two methods it was necessary to make a correct assumption as to the value of but one of the 26 letters of the cipher alphabet and the values of the remaining 25 letters at once become known; in the second method it was not necessary to assume a value for even a single cipher letter. The student should understand what constitutes the basis of this situation, *viz*, the fact that the two components of the cipher alphabet are composed of *known sequences*. What if one or both of these components are, for the cryptanalyst, *unknown sequences*? In other words, what difficulties will confront the cryptanalyst if the cipher component of the cipher alphabet is a mixed sequence? Will such an alphabet be solvable as a whole at one stroke, or will it be necessary to solve its values individually? Since the determination of the value of one cipher letter in this case gives no direct clues to the value of any other letter, it would seem that the solution of such a cipher should involve considerably more analysis and experiment than has the solution of either of the two types of ciphers so far examined occasioned. A typical example will be studied.

24. Preliminary steps in the analysis of a monoalphabetic, mixed alphabet cryptogram.—*a.* Note the following cryptogram:

```
SFDZF IOGHL PZFGZ DYSPF HBZDS GVHTF UPLVD FGYVJ VFVHT GADZZ AITYD
ZYFZJ ZTGPT VTZBD VFHTZ DFXSB GIDZY VTXOI YVTEF VMGZZ THLLV XZDFM
HTZAI TYDZY BDVFH TZDFK ZDZZJ SXISG ZYGAV FSLGZ DTHHT CDZRS VTYZD
OZFFH TZAIT YDZYG AVDGZ ZTKHI TYZYS DZGHU ZFZTG UPGDI XWGHX ASRUZ
DFUID EGHTV EAGXX
```

b. A casual inspection of the text discloses the presence of several long repetitions as well as of many letters of normally low frequency, such as F, G, V, X, and Z; on the other hand, letters of

normally high frequency, such as the vowels, and the consonants N and R, are relatively scarce. The cryptogram is obviously a substitution cipher and the usual mechanical tests for determining whether it is possibly of the monoalphabetic, standard-alphabet type are applied. The results being negative, a uniliteral frequency distribution is immediately constructed and is as shown in Figure 13.

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
8	4	1	23	3	19	19	15	10	3	2	5	2	0	3	5	0	2	10	22	5	16	1	8	14	35

FIGURE 13

c. The fact that the frequency distribution shows very marked crests and troughs means that the cryptogram is undoubtedly monoalphabetic; the fact that it has already been tested (by the method of completing the plain-component sequence) and found not to be of the monoalphabetic, standard-alphabet type, indicates with a high degree of probability that it involves a mixed cipher alphabet. A few moments might be devoted to making a careful inspection of the distribution to insure that it cannot be made to fit the normal; the object of this would be to rule out the possibility that the text resulting from substitution by a standard cipher alphabet had not subsequently been transposed. But this inspection in this case is hardly necessary, in view of the presence of long repetitions in the message.[1] (See Par. 13*g*.)

d. One might, of course, attempt to solve the cryptogram by applying the simple principles of frequency. One might, in other words, assume that Z_c (the letter of greatest frequency) represents E_p, D_c (the letter of next greatest frequency) represents T_p, and so on. If the message were long enough this simple procedure might more or less quickly give the solution. But the message is relatively short and many difficulties would be encountered. Much time and effort would be expended unnecessarily, because it is hardly to be expected that in a message of only 235 letters the relative order of frequency of the various cipher letters should exactly coincide with, or even closely approximate the relative order of frequency of letters of normal plain text found in a count of 50,000 letters. *It is to be emphasized that the beginner must repress the natural tendency to place too much confidence in the generalized principles of frequency and to rely too much upon them.* It is far better to bring into effective use certain other data concerning normal plain text which thus far have not been brought to notice.

25. Further data concerning normal plain text.—*a.* Just as the individual letters constituting a large volume of plain text have more or less characteristic or fixed frequencies, so it is found that digraphs and trigraphs have characteristic frequencies, when a large volume of text is studied statistically. In Appendix 1, Table 6, are shown the relative frequencies of all digraphs appearing in the 260 telegrams referred to in Paragraph 9*e*. It will be noted that 428 of the 676 possible pairs of letters occur in these telegrams, but whereas many of them occur but once or twice, there are a few which occur hundreds of times.

b. In Appendix 1 will also be found several other kinds of tables and lists which will be useful to the student in his work, such as the relative order of frequency of the 50 digraphs of greatest

[1] This possible step is mentioned here for the purpose of making it clear that the plain-component sequence completion method cannot solve a case in which transposition has followed or preceded monoalphabetic substitution with standard alphabets. Cases of this kind will be discussed in a later text. It is sufficient to indicate at this point that the frequency distribution for such a combined substitution-transposition cipher would present the characteristics of a standard alphabet cipher—and yet the method of completing the plain-component sequence would fail to bring out any plain text.

frequency, the relative order of frequency of doubled letters, doubled vowels, doubled consonants, and so on. It is suggested that the student refer to this appendix now, to gain an idea of the data available for his future reference. Just how these data may be employed will become apparent very shortly.

26. Preparation of the work sheet.—*a.* The details to be considered in this paragraph may at first appear to be superfluous but long experience has proved that systematization of the work, and preparation of the data in the most utilizable, condensed form is most advisable, even if this seems to take considerable time. In the first place if it merely serves to avoid interruptions and irritations occasioned by failure to have the data in an instantly available form, it will pay by saving mental wear and tear. In the second place, especially in the case of complicated cryptograms, painstaking care in these details, while it may not always bring about success, is often the factor that is of greatest assistance in ultimate solution. The detailed preparation of the data may be irksome to the student, and he may be tempted to avoid as much of it as possible, but, unfortunately, in the early stages of solving a cryptogram he does not know (nor, for that matter, does the expert always know) just which data are essential and which may be neglected. Even though not all of the data may turn out to have been necessary, as a general rule, time is saved in the end if all the usual data are prepared as a regular preliminary to the solution of most cryptograms.

b. First, the cryptogram is recopied in the form of a *work sheet.* This sheet should be of a good quality of paper so as to withstand considerable erasure. If the cryptogram is to be copied by hand, cross-section paper of ¼-inch squares is extremely useful. The writing should be in ink, and plain, carefully made roman capital letters should be used in all cases. If the cryptogram is to be copied on a typewriter, the ribbon employed should be impregnated with an ink that will not smear or smudge under the hand.

c. The arrangement of the characters of the cryptogram on the work sheet is a matter of considerable importance. If the cryptogram as first obtained is in groups of regular length (usually five characters to a group) and if the uniliteral frequency distribution shows the cryptogram to be monoalphabetic, the characters should be copied without regard to this grouping. It is advisable to allow two spaces between letters, and to write a constant number of letters per line, approximately 25. At least two spaces, preferably three spaces, should be left between horizontal lines. Care should be taken to avoid crowding the letters in any case, for this is not only confusing to the eye but also mentally irritating when later it is found that not enough space has been left for making various sorts of marks or indications. If the cryptogram is originally in what appears to be word lengths (and this is the case, as a rule, only with the cryptograms of amateurs), naturally it should be copied on the work sheet in the original groupings. If further study of a cryptogram shows that some special grouping is required, it is often best to recopy it on a fresh work sheet rather than to attempt to indicate the new grouping on the old work sheet.

d. In order to be able to locate or refer to specific letters or groups of letters with speed, certainty, and without possibility of confusion, it is advisable to use coordinates applied to the lines and columns of the text as it appears on the work sheet. To minimize possibility of confusion, it is best to apply letters to the horizontal lines of the text, numbers to the vertical columns. In referring to a letter the horizontal line in which the letter is located is usually given first. Thus, referring to the work sheet shown below, coordinates A17 designate the letter Y, the 17th letter in the first line. The letter I is usually omitted from the series of line indicators so as to avoid confusion with the figure 1. If lines are limited to 25 letters each, then each set of 100 letters of the text is automatically blocked off by remembering that 4 lines constitute 100 letters.

e. Above each character of the cipher text may be some indication of the frequency of that character in the whole cryptogram. This indication may be the actual number of times the

character occurs, or, if colored pencils are used, the cipher letters may be divided up into three categories or groups—high frequency, medium frequency, and low frequency. It is perhaps simpler, if clerical help is available, to indicate the actual frequencies. This saves constant reference to the frequency tables, which interrupts the train of thought, and saves considerable time in the end.

f. After the special frequency distribution, explained in Par. 27 below, has been constructed, repetitions of digraphs and trigraphs should be underscored. In so doing, the student should be particularly watchful of trigraphic repetitions which can be further extended into tetragraphs and polygraphs of greater length. Repetitions of more than ten characters should be set off by heavy vertical lines, as they indicate repeated phrases and are of considerable assistance in solution. If a repetition continues from one line to the next, put an arrow at the end of the underscore to signal this fact. Reversible digraphs should also be indicated by an underscore with an arrow pointing in both directions. Anything which strikes the eye as being peculiar, unusual, or significant as regards the distribution or recurrence of the characters should be noted. All these marks should, if convenient, be made with ink so as not to cause smudging. The work sheet will now appear as shown herewith (not all the repetitions are underscored):

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
	10	19	23	35	19	10	3	19	15	5	5	35	19	19	35	23	14	10	5	19	15	4	35	23	10
A	S	F	D	Z	F	I	O	G	H	L	P	Z	F	G	Z	D	Y	S	P	F	H	B	Z	D	S
	19	16	15	22	19	5	5	5	16	23	19	19	14	16	3	16	19	16	15	22	19	8	23	35	35
B	G	V	H	T	F	U	P	L	V	D	F	G	Y	V	J	V	F	V	H	T	G	A	D	Z	Z
	8	10	22	14	23	35	14	19	35	3	35	22	19	5	22	16	22	35	4	23	16	19	15	22	35
C	A	I	T	Y	D	Z	Y	F	Z	J	Z	T	G	P	T	V	T	Z	B	D	V	F	H	T	Z
	23	19	8	10	4	19	10	23	35	14	16	22	8	3	10	14	16	22	3	19	16	2	19	35	35
D	D	F	X	S	B	G	I	D	Z	Y	V	T	X	O	I	Y	V	T	E	F	V	M	G	Z	Z
	22	15	5	5	16	8	35	23	19	2	15	22	35	8	10	22	14	23	35	14	4	23	16	19	15
E	T	H	L	L	V	X	Z	D	F	M	H	T	Z	A	I	T	Y	D	Z	Y	B	D	V	F	H
	22	35	23	19	2	35	23	35	35	3	10	8	10	10	19	35	14	19	8	16	19	10	5	19	35
F	T	Z	D	F	K	Z	D	Z	Z	J	S	X	I	S	G	Z	Y	G	A	V	F	S	L	G	Z
	23	22	15	15	22	1	23	35	2	10	16	22	14	35	23	3	35	19	19	15	22	35	8	10	22
G	D	T	H	H	T	C	D	Z	R	S	V	T	Y	Z	D	O	Z	F	F	H	T	Z	A	I	T
	14	23	35	14	19	8	16	23	19	35	35	22	2	15	10	22	14	35	14	10	23	35	19	15	5
H	Y	D	Z	Y	G	A	V	D	G	Z	Z	T	K	H	I	T	Y	Z	Y	S	D	Z	G	H	U
	35	19	35	22	19	5	5	19	23	10	8	1	19	15	8	8	10	2	5	35	23	19	5	10	23
J	Z	F	Z	T	G	U	P	G	D	I	X	W	G	H	X	A	S	R	U	Z	D	F	U	I	D
	3	19	15	22	16	3	8	19	8	8															
K	E	G	H	T	V	E	A	G	X	X															

27. Triliteral-frequency distributions.—*a.* In what has gone before, a type of frequency distribution known as a uniliteral frequency distribution was used. This, of course, shows only the number of times each individual letter occurs. In order to apply the normal digraphic and trigraphic frequency data (given in Appendix 1) to the solution of a cryptogram of the type now being studied, it is obvious that the data with respect to digraphs and trigraphs occurring in the cryptogram should be compiled and should be compared with the data for normal plain text. In order to accomplish this in suitable manner, it is advisable to construct a slightly more complicated form of distribution termed a *triliteral frequency distribution*.[3]

b. Given a cryptogram of 50 or more letters and the task of determining what trigraphs are present in the cryptogram, there are three ways in which the data may be arranged or assembled. One may require that the data show (1) each letter with its two succeeding letters; (2) each letter with its two preceding letters; (3) each letter with one preceding letter and one succeeding letter.

c. A distribution of the first of the three foregoing types may be designated as a "triliteral frequency distribution showing two suffixes"; the second type may be designated as a "triliteral frequency distribution showing two prefixes"; the third type may be designated as a "triliteral frequency distribution showing one prefix and one suffix." Quadriliteral and pentaliteral frequency distributions may occasionally be found useful.

d. Which of these three arrangements is to be employed at a specific time depends largely upon what the data are intended to show. For present purposes, in connection with the solution of a monoalphabetic substitution cipher employing a mixed alphabet, possibly the third arrangement, that showing one prefix and one suffix, is most satisfactory.

e. It is convenient to use ¼-inch cross-section paper for the construction of a triliteral frequency distribution in the form of a distribution showing crests and troughs, such as that in Figure 14. In that figure the prefix to each letter to be recorded is inserted in the left half of the cell directly above the cipher letter being recorded; the suffix to each letter is inserted in the right half of the cell directly above the letter being recorded; and in each case the prefix and the suffix to the letter being recorded occupy the same cell, the prefix being directly to the left of the suffix. The number in parentheses gives the total frequency for each letter.

[3] Heretofore such a distribution has been termed a "trigraphic frequency table." It is thought that the word "triliteral" is more suitable, to correspond with the designation "uniliteral" in the case of the distribution of the single letters. A trigraphic distribution of A B C D E F would consider only the trigraphs A B C and D E F, whereas a triliteral distribution would consider the groups A B C, B C D, C D E, and D E F. (See also Par. 11*d*.) The use of the word "distribution" to replace the word "table" has already been explained.

CONDENSED TABLE OF REPETITIONS

Digraphs			*Trigraphs*		*Longer Polygraphs*
DZ–9	TZ–5	VF–4	DZY–4	FHT–3	HTZAITYDZY–2
ZD–9	TY–5	VT–4	HTZ–4	TYD–3	BDVFHTZDF–2
HT–8	FH–4	ZF–4	ITY–4	YDZ–3	ZAITYDZY–3
ZY–6	GH–4	ZT–4	ZDF–4	ZAI–3	FHTZ–3
DF–5	IT–4	ZZ–4	AIT–3		
GZ–5					

A	EG XS GV ZI GV ZI ZI GD	(8)
B	YD SG ZD HZ	(4)
C	TD	(1)
D	IE ZF GI SZ VG YZ ZO CZ ZT ZZ ZF BV YZ ZF IZ ZF BV YZ AZ VF ZS ZY FZ	(23)
E	VA DG TF	(3)
F	DU ZZ FH ZF VS DK VH DM EV DX VH YZ VV DG TU PH ZG ZI SD	(19)
G	AX EH WH PD TU ZH DZ YA LZ YA SZ MZ BI TP TA FY SV FZ OH	(19)
H	GT GX GU KI FT HT FH TT MT TL FT VT VT FB GL	(15)
I	UD DX HT AT XS AT OY GD AT FO	(10)
J	ZS ZZ VV	(3)
K	TH FZ	(2)
L	SG LV HL PV HP	(5)
M	FH VG	(2)
N		(0)
O	DZ XI IG	(3)
P	UG GT UL SF LZ	(5)
Q		(0)
R	SU ZS	(2)
S	AR YD RV FL IG JX XB DG YP –F	(10)
T	HV ZG IY ZK IY HZ VY HC DH HZ IY HZ ZH VE VX HZ VZ PV ZG IY HG HF	(22)
U	FI RZ GP HZ FP	(5)
V	TE AD ST AF DF LX FM YT YT DF TT FH JF YJ LD GH	(16)
W	XG	(1)
X	X– GX HA IW SI VZ TO FS	(8)
Y	ZS TZ ZG TD TZ ZG ZB TD IV ZV ZF TD GV DS	(14)
Z	UD FT UF DG YY ZT GZ DY TA OF YD DR GD GY ZJ DZ KD TD DY TA XD ZT GZ DY TD TB JT FJ DY ZA DZ BD GD PF DF	(35)

FIGURE 14.

148274—38——4

f. The triliteral frequency distribution is now to be examined with a view to ascertaining what digraphs and trigraphs occur two or more times in the cryptogram. Consider the pair of columns containing the prefixes and suffixes to D_c in the distribution, as shown in Fig. 14. This pair of columns shows that the following digraphs appear in the cryptogram:

Digraphs based on prefixes (arranged as one reads up the column)	*Digraphs based on suffixes (arranged as one reads up the column)*
FD, ZD, ZD, VD, AD, YD, BD,	DZ, DY, DS, DF, DZ, DZ, DV,
ZD, ID, ZD, YD, BD, ZD, ZD,	DF, DZ, DF, DZ, DV, DF, DZ,
ZD, CD, ZD, YD, VD, SD, GD,	DT, DZ, DO, DZ, DG, DZ, DI,
ZD, ID	DF, DE

The nature of the triliteral frequency distribution is such that in finding what digraphs are present in the cryptogram it is immaterial whether the prefixes or the suffixes to the cipher letters are studied, *so long as one is consistent in the study*. For example, in the foregoing list of digraphs based on the prefixes to D_c, the digraphs FD, ZD, ZD, VD, etc., are found; if now, the student will refer to the suffixes of F_c, Z_c, V_c, etc., he will find the very same digraphs indicated. This being the case, the question may be raised as to what value there is in listing both the prefixes and the suffixes to the cipher letters. The answer is that by so doing the trigraphs are indicated at the same time. For example, in the case of D_c, the following trigraphs are indicated:

FDZ, ZDY, ZDS, VDF, ADZ, YDZ, BDV, ZDF, IDZ, ZDF, YDZ, BDV, ZDF,
ZDZ, ZDT, CDZ, ZDO, YDZ, VDG, SDZ, GDI, ZDF, IDE.

g. The *repeated* digraphs and trigraphs can now be found quite readily. Thus, in the case of D_c, examining the list of digraphs based on suffixes, the following repetitions are noted:

DZ appears 9 times
DF appears 5 times
DV appears 2 times

Examining the trigraphs with D_c as central letter, the following repetitions are noted:

ZDF appears 4 times
YDZ appears 3 times
BDV appears 2 times

h. It is unnecessary, of course, to go through the detailed procedure set forth in the preceding subparagraphs in order to find all the repeated digraphs and trigraphs. The repeated trigraphs with D_c as central letter can be found merely from an inspection of the prefixes and suffixes opposite D_c in the distribution. It is necessary only to find those cases in which two or more prefixes are identical at the same time that the suffixes are identical. For example, the distribution shows at once that in four cases the prefix to D_c is Z_c at the same time that the suffix to this letter is F_c. Hence, the trigraph ZDF appears four times. The repeated trigraphs may all be found in this manner.

i. The most frequently repeated digraphs and trigraphs are then assembled in what is termed a *condensed table of repetitions*, so as to bring this information prominently before the eye. As a rule, digraphs which occur less than four or five times, and trigraphs which occur less than three or four times may be omitted from the condensed table as being relatively of no importance in the study of repetitions. In the condensed table the frequencies of the individual letters forming the most important digraphs, trigraphs, etc., should be indicated.

28. **Classifying the cipher letters into vowels and consonants.**—*a.* Before proceeding to a detailed analysis of the repeated digraphs and trigraphs, a very important step can be taken which will be of assistance not only in the analysis of the repetitions but also in the final solution of the cryptogram. This step concerns the classification of the high-frequency letters into two

groups—vowels and consonants. For if the cryptanalyst can quickly ascertain the equivalents of the four vowels, A, E, I, and O, and of only the four consonants, N, R, S, and T, he will then have the values of approximately two-thirds of all the cipher letters that occur in the cryptogram; the values of the remaining letters can almost be filled in automatically.

b. The basis for the classification will be found to rest upon a comparatively simple phenomenon: the associational or combinatory behavior of vowels is, in general, quite different from that of consonants. If an examination be made of Table 7–B in Appendix 1, showing the relative order of frequency of the 18 digraphs composing 25 percent of English telegraphic text, it will be seen that the letter E enters into the composition of 9 of the 18 digraphs; that is, in exactly half of all the cases the letter E is one of the two letters forming the digraph. The digraphs containing E are as follows:

ED	EN	ER	ES		
	NE	RE	SE	TE	VE

The remaining nine digraphs are as follows:

AN	ND	OR	ST
IN	NT		TH
ON			TO

c. None of the 18 digraphs is a combination of vowels. Note now that of the 9 combinations with E, 7 are with the consonants N, R, S, and T, one is with D, one is with V, and *none is with any vowel.* In other words, E_p combines most readily with consonants but not with other vowels, or even with itself. Using the terms often employed in the chemical analogy, E shows a great "affinity" for the consonants N, R, S, T, but not for the vowels. Therefore, if the letters of highest frequency occurring in a given cryptogram are listed, together with the number of times each of them combines with the cipher equivalent of E_p, those which show considerable combining power or affinity for the cipher equivalent of E_p may be assumed to be the cipher equivalents of N, R, S, T_p; those which do not show any affinity for the cipher equivalent of E_p may be assumed to be the cipher equivalents of A, I, O, U_p. Applying these principles to the problem in hand, and examining the triliteral frequency distribution, it is quite certain that $Z_c = E_p$, not only because Z_c is the letter of highest frequency, but also because it combines with *several* other high-frequency letters, such as D_c, F_c, G_c, etc. The nine letters of next highest frequency are:

23	22	19	19	16	15	14	10	10
D	T	F	G	V	H	Y	S	I

Let the combinations these letters form with Z_c be indicated in the following manner:

Number of times Z_c occurs as prefix	~~IIII~~ IIII	IIII	IIII	I			~~IIII~~ I		
Cipher Letter	D(23)	T(22)	F(19)	G(19)	V(16)	H(15)	Y(14)	S(10)	I(10)
Number of times Z_c occurs as suffix	~~IIII~~ IIII	~~IIII~~	II	~~IIII~~			II		

d. Consider D_c. It occurs 23 times in the message and 18 of those times it is combined with Z_c, 9 times in the form Z_cD_c ($=E\theta_p$), and 9 times in the form D_cZ_c ($=\theta E_p$). It is clear that D_c must be a consonant. In the same way, consider T_c, which shows 9 combinations with Z_c, 4 in the form Z_cT_c ($=E\theta_p$) and 5 in the form T_cZ_c ($=\theta E_p$). The letter T_c appears to represent a consonant, as do also the letters F_c, G_c, and Y_c. On the other hand, consider V_c, occurring in all 16 times but never in combination with Z_c; it appears to represent a vowel, as do also the letters H_c, S_c, and I_c. So far, then, the following classification would seem logical:

Vowels	*Consonants*
$Z_c(=E_p)$, V_c, H_c, S_c, I_c	D_c, T_c, F_c, G_c, Y_c

29. Further analysis of the letters representing vowels and consonants.—*a.* O_p is usually the vowel of second highest frequency. Is it possible to determine which of the letters V, H, S, I_c is the cipher equivalent of O_p? Let reference be made again to Table 6 in Appendix 1, where it is seen that the 10 most frequently occurring diphthongs are:

Diphthong	IO	OU	EA	EI	AI	IE	AU	EO	AY	UE
Frequency	41	37	35	27	17	13	13	12	12	11

If V, H, S, I_c are really the cipher equivalents of A, I, O, U_p (not respectively), perhaps it is possible to determine which is which *by examining the combinations they make among themselves and with* Z_c ($=E_p$). Let the combinations of V, H, S, I, and Z that occur in the message be listed. There are only the following:

ZZ_c—4	HI—1
VH —2	SV—1
HH —1	IS—1

ZZ_c is of course EE_p. Note the doublet HH_c; if H_c is a vowel, then the chances are excellent that $H_c=O_p$ because the doublets AA_p, II_p, UU_p, are practically non-existent, whereas the double vowel combination OO_p is of next highest frequency to the double vowel combination EE_p. If $H_c=O_p$, then V_c must be I_p because the digraph VH_c occurring two times in the message could hardly be AO_p, or UO_p, whereas the diphthong IO_p is the one of high frequency in English. So far then, the tentative (because so far unverified) results of the analysis are as follows:

$$Z_c=E_p \quad H_c=O_p \quad V_c=I_p$$

This leaves only two letters, I_c and S_c (already classified as vowels) to be separated into A_p and U_p. Note the digraphs:

$$HI_c=O\theta_p$$
$$SV_c=\theta I_p$$
$$IS_c=\theta\theta_p$$

Only two alternatives are open:

(1) Either $I_c=A_p$ and $S_c=U_p$,
(2) Or $I_c=U_p$ and $S_c=A_p$.

If the first alternative is selected, then

$$HI_c=OA_p$$
$$SV_c=UI_p$$
$$IS_c=AU_p$$

If the second alternative is selected, then

$$HI_c=OU_p$$
$$SV_c=AI_p$$
$$IS_c=UA_p$$

The eye finds it difficult to choose between these alternatives; but suppose the frequency values of the plain-text diphthongs as given in Table 6 of Appendix 1 are added for each of these alternatives, giving the following:

$HI_c=OA_p$, frequency value= 7	$HI_c=OU_p$, frequency value=37
$SV_c=UI_p$, frequency value= 5	$SV_c=AI_p$, frequency value=17
$IS_c=AU_p$, frequency value=13	$IS_c=UA_p$, frequency value= 5
Total 25	Total 59

Mathematically, the second alternative is more than twice as probable as the first. Let it be assumed to be correct and the following (still tentative) values are now at hand:

$$Z_c=E_p \quad H_c=O_p \quad V_c=I_p \quad S_c=A_p \quad I_c=U_p$$

b. Attention is now directed to the letters classified as consonants. How far is it possible to ascertain their values? The letter D_c, from considerations of frequency alone, would seem to be T_p, but its frequency, 23, is not considerably greater than that for T_c. It is not much greater than that for F_c or G_c, with a frequency of 19 each. But perhaps it is possible to ascertain not the value of one letter alone but of two letters at one stroke. To do this one may make use of a tetragraph of considerable importance in English, *viz*, $TION_p$. For if the analysis pertaining to the vowels is correct, and if $VH_c=IO_p$, then an examination of the letters immediately before and after the digraph VH_c in the cipher text might disclose both T_p and N_p. Reference to the text gives the following:

$GVHT_c$	$FVHT_c$
$\theta IO\theta_p$	$\theta IO\theta_p$

The letter T_c follows VH_c in both cases and very probably indicates that $T_c=N_p$; but as to whether G_c or F_c equals T_p cannot be decided. However, two conclusions are clear: first, the letter D_c is neither T_p nor N_p, from which it follows that it must be either R_p or S_p; second, the letters G_c and F_c must be either T_p and S_p, respectively, or S_p and T_p, respectively, because the only tetragraphs usually found (in English) containing the diphthong IO_p as central letters are $SION_p$ and $TION_p$. This in turn means that as regards D_c, the latter cannot be *either* R_p or S_p; it *must* be R_p, a conclusion which is corroborated by the fact that ZD_c ($=ER_p$) and DZ_c ($=RE_p$) occur 9 times each. Thus far, then, the identifications, when inserted in an *enciphering* alphabet, are as follows:

Plain	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
Cipher	S				Z				V					T	H			D	G	F	I					
																			F	G						

30. Substituting deduced values in the cryptogram.—*a.* Thus far the analysis has been almost purely hypothetical, for as yet not a single one of the values deduced from the foregoing analysis has been tried out in the cryptogram. It is high time that this be done, because the final test of the validity of the hypotheses, assumptions, and identifications made in any cryptographic study is, after all, only this: do these hypotheses, assumptions, and identifications ultimately yield verifiable, intelligible plain-text when *consistently* applied to the cipher text?

b. At the present stage in the process, since there are at hand the assumed values of but 9 out of the 25 letters that appear, it is obvious that a continuous "reading" of the cryptogram can certainly not be expected from a mere insertion of the values of the 9 letters. However, the substitution of these values should do two things. First, it should immediately disclose the fragments, outlines, or "skeletons" of "good" words in the text; and second, it should disclose no places in the text where "impossible" sequences of letters are established. By the first is meant that the partially deciphered text should show the outlines or skeletons of words such as may be expected to be found in the communication; this will become quite clear in the next subparagraph. By the second is meant that sequences, such as "AOOEN" or "TNRSENO" or the like, obviously not possible or extremely unusual in normal English text, must not result from the substitution of the tentative identifications resulting from the analysis. The appearance of several such extremely unusual or impossible sequences at once signifies that one or more of the assumed values is incorrect.

c. Here are the results of substituting the nine values which have been deduced by the reasoning based on a classification of the high-frequency letters into vowels and consonants and the study of the members of the two groups:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
	10	19	23	35	19	10	3	19	15	5	5	35	19	19	35	23	14	10	5	19	15	4	35	23	10
A	S	F	D	Z	F	I	O	G	H	L	P	Z	F	G	Z	D	Y	S	P	F	H	B	Z	D	S
	A	T	R	E	T			S	O			E	T	S	E	R		A		T	O		E	R	A
		S			S			T					S	T						S					
	19	16	15	22	19	5	5	5	16	23	19	19	14	16	3	16	19	16	15	22	19	8	23	35	35
B	G	V	H	T	F	U	P	L	V	D	F	G	Y	V	J	V	F	V	H	T	G	A	D	Z	Z
	S	I	O	N	T				I	R	T	S		I		I	T	I	O	N	S		R	E	E
	T				S						S	T					S				T				
	8	10	22	14	23	35	14	19	35	3	35	22	19	5	22	16	22	35	4	23	16	19	15	22	35
C	A	I	T	Y	D	Z	Y	F	Z	J	Z	T	G	P	T	V	T	Z	B	D	V	F	H	T	Z
			N		R	E		T	E		E	N	S		N	I	N	E		R	I	T	O	N	E
								S					T									S			
	23	19	8	10	4	19	10	23	35	14	16	22	8	3	10	14	16	22	3	19	16	2	19	35	35
D	D	F	X	S	B	G	I	D	Z	Y	V	T	X	O	I	Y	V	T	E	F	V	M	G	Z	Z
	R	T		A		S		R	E		I	N					I	N		T	I		S	E	E
		S				T														S			T		
	22	15	5	5	16	8	35	23	19	2	15	22	35	8	10	22	14	23	35	14	4	23	16	19	15
E	T	H	L	L	V	X	Z	D	F	M	H	T	Z	A	I	T	Y	D	Z	Y	B	D	V	F	H
	N	O			I		E	R	T		O	N	E			N		R	E			R	I	T	O
									S															S	
	22	35	23	19	2	35	23	35	35	3	10	8	10	10	19	35	14	19	8	16	19	10	5	19	35
F	T	Z	D	F	K	Z	D	Z	Z	J	S	X	I	S	G	Z	Y	G	A	V	F	S	L	G	Z
	N	E	R	T		E	R	E	E		A			A	S	E		S		I	T	A		S	E
				S											T			T			S			T	
	23	22	15	15	22	1	23	35	2	10	16	22	14	35	23	3	35	19	19	15	22	35	8	10	22
G	D	T	H	H	T	C	D	Z	R	S	V	T	Y	Z	D	O	Z	F	F	H	T	Z	A	I	T
	R	N	O	O	N		R	E		A	I	N		E	R		E	T	T	O	N	E			N
																		S	S						
	14	23	35	14	19	8	16	23	19	35	35	22	2	15	10	22	14	35	14	10	23	35	19	15	5
H	Y	D	Z	Y	G	A	V	D	G	Z	Z	T	K	H	I	T	Y	Z	Y	S	D	Z	G	H	U
		R	E		S		I	R	S	E	E	N		O		N		E		A	R	E	S	O	
					T				T														T		
	35	19	35	22	19	5	5	19	23	10	8	1	24	15	8	8	10	2	5	35	23	19	5	10	23
J	Z	F	Z	T	G	U	P	G	D	I	X	W	G	H	X	A	S	R	U	Z	D	F	U	I	D
	E	T	E	N	S			S	R				S	O			A			E	R	T			R
		S			T			T					T									S			
	3	19	15	22	16	3	8	19	8	8															
K	E	G	H	T	V	E	A	G	X	X															
		S	O	N	I			S																	
		T						T																	

d. No impossible sequences are brought to light, and, moreover, several long words, nearly complete, stand out in the text. Note the following portions:

	A21									
	H	B	Z	D	S	G	V	H	T	F
(1)	O	?	E	R	A	S	I	O	N	T
						T				S

	C15												
	T	V	T	Z	B	D	V	F	H	T	Z	D	F
(2)	N	I	N	E	?	R	I	T	O	N	E	R	T
								S					S

	F22								
	S	L	G	Z	D	T	H	H	T
(3)	A	?	S	E	R	N	O	O	N
			T						

The words are obviously OPERATIONS, NINE PRISONERS, and AFTERNOON. The value G_c si clearly T_p; that of F_c is S_p; and the following additional values are certain:

$$B_c=P_p \qquad L_c=F_p$$

31. Completing the solution.—*a.* Each time an additional value is obtained, substitution is at once made throughout the cryptogram. This leads to the determination of further values, in an ever-widening circle, until all the identifications are firmly and finally established, and the message is completely solved. In this case the decipherment is as follows:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
A	S	F	D	Z	F	I	O	G	H	L	P	Z	F	G	Z	D	Y	S	P	F	H	B	Z	D	S
	A	S	R	E	S	U	L	T	O	F	Y	E	S	T	E	R	D	A	Y	S	O	P	E	R	A
B	G	V	H	T	F	U	P	L	V	D	F	G	Y	V	J	V	F	V	H	T	G	A	D	Z	Z
	T	I	O	N	S	B	Y	F	I	R	S	T	D	I	V	I	S	I	O	N	T	H	R	E	E
C	A	I	T	Y	D	Z	Y	F	Z	J	Z	T	G	P	T	V	T	Z	B	D	V	F	H	T	Z
	H	U	N	D	R	E	D	S	E	V	E	N	T	Y	N	I	N	E	P	R	I	S	O	N	E
D	D	F	X	S	B	G	I	D	Z	Y	V	T	X	O	I	Y	V	T	E	F	V	M	G	Z	Z
	R	S	C	A	P	T	U	R	E	D	I	N	C	L	U	D	I	N	G	S	I	X	T	E	E
E	T	H	L	L	V	X	Z	D	F	M	H	T	Z	A	I	T	Y	D	Z	Y	B	D	V	F	H
	N	O	F	F	I	C	E	R	S	X	O	N	E	H	U	N	D	R	E	D	P	R	I	S	O
F	T	Z	D	F	K	Z	D	Z	Z	J	S	X	I	S	G	Z	Y	G	A	V	F	S	L	G	Z
	N	E	R	S	W	E	R	E	E	V	A	C	U	A	T	E	D	T	H	I	S	A	F	T	E
G	D	T	H	H	T	C	D	Z	R	S	V	T	Y	Z	D	O	Z	F	F	H	T	Z	A	I	T
	R	N	O	O	N	Q	R	E	M	A	I	N	D	E	R	L	E	S	S	O	N	E	H	U	N
H	Y	D	Z	Y	G	A	V	D	G	Z	Z	T	K	H	I	T	Y	Z	Y	S	D	Z	G	H	U
	D	R	E	D	T	H	I	R	T	E	E	N	W	O	U	N	D	E	D	A	R	E	T	O	B
J	Z	F	Z	T	G	U	P	G	D	I	X	W	G	H	X	A	S	R	U	Z	D	F	U	I	D
	E	S	E	N	T	B	Y	T	R	U	C	K	T	O	C	H	A	M	B	E	R	S	B	U	R
K	E	G	H	T	V	E	A	G	X	X															
	G	T	O	N	I	G	H	T	X	X															

Message: AS RESULT OF YESTERDAYS OPERATIONS BY FIRST DIVISION THREE HUNDRED SEVENTY NINE PRISONERS CAPTURED INCLUDING SIXTEEN OFFICERS ONE HUNDRED PRISONERS WERE EVACUATED THIS AFTERNOON REMAINDER LESS ONE HUNDRED THIRTEEN WOUNDED ARE TO BE SENT BY TRUCK TO CHAMBERSBURG TONIGHT

b. The solution should, as a rule, not be considered complete until an attempt has been made to discover all the elements underlying the general system and the specific key to a message. In this case, there is no need to delve further into the general system, for it is merely one of monoalphabetic substitution with a mixed cipher alphabet. It is necessary or advisable, however, to reconstruct the cipher alphabet because this may give clues that later may become valuable.

c. Cipher alphabets should, as a rule, be reconstructed by the cryptanalyst in the form of *enciphering* alphabets because they will then usually be in the form in which the encipherer used them. This is important for two reasons. First, if the sequence in the cipher component gives evidence of system in its construction or if it yields clues pointing toward its derivation from a keyword or a key-phrase, this may often corroborate the identifications already made and may lead directly to additional identifications. A word or two of explanation is advisable here. For example, refer to the skeletonized enciphering alphabet given at the end of par. 29*b:*

Plain	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
Cipher	S				Z				V					T	H			D	G	F	I					
																			F	G						

Suppose the cryptanalyst, looking at the sequence DGFI or DFGI in the cipher component, suspects the presence of a keyword-mixed alphabet. Then DFGI is certainly a more plausible sequence than DGFI. Again, noting the sequence S . . . Z . . . V . . . . TH . . D, he might have an idea that the keyword begins after the Z and that the TH is followed by AB or BC. This would mean that either P, Q_p=A, B_c or B, C_c. Assuming that P, Q_p=A, B_c, he refers to the frequency distribution and finds that the assumptions P_p=A_c and Q_p=B_c are not good; on the other hand, assuming that P, Q_p=B, C_c, the frequency distribution gives excellent corroboration. A trial of these values would materially hasten solution because it is often the case in cryptanalysis that if the value of a very low-frequency letter can be surely established it will yield clues to other values very quickly. Thus, if Q_p is definitely identified it almost invariably will identify U_p, and will give clues to the letter following the U_p, since it must be a vowel. In the case under discussion the identification PQ_p=BC_c would have turned out to be correct. For the foregoing reason an attempt should always be made in the early stages of the analysis to determine, if possible, the basis of construction or derivation of the cipher alphabet; as a rule this can be done only by means of the enciphering alphabet, and not the deciphering alphabet. For example, the skeletonized *deciphering* alphabet corresponding to the enciphering alphabet directly above is as follows:

Cipher	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
Plain				R		T	S	O	U										A	N		I				E
						S	T																			

Here no evidences of a keyword-mixed alphabet are seen at all. However, if the enciphering alphabet has been examined and shows no evidences of systematic construction, the deciphering alphabet should then be examined with this in view, because occasionally it is the deciphering alphabet which shows the presence of a key or keying element, or which has been systematically derived from a word or phrase. The second reason why it is important to try to discover the basis of construction or derivation of the cipher alphabet is that it affords clues to the general type of keywords or keying elements employed by the enemy. This is a psychological factor, of course, and may be of assistance in subsequent studies of his traffic. It merely gives a clue to the general type of thinking indulged in by certain of his cryptographers.

d. In the case of the foregoing solution, the complete enciphering alphabet is found to be as follows:

Plain	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
Cipher	S	U	X	Y	Z	L	E	A	V	N	W	O	R	T	H	B	C	D	F	G	I	J	K	M	P	

Obviously, the letter Q, which is the only letter not appearing in the cryptogram, should follow P in the cipher component. Note now that the latter is based upon the keyword LEAVENWORTH, and that this particular cipher alphabet has been composed by shifting the mixed sequence based upon this keyword five intervals to the right so that the key for the message is A_p=S_c. Note also that the deciphering alphabet fails to give any evidence of keyword construction based upon the word LEAVENWORTH.

Cipher	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
Plain	H	P	Q	R	G	S	T	O	U	V	W	F	X	J	L	Y	Z	M	A	N	B	I	K	C	D	E

e. If neither the enciphering or the deciphering alphabet exhibits characteristics which give indication of derivation from a keyword by some form of mixing or disarrangement, the latter is nevertheless not finally excluded as a possibility. The student is referred to Section IX of *Elementary Military Cryptography*, wherein will be found methods for deriving mixed alphabets by transposition methods applied to keyword-mixed alphabets. For the reconstruction of such mixed alphabets the cryptanalyst must use ingenuity and a knowledge of the more common methods of suppressing the appearance of keywords in the mixed alphabets.

32. General notes on the foregoing solution.—*a.* The example solved above is admittedly a more or less artificial illustration of the steps in analysis, made so in order to demonstrate general principles. It was easy to solve because the frequencies of the various cipher letters corresponded quite well with the normal or expected frequencies. However, all cryptograms of the same monoalphabetical nature can be solved along the same general lines, after more or less experimentation, depending upon the length of the cryptogram, the skill, and the experience of the cryptanalyst.

b. It is no cause for discouragement if the student's initial attempts to solve a cryptogram of this type require much more time and effort than were apparently required in solving the foregoing purely illustrative example. It is indeed rarely the case that *every* assumption made by the cryptanalyst proves in the end to have been correct; more often is it the case that a good many of his initial assumptions are incorrect, and that he loses much time in casting out the erroneous ones. The speed and facility with which this elimination process is conducted is in many cases all that distinguishes the expert from the novice.

c. Nor will the student always find that the initial classification into vowels and consonants can be accomplished as easily and quickly as was apparently the case in the illustrative example. The principles indicated are very general in their nature and applicability, and there are, in addition, some other principles that may be brought to bear in case of difficulty. Of these, perhaps the most useful are the following:

(1) In normal English it is unusual to find two or three consonants in succession, each of high frequency. If in a cryptogram a succession of three or four letters of high-frequency appear in succession, it is practically certain that at least one of these represents a vowel.[3]

[3] Sequences of seven consonants are not impossible, however, as in STRENGTH THROUGH.

(2) Successions of three vowels are rather unusual in English.[4] Practically the only time this happens is when a word ends in two vowels and the next word begins with a vowel.[5]

(3) When two letters already classified as vowel-equivalents are separated by a sequence of six or more letters, it is either the case that one of the supposed vowel-equivalents is incorrect, or else that one or more of the intermediate letters is a vowel-equivalent.[6]

(4) Reference to Table 7–B of Appendix 1 discloses the following:

Distribution of first 18 digraphs forming 25 percent of English text

Number of consonant-consonant digraphs	4
Number of consonant-vowel digraphs	6
Number of vowel-consonant digraphs	8
Number of vowel-vowel digraphs	0

Distribution of first 53 digraphs forming 50 percent of English text

Number of consonant-consonant digraphs	8
Number of consonant-vowel digraphs	23
Number of vowel-consonant digraphs	18
Number of vowel-vowel digraphs	4

The latter tabulation shows that of the first 53 digraphs which form 50 percent of English text, 41 of them, that is, over 75 percent, are combinations of a vowel with a consonant. In short, in normal English the vowels and the high-frequency consonants are in the long run distributed fairly evenly and regularly throughout the text.

(5) As a rule, repetitions of trigraphs in the cipher text are composed of high-frequency letters forming high-frequency combinations. The latter practically always contain at least one vowel; in fact, if reference is made to Table 10–A of Appendix 1, it will be noted that 36 of the 56 trigraphs having a frequency of 100 or more contain one vowel, 17 of them contain two vowels, and only three of them contain no vowel. In the case of tetragraph repetitions, Table 11–A of Appendix 1 shows that no tetragraph listed therein fails to contain at least one vowel; 28 of them contain one vowel, 25 contain two vowels, and 2 contain three vowels.

(6) Quite frequently when two known vowel-equivalents are separated by six or more letters none of which seems to be of sufficiently high frequency to represent one of the vowels A E I O, the chances are good that the cipher-equivalent of the vowel U or Y is present.

(7) The letter Q is invariably followed by U; the letters J and V are invariably followed by a vowel.

d. In the foregoing example the amount of experimentation or "cutting and fitting" was practically nil. (This is not true of real cases as a rule.) Where such experimentation is necessary, the underscoring of all repetitions of several letters is very essential, as it calls attention to peculiarities of structure that often yield clues.

[4] Note that the word RADIOED, past tense of the verb RADIO, is coming into usage.

[5] A sequence of seven vowels is not impossible, however, as in THE <u>WAY YOU EARN</u>.

[6] Some cryptanalysts place a good deal of emphasis upon this principle as a method of locating the remaining vowels after the first two or three have been located. They recommend that the latter be underlined throughout the text and then all sequences of five or more letters showing no underlines be studied attentively. Certain letters which occur in several such sequences are sure to be vowels. An arithmetical aid in the study is as follows: Take a letter thought to be a good possibility as the cipher equivalent of a vowel (hereafter termed a *possible vowel-equivalent*) and find the length of each interval from the possible vowel-equivalent to the next *known* (fairly surely determined) vowel-equivalent. Multiply the interval by the number of times this interval is found. Add the products and divide by the total number of intervals considered. This will give the *mean* interval for that possible vowel-equivalent. Do the same for all the other possible vowel-equivalents. The one for which the mean is the greatest is most probably a vowel-equivalent. Underline this letter throughout the text and repeat the process for locating additional vowel-equivalents, if any remain to be located.

e. After a few basic assumptions of values have been made, if short words or skeletons of words do not become manifest, it is necessary to make further assumptions for unidentified letters. This is accomplished most often by assuming a word.[7] Now there are two places in every message which lend themselves more readily to successful attack by the assumption of words than do any other places—the very beginning and the very end of the message. The reason is quite obvious, for although words may begin or end with almost any letter of the alphabet, they usually begin and end with but a few very common digraphs and trigraphs. Very often the association of letters in peculiar combinations will enable the student to note where one word ends and the next begins. For example suppose, E, N, S, and T have been definitely identified, and a sequence like the following is found in a cryptogram:

. . . E N T S N E . . .

Obviously the break between two words should fall either after the S of E N T S or after the T of E N T, so that two possibilities are offered: . . . E N T S / N E . . ., or . . . E N T / S N E . . .. Since in English there are very few words with the initial trigraph S N E, it is most likely that the proper division is . . . E N T S / N E . . .. Obviously, when several word divisions have been found, the solution is more readily achieved because of the greater ease with which assumptions of additional new values may be made.

33. The "probable word" method; its value and applicability.—*a.* In practically all cryptanalytic studies, short-cuts can often be made by assuming the presence of certain words in the message under study. Some writers attach so much value to this kind of an "attack from the rear" that they practically elevate it to the position of a method and call it the "intuitive method" or the "probable-word method." It is, of course, merely a refinement of what in every-day language is called "assuming" or "guessing" a word in the message. The value of making a "good guess" can hardly be overestimated, and the cryptanalyst should never feel that he is accomplishing a solution by an illegitimate subterfuge when he has made a fortunate guess leading to solution. A correct assumption as to plain text will often save hours or days of labor, and sometimes there is no alternative but to try to "guess a word", for occasionally a system is encountered the solution of which is absolutely dependent upon this artifice.

b. The expression "good guess" is used advisedly. For it is "good" in two respects. First, the cryptanalyst must use care in making his assumptions as to plain-text words. In this he must be guided by extraneous circumstances leading to the assumption of *probable* words—not just any words that come to his mind. Therefore he must use his imagination but he must nevertheless carefully control it by the exercise of *good* judgment. Second, only if the "guess" is correct and leads to solution, or at least puts him on the road to solution, is it a *good* guess. But, while realizing the usefulness and the time and labor-saving features of a solution by assuming a probable word, the cryptanalyst should exercise discretion in regard to how long he may continue in his efforts with this method. Sometimes he may actually waste time by adhering to the method too long, if straightforward, methodical analysis will yield results more quickly.

c. Obviously, the "probable-word" method has much more applicability when working upon material the general nature of which is known, than when working upon more or less isolated communications exchanged between correspondents concerning whom or whose activities

[7] This process does not involve anything more mysterious than ordinary, logical reasoning; there is nothing of the subnormal or supernormal about it. If cryptanalytic success seems to require processes akin to those of medieval magic, if "hocus-pocus" is much to the fore, the student should begin to look for items that the claimant of such success has carefully hidden from view, for the mystification of the uninitiated. (See Par. 33 in this connection.)

nothing is known. For in the latter case there is little or nothing that the imagination can seize upon as a background or basis for the assumptions.[8]

d. Very frequently, the choice of probable words is aided or limited by the number and positions of repeated letters. These repetitions may be *patent*—that is, externally visible in the cryptographic text as it originally stands—or they may be *latent*—that is, externally invisible but susceptible of being made patent as a result of the analysis. For example, in a monoalphabetic substitution cipher, such as that discussed in the preceding paragraph, the repeated letters are directly exhibited in the cryptogram; later the student will encounter many cases in which the repetitions are latent, but are made patent by the analytical process. When the repetitions are patent, then the *pattern* or *formula* to which the repeated letters conform is of direct use in assuming plain-text words; and when the text is in word-lengths, the pattern is obviously of even greater assistance. Suppose the cryptanalyst is dealing with military text, in which case he may expect such words as DIVISION, BATTALION, etc., to be present in the text. The positions of the repeated letter I in DIVISION, of the reversible digraph AT, TA in BATTALION, and so on, constitute for the experienced cryptanalyst tell-tale indications of the presence of these words, even when the text is not divided up into its original word lengths.

e. The important aid that a study of word patterns can afford in cryptanalysis warrants the use of definite terminology and the establishment of certain data having a bearing thereon. The phenomenon herein under discussion, namely, that many words are of such construction as regards the number and positions of repeated letters as to make them readily identifiable, will be termed *idiomorphism* (from the Greek "idios" = one's own, individual, peculiar + "morphe" = form). Words which show this phenomenon will be termed *idiomorphic*. It will be useful to deal with the idiomorphisms symbolically and systematically as described below.

f. When dealing with cryptograms in which the word lengths are determined or specifically shown, it is convenient to indicate their lengths and their repeated letters in some easily recognized manner or by formulas. This is exemplified, in the case of the word DIVISION, by the formula ABCBDBEF; in the case of the word BATTALION, by the formula ABCCBDEFG. If the cryptanalyst, during the course of his studies, makes note of striking formulas he has encountered, with the words which fit them, after some time he will have assembled a quite valuable body of data. And after more or less complete lists of such formulas have been established in some systematic arrangement, a rapid comparison of the idiomorphs in a specific cryptogram with those in his lists will be feasible and will often lead to the assumption of the correct word. Such lists can be arranged according to word length, as shown herewith:

```
3/aba  : DID, EVE, EYE.
  abb  : ADD, ALL, ILL, OFF, etc.
4/abac : ARAB, AWAY, etc.
  abca : AREA, BOMB, DEAD, etc.
  abbc : . . .
  abcb : . . .
  etc.   etc.
```

[8] General Givierge in his *Cours de Cryptographie* (p. 121) says: "However, expert cryptanalysts often employ such details as are cited above [in connection with assuming the presence of 'probable words'], and the experience of the years 1914 to 1918, to cite only those, prove that in practice one often has at his disposal elements of this nature, permitting assumptions much more audacious than those which served for the analysis of the last example. The reader would therefore be wrong in imagining that such fortuitous elements are encountered only in cryptographic works where the author deciphers a document that he himself enciphered. Cryptographic correspondence, if it is extensive, and if sufficiently numerous working data are at hand, often furnishes elements so complete that an author would not dare use all of them in solving a problem for fear of being accused of obvious exaggeration."

g. When dealing with cryptographic text in which the lengths of the words are not indicated or otherwise determinable, lists of the foregoing nature are not so useful as lists in which the words (or parts of words) are arranged according to the intervals between identical letters, in the following manner:

1 Interval	2 Intervals	3 Intervals	Repeated digraphs
–DiD–	AbbAcy	AbeyAnce	COCOa
–EvE–	ArAbiA	hAbitAble	dERER
–EyE–	AbiAtive	lAborAtory	ICICle
dIvIsion	AboArd	AbreAst	ININg
revIsIon	–AciA–	AbroAd	bAGgAGe
etc.	etc.	etc.	etc.

34. Solution of additional cryptograms produced by the same cipher component.—*a.* To return, after a rather long digression, to the cryptogram solved in pars. 28–31, once the cipher component of a cipher alphabet has been reconstructed, subsequent messages which have been enciphered by means of the same cipher component may be solved very readily, and without recourse to the principles of frequency, or application of the probable-word method. It has been seen that the illustrative cryptogram treated in paragraphs 24–31 was enciphered by juxtaposing the cipher component against the normal sequence so that $A_p = S_c$. It is obvious that the cipher component may be set against the plain component at any one of 26 different points of coincidence, each yielding a different cipher alphabet. After a cipher component has been reconstructed, however, it becomes a *known* sequence, and the method of converting the cipher letters into their plain-component equivalents and then completing the plain-component sequence begun by each equivalent can be applied to solve any cryptogram which has been enciphered by that cipher component.

b. An example will serve to make the process clear. Suppose the following message, passing between the same two stations as before, was intercepted shortly after the first message had been solved:

```
IYEWK  CERNW  OFOSE  LFOOH  EAZXX
```

It is assumed that the same cipher component was used, but with a different key letter. First the initial two groups are converted into their plain-component equivalents by setting the cipher component against the normal sequence at any arbitrary point of coincidence. The initial letter of the former may as well be set against A of the latter, with the following result:

```
Plain........... 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
Cipher.......... L E A V N W O R T H B C D F G I J K M P Q S U X Y Z

          Cryptogram.... I Y E W K   C E R N W   . . .
          Equivalents... P Y B F R   L B H E F   . . .
```

The normal sequence initiated by each of these conversion equivalents is now completed, with the results shown in Fig. 15. Note the plain-text generatrix, CLOSEYOURS, which manifests itself without further analysis. The rest of the message may be read either by continuing the same process, or, what is even more simple, the key letter of the message may now be determined quite readily and the message deciphered by its means.

I	Y	E	W	K	C	E	R	N	W
P	Y	B	F	R	L	B	H	E	F
Q	Z	C	G	S	M	C	I	F	G
R	A	D	H	T	N	D	J	G	H
S	B	E	I	U	O	E	K	H	I
T	C	F	J	V	P	F	L	I	J
U	D	G	K	W	Q	G	M	J	K
V	E	H	L	X	R	H	N	K	L
W	F	I	M	Y	S	I	O	L	M
X	G	J	N	Z	T	J	P	M	N
Y	H	K	O	A	U	K	Q	N	O
Z	I	L	P	B	V	L	R	O	P
A	J	M	Q	C	W	M	S	P	Q
B	K	N	R	D	X	N	T	Q	R
*C	L	O	S	E	Y	O	U	R	S
D	M	P	T	F	Z	P	V	S	T
E	N	Q	U	G	A	Q	W	T	U
F	O	R	V	H	B	R	X	U	V
G	P	S	W	I	C	S	Y	V	W
H	Q	T	X	J	D	T	Z	W	X
I	R	U	Y	K	E	U	A	X	Y
J	S	V	Z	L	F	V	B	Y	Z
K	T	W	A	M	G	W	C	Z	A
L	U	X	B	N	H	X	D	A	B
M	V	Y	C	O	I	Y	E	B	C
N	W	Z	D	P	J	Z	F	C	D
O	X	A	E	Q	K	A	G	D	E

c. In order that the student may understand without question just what is involved in the latter step, that is, discovering the key letter after the first two or three groups have been deciphered by the conversion-completion process, the foregoing example will be used. It was noted that the first cipher group was finally deciphered as follows:

Cipher............ I Y E W K
Plain............. C L O S E

Now set the cipher component against the normal sequence so that $C_p=I_c$. Thus:

Plain.............. 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
Cipher............ F G I J K M P Q S U X Y Z L E A V N W O R T H B C D

It is seen here that when $C_p=I_c$ then $A_p=F_c$. This is the key for the entire message. The decipherment may be completed by direct reference to the foregoing cipher alphabet. Thus:

Cipher................... I Y E W K   C E R N W   O F O S E   L F O O H   E A Z X X
Plain..................... C L O S E   Y O U R S   T A T I O   N A T T W   O P M X X

Message: CLOSE YOUR STATION AT TWO PM

d. The student should make sure that he understands the fundamental principles involved in this quick solution, for they are among the most important principles in cryptanalytics. How useful they are will become clear as he progresses into more and more complex cryptanalytic studies.

SECTION VII

MULTILITERAL SUBSTITUTION WITH SINGLE-EQUIVALENT CIPHER ALPHABETS

35. Analysis of multiliteral, monoalphabetic substitution systems.—*a.* Substitution methods in general may be classified into uniliteral and multiliteral systems.[1] In the former there is a strict "one-to-one" correspondence between the length of the units of the plain and those of the cipher text; that is, each letter of the plain text is replaced by a single character in the cipher text. In the latter this correspondence is no longer $1_p:1_c$ but may be $1_p:2_c$, where each letter of the plain text is replaced by a combination of two characters in the cipher text; or $1_p:3_c$, where a 3-character combination in the cipher text represents a single letter of the plain text, and so on. A cipher in which the correspondence of the $1_p:1_c$ type is termed uniliteral in character; one in which it is of the $1_p:2_c$ type, biliteral; $1_p:3_c$, triliteral, and so on. Those beyond the $1_p:1_c$ type are classed together as *multiliteral.*

b. When a multiliteral system employs biliteral equivalents, the cipher alphabet is said to be bipartite. Such alphabets are composed of a set of 25 or 26 combinations of a limited number of characters taken in pairs. An example of such an alphabet is the following.

Plain...............	A	B	C	D	E	F	G	H	I	J	K	L	M
Cipher...............	WW	WH	WI	WT	WE	HW	HH	HI	HT	HT	HE	IW	IH
Plain...............	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
Cipher...............	II	IT	IE	TW	TH	TI	TT	TE	EW	EH	EI	ET	EE

This alphabet is derived from the square shown in Fig. 15.

(2)

		W	H	I	T	E
	W	A	B	C	D	E
	H	F	G	H	I–J	K
(1)	I	L	M	N	O	P
	T	Q	R	S	T	U
	E	V	W	X	Y	Z

FIGURE 15.

c. If a message is enciphered by means of the foregoing bipartite alphabet the cryptogram is still monoalphabetic in character. A frequency distribution based upon pairs of letters will

[1] See Sec. VII, *Advanced Military Cryptography.*

obviously have all the characteristics of a simple, uniliteral distribution for a monoalphabetic substitution cipher.

d. Ciphers of this type, as well as of those of the multiliteral (triliteral, quadraliteral, . . .) type are readily detected externally by virtue of the fact that the cryptographic text is composed of but a very limited number of different characters. They are handled in exactly the same manner as are uniliteral, monoalphabetic substitution ciphers. So long as the same character, or combination of characters, is always used to represent the same plain-text letter, and so long as a given letter of the plain text is always represented by the same character or combination of characters, the substitution is strictly monoalphabetic and can be handled in the simple manner described under Par. 31 of this text.

e. An interesting example in which the cipher equivalents are quinqueliteral groups and yet the resulting cipher is strictly monoalphabetic in character is found in the cipher system invented by Sir Francis Bacon over 300 years ago. Despite its antiquity the system possesses certain features of merit which are well worth noting. Bacon[2] proposed the following cipher alphabet, composed of permutations of two elements taken five at a time:[3]

A=aaaaa	I–J=abaaa	R=baaaa
B=aaaab	K=abaab	S=baaab
C=aaaba	L=ababa	T=baaba
D=aaabb	M=ababb	U–V=baabb
E=aabaa	N=abbaa	W=babaa
F=aabab	O=abbab	X=babab
G=aabba	P=abbba	Y=babba
H=aabbb	Q=abbbb	Z=babbb

If this were all there were to Bacon's invention it would be hardly worth bringing to attention. But what he pointed out, with great clarity and simple examples, was how such an alphabet might be used to convey a secret message by enfolding it in an innocent, external message which might easily evade the strictest kind of censorship. As a very crude example, suppose that a message is written in capital and lower case letters, any capital letter standing for an "a" element of the cipher alphabet, and any small letter, for a "b" element. Then the external sentence "All is well with me today" can be made to contain the secret message "Help." Thus:

```
ALl  is  WElL  WItH  mE  TodaY
aab  bb  aaba  aaba  ba  abbba
   H        E        L      P
```

Instead of employing such an obvious device as capital and small letters, suppose that an "a" element be indicated by a very slight shading, or a very slightly heavier stroke. Then a secret message might easily be thus enfolded within an external message of exactly opposite meaning. The number of possible variations of this basic scheme is very high. The fact that the characters

[2] For a true picture of this cipher, the explanation of which is often distorted beyond recognition even by cryptographers, see Bacon's own description of it as contained in his *De Augmentis Scientiarum* (*The Advancement of Learning*), as translated by any first-class editor, such as Gilbert Watts (1640) or Ellis, Spedding, and Heath (1857, 1870). The student is cautioned, however, not to accept as true any alleged "decipherments" obtained by the application of Bacon's cipher to literary works of the 16th century. These readings are purely subjective.

[3] In the 16th Century, the letters I and J were used interchangeably, as were also U and V. Bacon's alphabet was called by him a "biliteral alphabet" because it employs permutations of two letters. But from the cryptanalytic standpoint the significant point is that each plain-text letter is represented by a 5-character equivalent. Hence, present terminology requires that this alphabet be referred to as a *quinqueliteral alphabet.*

of the cryptographic text are hidden in some manner or other has, however, no effect upon the strict monoalphabeticity of the scheme.

36. Historically interesting examples.—*a.* Two examples of historical interest will be cited in this connection as illustrations. During the campaign for the presidential election of 1876 many cipher messages were exchanged between the Tilden managers and their agents in several states where the voting was hotly contested. Two years later the New York Tribune[4] exposed many irregularities in the campaign by publishing the decipherments of many of these messages. These decipherments were achieved by two investigators employed by the Tribune, and the plain text of the messages seems to show that illegal attempts and measures to carry the election for Tilden were made by his managers. Here is one of the messages:

```
                                         JACKSONVILLE, Nov. 16 (1876).
GEO. F. RANEY, Tallahassee.
    P p y y e m n s n y y y p i m a s h n s y y s s i t e p a a e n s h n s
p e n s s h n s m m p i y y s n p p y e a a p i e i s s y e s h a i n s s s p
e e i y y s h n y n s s s y e p i a a n y i t n s s h y y s p y y p i n s y y
s s i t e m e i p i m m e i s s e i y y e i s s i t e i e p y y p e e i a a s s
i m a a y e s p n s y y i a n s s s e i s s m m p p n s p i n s s n p i n s i m
i m y y i t e m y y s s p e y y m m n s y y s s i t s p y y p e e p p p m a
a a y y p i i t
L'Engle goes up tomorrow.

                                                                 DANIEL.
```

Examination of the message discloses that only ten different letters are used. It is probable, therefore, that what one has here is a cipher which employs a bipartite alphabet and in which combinations of two letters represent single letters of the plain text. The message is therefore rewritten in pairs and substitution of arbitrary letters for the pairs is made, as seen below:

PP	YY	EM	NS	NY	YY	PI	MA	SH	NS	YY	SS	etc.
A	B	C	D	E	B	F	G	H	D	B	I	etc.

A triliteral frequency distribution is then made and analysis of the message along the lines illustrated in the preceding section of this text yields solution, as follows:

JACKSONVILLE, *Nov. 16.*

GEO. F. RANEY, *Tallahassee:*

Have Marble and Coyle telegraph for influential men from Delaware and Virginia. Indications of weakening here. Press advantage and watch Board. L'Engle goes up tomorrow.

DANIEL.

b. The other example, using numbers, is as follows:

JACKSONVILLE, Nov. 17.

S. PASCO and E. M. L'ENGLE:

84	55	84	25	93	34	82	31	31	75	93	82	77	33	55	42
93	20	93	66	77	66	33	84	66	31	31	93	20	82	33	66
52	48	44	55	42	82	48	89	42	93	31	82	66	75	31	93

DANIEL.

[4] New York Tribune, Extra No. 44, *The Cipher Dispatches*, New York, 1879.

There were, of course, several messages of like nature, and examination disclosed that only 26 different numbers in all were used. Solution of these ciphers followed very easily, the decipherment of the one given above being as follows:

JACKSONVILLE, *Nov. 17.*

S. PASCO and E. M. L'ENGLE:

Cocke will be ignored, Eagan called in. Authority reliable.

DANIEL.

c. The Tribune experts gave the following alphabets as the result of their decipherments:

AA=O	EN=Y	IT=D	NS=E	PP=H	SS=N
AI=U	EP=C	MA=B	NY=M	SH=L	YE=F
EI=I	IA=K	MM=G	PE=T	SN=P	YI=X
EM=V	IM=S	NN=J	PI=R	SP=W	YY=A
20=D	33=N	44=H	62=X	77=G	89=Y
25=K	34=W	48=T	66=A	82=I	93=E
27=S	39=P	52=U	68=F	84=C	96=M
31=L	42=R	55=O	75=B	87=V	99=J

They did not attempt to correlate these alphabets, or at least they say nothing about a possible relationship. The present author has, however, reconstructed the rectangle upon which these alphabets are based, and it is given below (fig. 16).

2d Letter or Number

1st Letter or Number	H 1	I 2	S 3	P 4	A 5	Y 6	M 7	E 8	N 9	T 0
H 1										
I 2					K		S			D
S 3	L		N	W					P	
P 4		R		H				T		
A 5		U			O					
Y 6		X				A		F		
M 7					B		G			
E 8		I		C			V		Y	
N 9			E			M			J	
T 0										

FIGURE 16.

It is amusing to note that the conspirators selected as their key a phrase quite in keeping with their attempted illegalities—HIS PAYMENT—for bribery seems to have played a considerable part in that campaign. The blank squares in the diagram probably contained proper names, numbers, etc.

SECTION VIII

MULTILITERAL SUBSTITUTION WITH MULTIPLE-EQUIVALENT CIPHER ALPHABETS

37. Purpose of providing multiple-equivalent cipher alphabets.—*a.* It has been seen that the characteristic frequencies of letters composing normal plain text, the associations they form in combining to form words, and the peculiarities certain of them manifest in such text all afford direct clues by means of which ordinary monoalphabetic substitution encipherments of such plain text may be more or less speedily solved. This has led to the introduction of simple methods for disguising or suppressing the manifestations of monoalphabeticity, so far as possible. Basically these methods are multiliteral and they will now be presented.

b. Multiliteral substitution may be of two types: (1) That wherein each letter of the plain text is represented by one and only one multiliteral equivalent. For example, in the Francis Bacon cipher described in Par. 35*e*, the letter K_p is invariably represented by the permutation abaab. For this reason this type of system may be more completely described as *monoalphabetic, multiliteral substitution with single-equivalent cipher alphabets.*

(2) That wherein, because of the large number of equivalents made available by the combinations and permutations of a limited number of elements, each letter of the plain text may be represented by several multiliteral equivalents which may be selected at random. For example, if 3-letter combinations are employed there are available 26^3 or 17,576 equivalents for the 26 letters of the plain text; they may be assigned in equal numbers of different equivalents for the 26 letters, in which case each letter would be representable by 676 different 3-letter equivalents; or they may be assigned on some other basis, for example, proportionately to the relative frequencies of plain-text letters. For this reason this type of system may be more completely described as *monoalphabetic, multiliteral substitution with multiple-equivalent cipher alphabets.* Some authors term such a system "simple substitution with multiple equivalents"; others term it *monoalphabetic substitution with variants.* For the sake of brevity, the latter designation will be employed in this text.

c. The primary object of monoalphabetic substitution with variants is, as has been mentioned above, to provide several values which may be employed at random in a simple substitution of cipher equivalents for the plain-text letters. In this connection, reference is made to Section X of *Elementary Military Cryptography*, wherein several of the most common methods for producing and using variants are set forth.

d. A word or two concerning the underlying theory from the cryptanalytic point of view of monoalphabetic substitution with variants, may not be amiss. Whereas in simple or single-equivalent, monoalphabetic substitution it is seen that—

(1) The same letter of the plain text is invariably represented by but one and always the same character of the cryptogram, and

(2) The same character of the cryptogram invariably represents one and always the same letter of the plain text;

In multiliteral substitution with multiple equivalents (monoalphabetic substitution with variants) it is seen that—

(1) The same letter of the plain text may be represented by one or more different characters of the cryptogram, but

(2) The same character of the cryptogram nevertheless invariably represents one and always the same letter of the plain text.

38. Solution of a simple example.—*a.* The following cryptogram has been enciphered by a set of four alphabets similar to the following:

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
08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	01	02	03	04	05	06	07
35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	26	27	28	29	30	31	32	33	34
68	69	70	71	72	73	74	75	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67
87	88	89	90	91	92	93	94	95	96	97	98	99	00	76	77	78	79	80	81	82	83	84	85	86

The keyword here is TRIP[1]. In enciphering a message the equivalents are to be selected at random from among the four variants for each letter. The steps in solving a message produced by such a scheme will now be scrutinized.

CRYPTOGRAM

```
68321 09022 48057 65111 88648 42036 45235 09144 05764 22684
00225 57003 97357 14074 82524 40768 51058 93074 92188 47264
09328 04255 06186 79882 85144 45886 32574 55136 56019 45722
76844 68350 45219 71649 90528 65106 11886 44044 89669 70553
18491 06985 48579 33684 50957 70612 09795 29148 56109 08546
62062 65509 32800 32568 97216 44282 34031 84989 68564 53789
12530 77401 68494 38544 11368 87616 56905 20710 58864 67472
22490 09136 62851 24551 35180 14230 50886 44084 06231 12876
05579 58980 29503 99713 32720 36433 82689 04516 52263 21175
06445 72255 68951 86957 76095 67215 53049 08567 9730
```

b. Assuming that the foregoing remarks had not been made and that the cryptogram has just been submitted for solution with no information concerning it, the first step is to make a preliminary study to determine whether the cryptogram involves cipher or code. The cryptogram appears in 5-figure groups, which may indicate either cipher or code. A few remarks will be made at this point with reference to the method of determining whether a cryptogram composed of figure groups is in code or cipher, using the foregoing example.

c. In the first place, if the cryptogram contains an even number of digits, as for example 494 in the foregoing message, this leaves open the possibility that it may be cipher, composed of 247 pairs of digits; were the number of digits an exact *odd* multiple of five, such as 125, 135, etc., the possibility that the cryptogram is in code of the 5-figure group type must be considered. Next, a preliminary study is made to see if there are many repetitions, and what their characteristics

[1] The letter corresponding to the lowest number in each line of the diagram showing the cipher alphabets is a key letter. Thus, in the 1st line 01=T; in the 2d line 26=R; etc.

are. If the cryptogram is code of the 5-figure group type, then such repetitions as appear should *generally* be in whole groups of five digits, and they should be visible in the text just as the message stands, unless the code message has undergone encipherment also. If the cryptogram is in cipher, then the repetitions should extend beyond the 5-digit groupings; if they conform to any definite groupings at all they should for the most part contain even numbers of digits since each letter is probably represented by a pair of digits. If no clues of the foregoing nature are present, doubts will be dissolved by making a detailed study of frequencies.

d. A simple 4-part frequency distribution is therefore decided upon. Shall the alphabet be assumed to be a 25- or a 26-character one? If the former, then the 2-digit pairs from 01 to 00 fall into exactly four groups each corresponding to an alphabet. Since this is the most common scheme of drawing up such alphabets, let it be assumed to be true of the present case. The following distributions result from the breaking up of the text into 2-digit pairs.

01—///	26—///	51—~~////~~	76—~~////~~ /
02—	27—	52—~~////~~	77—/
03—////	28—/	53—///	78—
04—/	29—/	54—	79—/
05—~~////~~	30—///	55—////	80—///
06—~~////~~ /	31—	56—~~////~~	81—
07—///	32—~~////~~ /	57—~~////~~ /	82—////
08—	33—/	58—//	83—/
09—////	34—/	59—	84—~~////~~ /
10—////	35—//	60—	85—~~////~~ /
11—~~////~~	36—~~////~~	61—	86—///
12—///	37—/	62—//	87—
13—/	38—	63—	88—////
14—/	39—/	64—~~////~~ /	89—~~////~~
15—/	40—///	65—	90—~~////~~ /
16—///	41—	66—/	91—///
17—	42—////	67—//	92—/
18—~~////~~ /	43—/	68—~~////~~ //	93—/
19—	44—~~////~~ /	69—//	94—/
20—/	45—~~////~~ /	70—/	95—///
21—//	46—///	71—/	96—
22—~~////~~	47—	72—////	97—~~////~~ /
23—//	48—///	73—	98—/
24—	49—~~////~~	74—////	99—
25—/	50—~~////~~	75—/	00—//

e. If the student will bring to bear upon this problem the principles he learned in Section V of this text, he will soon realize that what he now has before him are four, simple, monoalphabetic frequency distributions similar to those involved in a monoalphabetic substitution cipher using standard cipher alphabets. The realization of this fact immediately provides the clue to the next step: "fitting each of the distributions to the normal." (See Par. 17*b*). This can be done without difficulty in this case (remembering that a 25-letter alphabet is involved and assuming that I and J are the same letter) and the following alphabets result:

01—I–J	26—U	51—N	76—E
02—K	27—V	52—O	77—F
03—L	28—W	53—P	78—G
04—M	29—X	54—Q	79—H
05—N	30—Y	55—R	80—I–J
06—O	31—Z	56—S	81—K
07—P	32—A	57—T	82—L
08—Q	33—B	58—U	83—M
09—R	34—C	59—V	84—N
10—S	35—D	60—W	85—O
11—T	36—E	61—X	86—P
12—U	37—F	62—Y	87—Q
13—V	38—G	63—Z	88—R
14—W	39—H	64—A	89—S
15—X	40—I–J	65—B	90—T
16—Y	41—K	66—C	91—U
17—Z	42—L	67—D	92—V
18—A	43—M	68—E	93—W
19—B	44—N	69—F	94—X
20—C	45—O	70—G	95—Y
21—D	46—P	71—H	96—Z
22—E	47—Q	72—I–J	97—A
23—F	48—R	73—K	98—B
24—G	49—S	74—L	99—C
25—H	50—T	75—M	00—D

f. The keyword is seen to be JUNE and the first few groups of the cryptogram decipher as follows:

68	32	10	90	22	48	05	76	51	11	88	64	84	20	36	45	23
E	A	S	T	E	R	N	E	N	T	R	A	N	C	E	O	F

g. From the detailed procedure given above, the student should be able to draw his own conclusions as to the procedure to be followed in solving cryptograms produced by methods which are more or less simple variations of that just discussed. In this connection he is referred to Section X of *Elementary Military Cryptography*, wherein a few of these variations are mentioned.

h. Possibly the most important of the variations is that in which a rectangle such as that shown in Fig. 17 is employed.

	1	2	3	4	5	6	7	8	9	0
1, 4, 7	A	B	C	D	E	F	G	H	I	J
2, 5, 8	K	L	M	N	O	P	Q	R	S	T
3, 6, 9	U	V	W	X	Y	Z	–	,	:	;

FIGURE 17

In the solution of cases of this kind, repetitions would play their usual role, with the modifications noted below in Par. 39. Once an entering wedge has been forced, through the identification of one or more repeated words such as BATTALION, DIVISION, etc., the entire enciphering rectangle would soon be reconstructed. It may be added that the frequency distribution for the text of a single long message or several short ones enciphered by such a system would show characteristic phenomena, the most important of which are, first, that the distribution for a rectangle such as shown in Fig. 17 would practically follow the normal and, second, that the distribution for the 2d digit of pairs would show more marked crests and troughs than the distribution for the 1st digit. For example, the initial digits 1, 4, and 7 (for the numbers 10–19, 40–49, and 70–79, inclusive) would apply to the distribution for the letters A to J, inclusive; the initial digits 2, 5, and 8 would apply to the distribution for the letters K to T, inclusive. The total weighted frequency values for these two groups of letters are about equal. Therefore, the frequencies of the initial digits 1, 2, 4, 5, 7, and 8 would be approximately equal. But consider the final digit 5 in the numbers 15, 45, 75, 25, 55, and 85; its total frequency is composed of the frequency of E_p plus the frequency of O_p; whereas in the case of the final digit 6, its total frequency is composed of the frequency of F_p plus the frequency of Q_p. The two cases would show a marked difference in frequency. Of course, the letters may be inserted within the enciphering rectangle in a keyword-mixed or even in a random order; the numbers may be applied to the rectangle in a random order. But these variations, while increasing the difficulty in solution, by no means make the latter as great as may be thought by the novice.

39. Solution of a more complicated example.—*a*. As soon as a beginner in cryptography realizes the consequences of the fact that letters are used with greatly varying frequencies in normal plain text, a brilliant idea very speedily comes to him. Why not disguise the natural frequencies of letters by a system of substitution using many equivalents, and let the numbers of equivalents assigned to the various letters be more or less in direct proportion to the normal frequencies of the letters? Let E, for example, have 13 or more equivalents; T, 10; N, 9; etc., and thus (he thinks) the enemy cryptanalyst can have nothing in the way of tell-tale or characteristic frequencies to use as an entering wedge.

b. If the text available for study is small in amount and if the variant values are wholly independent of one another, the problem can become exceedingly difficult. But in practical military communications such methods are rarely encountered, *because the volume of text is usually great enough to permit of the establishment of equivalent values*. To illustrate what is meant, suppose a set of cryptograms produced by the monoalphabetic-variant method described above shows the following two sets of groupings in the text:

SET A	SET B
12–37–02–79–68–13–03–37–77	71–12–02–51–23–05–77
82–69–03–79–13–68–23–37–35	11–82–51–02–03–05–35
82–69–51–16–13–13–78–05–35	11–91–02–02–23–37–35
91–05–02–01–68–42–78–37–77	97–12–51–03–78–69–77

An examination of these groupings would lead to the following tentative conclusions with regard to probable equivalents:

12, 82, 91	01, 16, 79	03, 23, 78
05, 37, 69	13, 42, 68	35, and 77
02, and 51		

The establishment of these equivalencies would sooner or later lead to the finding of additional sets of equal values. The completeness with which this can be accomplished will determine the ease or difficulty of solution. Of course, if many equivalencies can be established the problem can then be reduced practically to monoalphabetic terms and a speedy solution can be attained.

c. Theoretically, the determination of equivalencies may seem to be quite an easy matter, but practically it may be very difficult, because the cryptanalyst can never be *certain* that a combination showing what may appear to be a variant value is really such, and is not a different word. For example, take the groups—

17–82–31–82–14–63, and
27–82–40–82–14–63

Here one might suspect that 17 and 27 represent the same letter, 31 and 40 another letter. But it happens that one group represents the word MANAGE, the other DAMAGE.

d. When reversible combinations are used as variants, the problem is perhaps a bit more simple. For example, using the accompanying Fig. 18 for encipherment, two messages with the same initial words, REFERENCE YOUR, may be enciphered as follows:

	K,Z	Q,V	B,H	M,R	D,L
W,S	N	H	A	O	E
F,X	D	T	M	F	P
G,J	Q	B	U	I	V
C,N	G	X	R	C	S
P,T	Z	L	Y	W	K

FIGURE 18.

```
     R E  F  E R   E N  C  E Y   O U  R
(1)  NHWDR   XLSHC   DWWZN   RSLHP   SRBJC   H
(2)  CHDWR   XSLHN   DWZWN   RLSHP   RWJBN   H
```

The experienced cryptanalyst, noting the appearance of the very first few groups, assumes that he is here confronted with a case involving biliteral reversible equivalents, with variants.

e. The probable-word method of solution may be used, but with a slight variation introduced by virtue of the fact that, regardless of the system, *letters of low frequency in plain text remain infrequent*. Hence, suppose a word containing low-frequency letters, but in itself a rather common word strikingly idiomorphic in character is sought as a "probable word"; for example, words such as CAVALRY, ATTACK, and PREPARE. Writing such a word on a slip of paper, it is slid one interval at a time under the text, which has been marked so that the high and low-frequency characters are indicated. Each coincidence of a low-frequency letter of the text with a low-frequency letter of the assumed word is examined carefully to see whether the adjacent text letters correspond in frequency with the other letters of the assumed word; or, if the latter presents repetitions, whether there are correspondences between repetitions in the text and those in the word. Many trials are necessary but this method will produce results when the difficulties are otherwise too much for the cryptanalyst to overcome.

40. A subterfuge to prevent decomposition of cipher text into component units.—*a*. A few words should be added with regard to certain subterfuges which are sometimes encountered in monoalphabetic substitution with variants, and which, if not recognized in time, cause considerable delays. These have to deal with the insertion of nulls so as to prevent the cryptanalyst from breaking up the text into its real cryptographic units. The student should take careful

note of the last phrase; the mere insertion of symbols having the same characteristics as the symbols of the cryptographic text, except that they have no meaning, is not what is meant. This class of nulls rarely achieves the purpose for which they are intended. What is really meant can best be explained in connection with an example. Suppose that a 5 x 5 checkerboard design with the row and column indicators shown in Fig. 19 is adopted for encipherment. Normally, the cipher units would consist of 2-letter combinations of the indicators, invariably giving the row indicator first (by agreement).

	V	G	I	W	D
	A	H	P	S	M
	T	O	E	B	N
	F	U	R	L	C
V,A,T,F	A	B	C	D	E
G,H,O,U	F	G	H	I–J	K
I,P,E,R	L	M	N	O	P
W,S,B,L	Q	R	S	T	U
D,M,N,C	V	W	X	Y	Z

FIGURE 19.

The phrase COMMANDER OF SPECIAL TROOPS might be enciphered thus:

```
C   O   M   M   A   N   D   E   R   O   F   . . .
VI  EB  PH  IU  FT  IE  AB  TM  WO  PW  GT  . . .
```

These would normally then be arranged in 5-letter groups, thus:

VIEBP HIUFT IEABT MWOPW GT . . .

b. It will be noted, however, that only 20 of the 26 letters of the alphabet have been employed as row and column indicators, leaving J, K, Q, X, Y, and Z unused. Now, suppose these five letters are used as nulls, *not in pairs, but as individual letters inserted at random* just before the real text is arranged in 5-letter groups. Occasionally, a pair of nulls is inserted. Thus, for example:

VIEXB PHKIU FJXTI EAJBT MWOQP WGKTY

The cryptanalyst, after some study, suspecting a biliteral cipher, proceeds to break up the text into pairs:

VI EX BP HK IU FJ XT IE AJ BT MW OQ PW GK TY

Compare this set of 2-letter combinations with the correct set. Only 4 of the 15 pairs are "proper" units. It is easy to see that without a knowledge of the *existence* of the nulls, and even with a knowledge, if he does not know *which* letters are nulls, the cryptanalyst would be confronted with a problem for the solution of which a fairly large amount of text might be necessary. The careful employment of the variants also very materially adds to the security of the method because repetitions can be rather effectively suppressed.

c. From the cryptographic standpoint, the fact that in this system the cryptographic text is more than twice as long as the plain text constitutes a serious disadvantage. From the cryptanalytic standpoint, the masking of the cipher units constitutes the most important source of strength of the system; this, coupled with the use of variants, makes it a bit more difficult system to solve, despite its monoalphabeticity.

SECTION IX

POLYGRAPHIC SUBSTITUTION SYSTEMS

41. Monographic and polygraphic substitution systems.—*a.* The student is now referred to Sections VII and VIII of *Advanced Military Cryptography*, wherein polygraphic systems of substitution are discussed from the cryptographic point of view. These will now be discussed from the cryptanalytic point of view.

b. Although the essential differences between polyliteral and polygraphic substitution are treated with some detail in Section VII of *Advanced Military Cryptography*, a few additional words on the subject may not be amiss at this point.

c. The two primary divisions of substitution systems into (1) uniliteral and multiliteral methods and into (2) monographic and polygraphic methods are both based upon considerations as to the *number of elements* constituting the plain-text and the equivalent cipher-text units. In uniliteral as well as in monographic substitution, each plain-text unit consists of a single element and each cipher-text unit consists of a single element. The two terms uniliteral and monographic are therefore identical in significance, as defined cryptographically. It is when the terms multiliteral and polygraphic are examined that an essential difference is seen. In multiliteral substitution the plain-text unit always consists of a single element (one letter) and the cipher-text unit consists of a group of two or more elements; when biliteral, it is a pair of elements, when triliteral, it is a set of three elements, and so on. In what will herein be designated as true or complete polygraphic substitution the plain-text unit consists of two or more elements forming an *indivisible compound;* the cipher-text unit usually consists of a corresponding number of elements.[1] When the number of elements comprising the plain-text units is fixed and always two, the system is *digraphic;* when it is three, the system is *trigraphic;* when it is four, *tetragraphic;* and so on.[2] It is important to note that in true or complete polygraphic substitution the elements combine to form indivisible compounds having properties different from those of either of the constituent letters. For example, in uniliteral substitution AB_p may yield XY_c and AC_p may yield XZ_c; but in true digraphic substitution $\overline{AB}_p$ may yield $\overline{XY}_c$ and $\overline{AC}_p$ may yield $\overline{QN}_c$. A difference in identify of one letter affects the whole result.[3] An analogy is found in chemistry, when two elements combine to form a molecule, the latter usually having properties quite different from those of either of the constituent elements. For example: sodium, a metal, and

[1] The qualifying adverb "usually" is employed because this correspondence is not essential. For example, if one should draw up a set of 676 arbitrary single signs, it would be possible to represent the 2-letter pairs from AA to ZZ by single symbols. This would still be a digraphic system.

[2] In this sense a code system is merely a polygraphic substitution system in which the number of elements constituting the plain-text units is variable.

[3] For this reason the two letters are marked by a ligature, that is, by a bar across their tops.

chlorine, a gas, combine to form sodium chloride, common table salt. Furthermore, sodium and fluorine, also a gas similar in many respects to chlorine, combine to form sodium fluoride, which is much different from table salt. Partial and pseudo-polygraphic substitution will be treated under subparagraphs *d* and *e* below.

d. Another way of looking at polygraphic substitution is to regard the elements comprising the plain-text units as being enciphered individually and polyalphabetically by a fairly large number of separate alphabets. For example, in a digraphic system in which 676 pairs of plain-text letters are representable by 676 cipher-text pairs assigned at random, this is equivalent to having a set of 26 different alphabets for enciphering one member of the pairs, and another set of 26 different alphabets for enciphering the other member of the pairs. According to this viewpoint the different alphabets are brought into play by the particular combination of letters forming each plain-text pair. This is, of course, quite different from systems wherein the various alphabets are brought into play by more definite rules; it is perhaps this very absence of definite rules guiding the selection of alphabets which constitutes the cryptographic strength of this type of polygraphic system.

e. When regarded in the light of the preceding remarks, certain systems which at first glance seem to be polygraphic, in that groupings of plain-text letters are treated as units, on closer inspection are seen to be only partially polygraphic, or pseudo-polygraphic in character. For example, in a system in which encipherment is by pairs and yet one of the letters in each pair is enciphered monoalphabetically, the other letter, polyalphabetically, the method is only *psuedo*-polygraphic. Cases of this type are shown in Section VII of *Advanced Military Cryptography*. Again, in a system in which encipherment is by pairs and the encipherments of the left-hand and right-hand members of the pairs show group relationships, this is not pseudo-polygraphic but only *partially* polygraphic. Cases of this type are also shown in the text referred to above.

f. The fundamental purpose of polygraphic substitution is again the suppression of the frequency characteristics of plain text, just as is the case in monoalphabetic substitution with variants; but here this is accomplished by a different method, the latter arising from a somewhat different approach to the problem involved in producing cryptographic security. When the substitution involves replacement of *single* letters in a monoalphabetic system, the cryptogram can be solved rather readily. Basically the reason for this is that the principles of frequency and the laws of probability, applied to individual units of the text (single letters), have a very good opportunity to manifest themselves. A given volume of text of say n plain-text letters, enciphered purely monoalphabetically, affords n cipher characters, and the same number of cipher units. The same volume of text, enciphered digraphically, still affords n cipher characters but only $\frac{n}{2}$ cipher units. Statistically speaking, the sample within which the laws of probability now apply has been cut in half. Furthermore, from the point of view of frequency, the very noticeable diversity in the frequencies of individual letters, leading to the marked crests and troughs of the uniliteral frequency distribution, is no longer so strikingly in evidence in the frequencies of digraphs. Therefore, although true digraphic encipherment, for example, cuts the cryptographic textual units in half, the difficulty of solution is not doubled, but, if a matter of judgment arising from practical experience can be expressed or approximated mathematically, squared or cubed.

g. Sections VII and VIII of *Advanced Military Cryptography* show various methods for the derivation of polygraphic equivalents and for handling these equivalents in cryptographing and decryptographing messages. The most practicable of those methods are digraphic in character and for this reason their solution will be treated in a somewhat more detailed manner than will trigraphic methods. The latter can be passed over with the simple statement that their analysis requires much text to permit of solution by the frequency method, and hard labor. Fortunately, they are infrequently encountered because they are difficult to manipulate without extensive tables.[4] If the latter are required they must be compiled in the form of a book or pamphlet. If one is willing to go that far, one might as well include in such document more or less extensive lists of words and phrases, in which case the system falls under the category of code and not cipher.

42. Tests for identifying digraphic substitution.—*a.* The tests which are applied to determine whether a given cryptogram is digraphic in character are usually rather simple. If there are many repetitions in the cryptogram and yet the uniliteral-frequency distribution gives no clear-cut indications of monoalphabeticity; if most of the repetitions contain an even number of letters; and if the cryptogram contains an even number of letters, it may be assumed to be digraphic in nature.

b. The student should first try to determine whether the substitution is completely digraphic, or only partially digraphic, or pseudo-digraphic in character. As mentioned above, there are cases in which, although the substitution is effected by taking pairs of letters, one of the members of the pairs is enciphered monoalphabetically, the other member, polyalphabetically. A distribution based upon the letters in the odd positions and one based upon those in the even positions should be made. If one of these is clearly monoalphabetic, then this is evidence that the message represents a case of pseudo-digraphism of the type here described. By attacking the monoalphabetic portion of the messages, solution can soon be reached by slight variation of the usual method, the polyalphabetic portion being solved by the aid of the context and considerations based upon the probable nature of the substitution chart. (See Tables 2, 3, and 4 of *Advanced Military Cryptography.*) It will be noted that the charts referred to show definite symmetry in their construction.

c. On the other hand, if the foregoing steps prove fruitless, it may be assumed that the cryptogram is completely digraphic in character.

d. Just as certain statistical tests may be applied to a cryptogram to establish its monoalphabeticity, so also may a statistical test be applied to a cryptogram for the purpose of establishing its digraphicity. The nature of this test and its method of application will be discussed in a subsequent text.

43. General procedure in the analysis of digraphic substitution ciphers.—*a.* The analysis of cryptograms which have been produced by digraphic substitution is accomplished largely by the application of the simple principles of frequency of digraphs, with the additional aid of such special circumstances as may be known to or suspected by the cryptanalyst. The latter refer to peculiarities which may be the result of the particular method employed in obtaining the equivalents of the plain-text digraphs in the cryptographing process. In general, however, only if there is sufficient text to disclose the normal phenomena of repetition will solution be feasible or possible.

b. However, when a digraphic system is employed in regular service, there is little doubt but that traffic will rapidly accumulate to an amount more than sufficient to permit of solution by simple principles of frequency. Sometimes only two or three long messages, or a half dozen of average length are sufficient. For with the identification of only a few cipher digraphs, larger portions of messages may be read because the skeletons of words formed from the few high-frequency digraphs very definitely limit the values that can be inserted for the intervening unidentified digraphs. For example, suppose that the plain-text digraphs TH, ER, IN, IS, OF, NT, and TO have been identified by frequency considerations, corroborated by a tentatively identified long repetition; and suppose also that the enemy is known to be using a quadricular

[4] A patent has been granted upon a rather ingenious machine for automatically accomplishing true polygraphic substitution, but it has not been placed upon the market. See U. S. Patent No. 1845947 issued in 1932 to Weisner and Hill. In U. S. Patent No. 1515680 issued to Henkels in 1924, there is described a mechanism which also produces polygraphic substitution.

table of 676 cells containing digraphs showing reciprocal equivalence between plain and cipher-text digraphs. Suppose the message begins as follows (in which the assumed values have been inserted):

XQ	VO	ZI	LK	AP	OL	ZX	PV	QN	IK	OL	UK	AL	HN	LK	VL
FO		TH	IN		NT		RE			NT	NO			IN	

BN	OZ	KU	DY	EL	LE	YW
SI		ON	TO			

The words FOURTH INFANTRY REGIMENT are readily recognized. The reciprocal pairs EL_c and LE_c suggest ATTACK. The beginning of the message is now completely disclosed: FOURTH INFANTRY REGIMENT NOT YET IN POSITION TO ATTACK. The values more or less automatically determined are $VO_c=UR_p$, $AL_c=TY_p$, $HN_c=ET_p$, $VL_c=PO_p$, $OZ_c=TI_p$, $YW_c=CK_p$.

c. Once a good start has been made and a few words have been solved, subsequent work is quite simple and straightforward. A knowledge of enemy correspondence, including data regarding its most common words and phrases, is of great assistance in breaking down new digraphic tables of the same nature but with different equivalents.

d. The foregoing remarks also apply to the details of solution in cases of partially digraphic substitution.

44. Analysis of digraphic substitution ciphers based upon 4-square checkerboard designs.—*a.* In Section VIII of *Advanced Military Cryptography* there are shown various examples of digraphic substitution based upon the use of checkerboard designs. These may be considered cases of partially digraphic substitution, in that in the checkerboard system there are certain relationships between plain-text digraphs having common elements and their corresponding cipher-text digraphs, which will also have common elements. For example, take the following 4-square checkerboard design:

	B	W	G	R	M	O	P	A	U	L	
	N	Y	V	X	E	H	Z	Q	D	F	
1	S	I	C	T	K	K	I	T	S	C	3
	U	P	L	A	O	M	W	R	B	G	
	D	Z	F	Q	H	E	Y	X	N	V	
	W	A	L	E	S	C	X	K	P	B	
	F	H	U	I	T	O	M	Y	D	V	
2	P	X	B	K	C	S	A	E	W	L	4
	N	Z	R	Q	G	G	Z	Q	N	R	
	D	M	V	Y	O	T	H	I	F	U	

FIGURE 20.

Here $BC_p=OW_c$, $BO_p=OF_c$, $BS_p=OP_c$, $BG_p=ON_c$ and $BT_p=OD_c$. In each case when B_p is the initial letter of the plain-text pair, the initial letter of the cipher-text equivalent is O_c. This, of course, is the direct result of the method; it means that the encipherment is monoalphabetic for the first half of each of these *five* plain-text pairs, polyalphabetic for the second half. This relationship holds true for *four* other groups of pairs beginning with B_p. In other words, there are five alphabets employed, not 25. Thus, this case differs from the case discussed under Par. 42*b* only in that the monoalphabeticity is not complete for one-half of all the pairs, but only among the members of certain groups of pairs. In a completely digraphic system using a 676-cell randomized square, such relationships are entirely absent and for this reason the system is cryptographically more secure than the checkerboard system.

b. From the foregoing, it is clear that when solution has progressed sufficiently to disclose a few values, the insertion of letters within the cells of the checkerboard design to give the plain-text and cipher relationships indicated by the solved values immediately leads to the disclosure of additional values. Thus, the solution of only a few values soons leads to the breakdown of the entire checkerboard design.

c. (1) The following example will serve to illustrate the procedure. Let the message be as follows:

```
    1 2 3 4 5   6 7 8 9 10  11 12 13 14 15  16 17 18 19 20  21 22 23 24 25  26 27 28 29 30
A.  H F C A P   G O Q I L   B S P K M       N D U K E       O H Q N F       B O R U N
B.  Q C L C H   Q B Q B F   H M A F X       S I O K O       Q Y F N S       X M C G Y
C.  X I F B E   X A F D X   L P M X H       H R G K G       Q K Q M L       F E Q Q I
D.  G O I H M   U E O R D   C L T U F       E Q Q C G       Q N H F X       I F B E X
E.  F L B U Q   F C H Q O   Q M A F T       X S Y C B       E P F N B       S P K N U
F.  Q I T X E   U Q M L F   E Q Q I G       O I E U E       H P I A N       Y T F L B
G.  F E E P I   D H P C G   N Q I H B       F H M H F       X C K U P       D G Q P N
H.  C B C Q L   Q P N F N   P N I T O       R T E N C       O B C N T       F H H A Y
I.  Z L Q C I   A A I Q U   C H T P C       B I F G W       K F C Q S       L Q M C B
J.  O Y C R Q   Q D P R X   F N Q M L       F I D G C       C G I O G       O I H H F
K.  I R C G G   G N D L N   O Z T F G       E E R R P       I F H O T       F H H A Y
L.  Z L Q C I   A A I Q U   C H T P
```

(2) The cipher having been tested for standard alphabets (by the method of completing the normal components) and found to give negative results, a uniliteral-frequency distribution is made. It is as follows:

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
11	15	26	8	16	30	17	22	24	0	8	14	11	18	15	16	33	9	6	11	11	0	1	12	7	3

FIGURE 21.

(3) At first glance this may appear to the untrained eye to be a monoalphabetic frequency distribution but upon closer inspection it is noted that aside from the frequencies of four or five letters the frequencies for the remaining letters are not very dissimilar. There are, in reality, no very marked crests and troughs, certainly not as many as would be expected in a monoalphabetic substitution cipher of equal length.

(4) The message having been carefully examined for repetitions of 4 or more letters, all of them are listed:

	Frequency	Located in lines
TFHHAYZLQCIAAIQUCHTP (20 letters)	2	H and K.
QMLFEQQIGOI (11 letters)	2	C and F.
XIFBEX (6 letters)	2	C and D.
FEQQ	3	C, D, F.
QMLF	3	C, F, J.
BFHM	2	B and G.
BSPK	2	A and E.
GOIH	2	D and J.

Since there are quite a few repetitions, two of considerable length, since all but one of them contain an even number of letters, and since the message also contains an even number of letters, 344, digraphic substitution is suspected. The cryptogram is transcribed in 2-letter groups, for greater convenience in study. It is as follows:

Message transcribed in pairs

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A.	HF	CA	PG	OQ	IL	BS	PK	MN	DU	KE	OH	QN	FB	OR	UN
B.	QC	LC	HQ	BQ	BF	HM	AF	XS	IO	KO	QY	FN	SX	MC	GY
C.	XI	FB	EX	AF	DX	LP	MX	HH	RG	KG	QK	QM	LF	EQ	QI
D.	GO	IH	MU	EO	RD	CL	TU	FE	QQ	CG	QN	HF	XI	FB	EX
E.	FL	BU	QF	CH	QO	QM	AF	TX	SY	CB	EP	FN	BS	PK	NU
F.	QI	TX	EU	QM	LF	EQ	QI	GO	IE	UE	HP	IA	NY	TF	LB
G.	FE	EP	ID	HP	CG	NQ	IH	BF	HM	HF	XC	KU	PD	GQ	PN
H.	CB	CQ	LQ	PN	FN	PN	IT	OR	TE	NC	CB	CN	TF	HH	AY
J.	ZL	QC	IA	AI	QU	CH	TP	CB	IF	GW	KF	CQ	SL	QM	CB
K.	OY	CR	QQ	DP	RX	FN	QM	LF	ID	GC	CG	IO	GO	IH	HF
L.	IR	CG	GG	ND	LN	OZ	TF	GE	ER	RP	IF	HO	TF	HH	AY
M.	ZL	QC	IA	AI	QU	CH	TP								

It is noted that all the repetitions listed above break up properly into digraphs except in one case, viz, FEQQ in lines C, D, and F. This seems rather strange, and at first thought one might suppose that a letter was dropped out or was added in the vicinity of the FEQQ in line D. But it is immediately seen that the FE QQ in line D has no relation at all to the .F EQ Q. in lines C and F, and that the F EQ Q in line D is merely an accidental repetition.

(5) A digraphic frequency distribution [5] is made and is shown in Fig. 22.

	A	B	C	D	E	F	G	H	I	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A						3			2															2	
B						2										1		2		1					
C	1	5					4	3			1		1			2	1								
D															1					1			1		
E														1	2	2	1			1			2		
F		3			2						1		4												
G			1		1		1							3		1						1		1	
H						4		3				2		1	2	1									
I	3			2	1	2		3			1			2			1		1						
K					1	1	1							1						1					
L		1	1			3							1		1	1									
M			1										1							1			1		
N			1	1												1				1				1	
O								1								1	2							1	1
P				1			1			2			3												
Q			3			1			3	1		5	2	1		2				2				1	
R				1			1								1								1		
S											1												1	1	
T					1	4									2					1			2		
U					1								1												
V																									
W																									
X			1						2									1							
Y																									
Z											2														

FIGURE 22.

(6) The appearance of the foregoing distribution for this message is quite characteristic of that for a digraphic substitution cipher. There are many blank cells; although there are many cases in which a digraph appears only once, there are quite a few in which a digraph appears two or three times, four cases in which a digraph appears four times, and two cases in which a digraph appears five times. The absence of the letter J is also noted; this is often the case in a digraphic system based upon a checkerboard design.

[5] The distinction between "digraphic" and "biliteral" is based upon the following consideration. In a biliteral (or diliteral) distribution every two successive letters of the text would be grouped together to form a pair. For example, a biliteral distribution of ABCDEF would tabulate the pairs AB, BC, CD, DE, and EF. In a digraphic distribution only successive pairs of the text are tabulated. For example, ABCDEF would yield only AB, CD, and EF.

(7) In another common type of checkerboard system known as the Playfair cipher, described in Par. 46, one of the telltale indications besides the absence of the letter J is the absence of double letters, that is, two successive identical letters. The occurrence of the double letters GG, HH, and QQ in the message under investigation eliminates the possibility of its being a Playfair cipher. The simplest thing to assume is that a 4-square checkerboard is involved. One with normal alphabets in Sections 1 and 2 is therefore set down (Fig. 23*a*).

	A	B	C	D	E						
	F	G	H	I–J	K						
1	L	M	N	O	P						3
	Q	R	S	T	U						
	V	W	X	Y	Z						
						A	B	C	D	E	
						F	G	H	I–J	K	
4						L	M	N	O	P	2
						Q	R	S	T	U	
						V	W	X	Y	Z	

FIGURE 23*a*.

(8) The recurrence of the group QMLF, three times, and at intervals suggesting that it might be a sentence separator, leads to the assumption that it is the word STOP. The letters Q, M, L, and F are therefore inserted in the appropriate cells in Sections 3 and 4 of the diagram. Thus (Fig. 23*b*):

	A	B	C	D	E						
	F	G	H	I–J	K						
1	L	M	N	O	P					L	3
	Q	R	S	T	U				Q		
	V	W	X	Y	Z						
						A	B	C	D	E	
						F	G	H	I–J	K	
4				F		L	M	N	O	P	2
			M			Q	R	S	T	U	
						V	W	X	Y	Z	

FIGURE 23*b*.

These placements seem logical. Moreover, in Section 3 the number of cells between L and Q is just one less than enough to contain all the letters M to P, inclusive, and suggests that either N or O is in the keyword portion of the sequence, that is, near the top of Section 3. Without making a commitment in the matter, suppose both N and O, for the present, be inserted in the cell between M and P. Thus (Fig. 23c):

	A	B	C	D	E						
	F	G	H	I–J	K						
1	L	M	N	O	P					L	3
	Q	R	S	T	U	M	N/O	P	Q		
	V	W	X	Y	Z						
						A	B	C	D	E	
						F	G	H	I–J	K	
4				F		L	M	N	O	P	2
			M			Q	R	S	T	U	
						V	W	X	Y	Z	

FIGURE 23c.

(9) Now, if the placement of P in Section 3 is correct, the cipher equivalent of TH_p will be $P\Theta_c$, and there should be a group of adequate frequency to correspond. Noting that PN_c occurs three times, it is assumed to be TH_p and the letter N is inserted in the appropriate cell in Section 4. Thus (Fig. 23d):

	A	B	C	D	E						
	F	G	H	I–J	K						
1	L	M	N	O	P					L	3
	Q	R	S	T	U	M	N/O	P	Q		
	V	W	X	Y	Z						
						A	B	C	D	E	
				N		F	G	H	I–J	K	
4				F		L	M	N	O	P	2
			M			Q	R	S	T	U	
						V	W	X	Y	Z	

FIGURE 23d.

(10) It is about time to try out these assumed values in the message. The proper insertions are made, with the following results:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A.	HF	CA	PG	OQ	IL	BS	PK	MN	DU	KE	OH	QN	FB	OR	UN
B.	QC	LC	HQ	BQ	BF	HM	AF	XS	IO	KO	QY	FN	SX	MC	GY
C.	XI	FB	EX	AF	DX	LP	MX	HH	RG	KG	QK	QM ST	LF OP	EQ	QI
D.	GO	IH	MU	EO	RD	CL	TU	FE	QQ	CG	QN	HF	XI	FB	EX
E.	FL	BU	QF	CH	QO	QM ST	AF	TX	SY	CB	EP	FN	BS	PK	NU
F.	QI	TX	EU	QM ST	LF OP	EQ	QI	GO	IE	UE	HP	IA	NY	TF	LB
G.	FE	EP	ID	HP	CG	NQ	IH	BF	HM	HF	XC	KU	PD	GQ	PN TH
H.	CB	CQ	LQ	PN TH	FN	PN TH	IT	OR	TE	NC	CB	CN	TF	HH	AY
J.	ZL	QC	IA	AI	QU	CH	TP	CB	IF	GW	KF	CQ	SL	QM ST	CB
K.	OY	CR	QQ	DP	RX	FN	QM ST	LF OP	ID	GC	CG	IO	GO	IH	HF
L.	IR	CG	GG	ND	LN	OZ	TF	GE	ER	RP	IF	HO	TF	HH	AY
M.	ZL	QC	IA	AI	QU	CH	TP								

(11) So far no impossible combinations are in evidence. Beginning with group H4 in the message is seen the following sequence:

P N F N P N
T H . . T H

Assume it to be THAT THE. Then $AT_p=FN_c$, and the letter N is to be inserted in row 4 column 1. But this is inconsistent with previous assumptions, since N in Section 4 has already been tentatively placed in row 2 column 4 of Section 4. Other assumptions for FN_c are made: that it is, IS_p (THIS TH. . .); that it is EN_p (THEN TH. . .); but the same inconsistency is apparent. In fact the student will see that FN_c must represent a digraph ending in F, G, H, I–J, or K, since N_c is tentatively located on the same line as these letters in Section 2. Now FN_c occurs 4 times in the message. The digraph it represents *must* be one of the following:

DF, DG, DH, DI, DJ, DK
IF, IG, IH, II, IJ, IK
JF, JG, JH, JI, JJ, JK
OF, OG, OH, OI, OJ, OK
TK,
YF, YG, YH, YI, YJ, YK

Of these the only one likely to be repeated 4 times is OF, yielding T H O F T H which may be
P N F N P N
a part of

. N O R T H O F T H E .   or   . S O U T H O F T H E .
C Q L Q P N F N P N I T   C Q L Q P N F N P N I T

In either case, the position of the F in Section 3 is excellent: F . . . L in row 3. There are 3 cells intervening between F and L, into which G, H, I–J, and K may be inserted. It is not nearly so likely that G, H, and K are in the keyword as that I should be in it. Let it be assumed that this is the case, and let the letters be placed in the appropriate cells in Section 3. Thus (Fig. 23*e*):

	A	B	C	D	E						
	F	G	H	I–J	K						
1	L	M	N	O	P	F	G	H	K	L	3
	Q	R	S	T	U	M	N/O	P	Q		
	V	W	X	Y	Z						
						A	B	C	D	E	
				N		F	G	H	I–J	K	
4				F		L	M	N	O	P	2
			M	Q		Q	R	S	T	U	
						V	W	X	Y	Z	

FIGURE 23*e*.

Let the resultant derived values be checked against the frequency distribution. If the position of H in Section 3 is correct, then the digraph ON_p, normally of high frequency should be represented several times by HF_c. Reference to Fig. 22 shows a frequency of 4 times. And HM_c, with 2 occurrences, represents NS_p. There is no need to go through all the possible corroborations.

(12) Going back to the assumption that T H . . T H
P N F N P N
is part of the expression

. N O R T H O F T H E .   or   . S O U T H O F T H E . .
C Q L Q P N F N P N I T   C Q L Q P N F N P N I T

it is seen at once from Fig. 23*e* that the latter is apparently correct and not the former, because LQ_c equals OU_p and not OR_p. If $OS_p=CQ_c$, this means that the letter C of the digraph CQ_c must be placed in row 1 column 3 or row 2 column 3 of Section 3. Now the digraph CB_c occurs 5 times, CG_c, 4 times, CH_c, 3 times, CQ_c, 2 times. Let an attempt be made to deduce the exact position of C in Section 3 and the positions of B, G, and H in Section 4. Since F is already placed in Section

4, assume G and H directly follow it, and that B comes before it. How much before? Suppose a trial be made. Thus (Fig. 23*f*):

	A	B	C	D	E			C			
	F	G	H	I–J	K			C			
1	L	M	N	O	P	F	G	H	K	L	3
	Q	R	S	T	U	M	N/O	P	Q		
	V	W	X	Y	Z						
						A	B	C	D	E	
				N		F	G	H	I–J	K	
4	B	B	B	F	G	L	M	N	O	P	2
	H		M	Q		Q	R	S	T	U	
						V	W	X	Y	Z	

FIGURE 23*f*.

By referring now to the frequency distribution, Fig. 22, after a very few minutes of experimentation it becomes apparent that the following is correct:

	A	B	C	D	E			C			
	F	G	H	I–J	K						
1	L	M	N	O	P	F	G	H	K	L	3
	Q	R	S	T	U	M	N/O	P	Q		
	V	W	X	Y	Z						
						A	B	C	D	E	
				N		F	G	H	I–J	K	
4	B			F	G	L	M	N	O	P	2
	H		M	Q		Q	R	S	T	U	
						V	W	X	Y	Z	

FIGURE 23*g*.

(13) The identifications given by these placements are inserted in the text, and solution is very rapidly completed. The final checkerboard and deciphered text are given below.

	A	B	C	D	E	S	O	C	I	E	
	F	G	H	I–J	K	T	Y	A	B	D	
1	L	M	N	O	P	F	G	H	K	L	3
	Q	R	S	T	U	M	N	P	Q	R	
	V	W	X	Y	Z	U	V	W	X	Z	
	E	X	P	U	L	A	B	C	D	E	
	S	I	O	N	A	F	G	H	I–J	K	
	B	C	D	F	G	L	M	N	O	P	
4	H	K	M	Q	R	Q	R	S	T	U	2
	T	V	W	Y	Z	V	W	X	Y	Z	

FIGURE 23A.

A. H F C A P G O Q I L B S P K M N D U K E O H Q N F B O R U N
O N E H U N D R E D F I R S T F I E L D A R T I L L E R Y F

B. Q C L C H Q B Q B F H M A F X S I O K O Q Y F N S X M C G Y
R O M P O S I T I O N S I N V I C I N I T Y O F B A R L O W

C. X I F B E X A F D X L P M X H H R G K G Q K Q M L F E Q Q I
W I L L B E I N G E N E R A L S U P P O R T S T O P D U R I

D. G O I H M U E O R D C L T U F E Q Q C G Q N H F X I F B E X
N G A T T A C K S P E C I A L A T T E N T I O N W I L L B E

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

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

G. F E E P I D H P C G N Q I H B F H M H F X C K U P D G Q P N
L A C E C O N C E N T R A T I O N S O N W O O D S N O R T H

H. C B C Q L Q P N F N P N I T O R T E N C C B C N T F H H A Y
A N D S O U T H O F T H A Y E R F A R M A N D H I L L S I X

J. Z L Q C I A A I Q U C H T P C B I F G W K F C Q S L Q M C B
Z E R O E I G H T D A S H A A N D O N W O O D S E A S T A N

K. O Y C R Q Q D P R X F N Q M L F I D G C C G I O G O I H H F
D W E S T T H E R E O F S T O P C O M M E N C I N G A T O N

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

M. Z L Q C I A A I Q U C H T P
Z E R O E I G H T D A S H A

d. (1) It is interesting to note how much simpler the matter becomes when the positions of the plain-text and cipher-text sections are reversed, or, what amounts to the same thing, when in encipherment the plain-text pairs are sought in the sections containing the mixed alphabets, and their cipher equivalents are taken from the sections containing the normal alphabets. For example, referring to Fig. 23*h*, suppose that sections 3–4 be used as the source of the plain-text pairs, and sections 1–2 as the source of the cipher-text pairs. Then $ON_p=DG_c$, $EH_p=AU_c$, etc.

(2) To solve a message enciphered in that manner, it is necessary merely to make a square in which all four sections are normal alphabets, and then perform two steps. First, the cipher text pairs are converted into their normal alphabet equivalents merely by "deciphering" the message with that square; the result of this operation yields two monoalphabets, one composed of the odd letters, the other of the even letters. The second step is to solve these two mono-alphabets.

(3) Where the same mixed alphabet is inserted in sections 3 and 4, the problem is still easier, since the letters resulting from the conversion into normal-alphabet equivalents all belong to the same, single-mixed alphabet.

45. Analysis of ciphers based upon other types of checkerboard designs.—The solution of cryptograms enciphered by other types of checkerboard designs is accomplished along lines very similar to those set forth in the foregoing example of the solution of a message prepared by means of a 4-square checkerboard design. There are, unfortunately, no means or tests which can be applied to determine in the early stages of the analysis exactly what type of design is involved in the *first* case under study. The author freely admits that the solution outlined in subparagraph *c* is quite artificial in that nothing is demonstrated in step (7) that obviously leads to or warrants the assumption that a 4-square checkerboard is involved. This point was passed over with the quite bald statement that this was "the simplest thing to assume"—and then the solution proceeds exactly as though this mere *hypothesis* has been definitely established. For example, the very first results obtained were based upon assuming that a certain 4-letter repetition represented the word STOP and *immediately inserting certain letters in appropriate cells in a 4-square checkerboard.* Several more assumptions were built on top of that and very rapid strides were made. What if it had not been a 4-square checkerboard at all? What if it had been a 2-square checkerboard of the type shown in Fig. 24?

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

FIGURE 24.

The only defense that can be made of what may seem to the student to be purely arbitrary procedure based upon the author's advance information or knowledge is the following: In the first place, in order to avoid making the explanation a too-long-drawn-out affair, it is necessary (and pedagogical experience warrants) that certain alternative hypotheses be passed over in silence. In the second place, it may now be added, *after* the principles and procedure have been elucidated (which at this stage is the primary object of this text) that if good results do not follow from a first hypothesis, the only thing the cryptanalyst can do is to reject that hypothesis, and formulate a second hypothesis. In actual practice he may have to reject a second, third, fourth, . . . *nth* hypothesis. In the end he may strike the right one—or he may not. There is no guaranty of success in the matter. In the third place, one of the objects of this text is to show how certain systems, if employed for military purposes, can readily be broken down. Assuming that a checkerboard system is in use, and that daily changes in keywords are made, it is possible that the traffic of the first day might give considerable difficulty in solution, if the type of checkerboard were not known to the cryptanalyst. But the second or third day's traffic would be easy to solve, because by that time the cryptanalytic personnel would have analyzed the system and thus learned what type of checkerboard the enemy is using.

46. Analysis of the Playfair cipher system.—*a.* An excellent example of a practical, partially digraphic system is the Playfair cipher.[6] It was used for a number of years as a field cipher by the British Army, before and during the World War, and for a short time, also during that war, by certain units of the American Expeditionary Forces.

b. Published solutions[7] for this cipher are quite similar basically and vary only in minor details. The earliest, that by Lieut. Mauborgne, used straightforward principles of frequency to establish the values of three or four of the most frequent digraphs. Then, on the assumption that in most cases in which a keyword appears on the first and second rows the last five letters of the normal alphabet, VWXYZ, will rarely be disturbed in sequence and will occupy the last row of the square, he "juggles" the letters given by the values tentatively established from frequency considerations, placing them in various positions in the square, together with VWXYZ, to correspond to the plain-text cipher relationships tentatively established. A later solution by Lieut. Frank Moorman, as described in Hitt's Manual, assumes that in a Playfair cipher prepared by means of a square in which the keyword occupies the first and second rows, if a digraphic frequency distribution is made, it will be found that the letters having the greatest combining power are very probably letters of the key. A still later solution, by Lieut. Commander Smith, is perhaps the most lucid and systematized of the three. He sets forth in definite language certain considerations which the other two writers certainly entertained but failed to indicate.

c. The following details have been summarized from Commander Smith's solution:

(1) The Playfair cipher may be recognized by virtue of the fact that it always contains an even number of letters, and that when divided into groups of two letters each, no group contains a repetition of the same letter, as NN or EE. Repetitions of digraphs, trigraphs, and polygraphs will be evident in fairly long messages.

(2) Using the square[8] shown in Fig. 25*a*, there are two general cases to be considered, as regards the results of encipherment:

B	A	N	K	R
D	E	F	G	H
I–J	L	M	O	Q
U	P	T	C	Y
S	V	W	X	Z

FIGURE 25*a*.

[6] This cipher was really invented by Sir Charles Wheatstone but receives its name from Lord Playfair, who apparently was its sponsor before the British Foreign Office. See Wemyss Reid, *Memoirs of Lyon Playfair*, London, 1899. A detailed description of this cipher will be found in Sec. VIII, *Advanced Military Cryptography*.

[7] Mauborgne, Lieut. J. O., U. S. A. *An advanced problem in cryptography and its solution*, Leavenworth, 1914.
Hitt, Captain Parker, U. S. A. *Manual for the solution of military ciphers*, Leavenworth, 1918.
Smith, Lieut. Commander W. W., U. S. N. In *Cryptography* by André Langie, translated by J. C. H. Macbeth, New York, 1922.

[8] The Playfair square accompanying Commander Smith's solution is based upon the keyword BANKRUPTCY, "to be distributed between the first and fourth lines of the square." This is a simple departure from the original Playfair scheme in which the letters of the keyword are written from left to right and in consecutive lines from the top downward.

CASE 1. Letters at opposite corners of a rectangle. The following illustrative relationships are found:

$$TH_p = YF_c$$
$$HT_p = FY_c$$
$$YF_p = TH_c$$
$$FY_p = HT_c$$

Reciprocity is complete.

CASE 2. Two letters in the same line or column. The following illustrative relationships are found:

$$AN_p = NK_c$$
$$NA_p = KN_c$$

But NK_p does not $= AN_c$, nor does $KN_p = NA_c$.

Reciprocity is only partial.

(3) The foregoing gives rise to the following:

RULE I. (*a*) Regardless of the position of the letters in the square, if

$$1.2 = 3.4, \text{ then}$$
$$2.1 = 4.3$$

(*b*) If 1 and 2 form opposite corners of a rectangle, the following equations obtain:

$$1.2 = 3.4$$
$$2.1 = 4.3$$
$$3.4 = 1.2$$
$$4.3 = 2.1$$

(4) A letter considered as occupying a position in a row can be combined with but four other letters in the same row; the same letter considered as occupying a position in a column can be combined with but four other letters in the same column. Thus, this letter can be combined with only 8 other letters all told, under Case 2, above. But the same letter considered as occupying a corner of a rectangle can be combined with 16 other letters, under Case 1, above. Commander Smith derives from these facts the conclusion that "it would appear that Case 1 is twice as probable as Case 2." He continues thus (notation my own):

"Now in the square, note that:

$AN_p = NK_c$		$EN_p = FA_c$
$GN_p = FK_c$		$EM_p = FL_c$
$ON_p = MK_c$	also	$ET_p = FP_c$
$CN_p = TK_c$		$EW_p = FV_c$
$XN_p = WK_c$		$EF_p = FG_c$

"From this it is seen that of the 24 equations that can be formed when each letter of the square is employed either as the initial or final letter of the group, five will indicate a repetition of a corresponding letter of plain text.

"Hence, RULE II. After it has been determined, in the equation $1.2 = 3.4$, that, say, $EN_p = FA_c$, there is a probability of one in five that any other group beginning with F_c indicates $E\Theta_p$, and that any group ending in A_c indicates ΘN_p.

"After such combinations as ER_p, OR_p, and EN_p have been assumed or determined, the above rule may be of use in discovering additional digraphs and partial words." [9]

Rule III. In the equation 1.2=3.4, 1 and 3 can never be identical, nor can 2 and 4 ever be identical. Thus, AN_p could not possibly be represented by AY_c, nor could ER_p be represented by KR_c. This rule is useful in elimination of certain possibilities when a specific message is being studied.

Rule IV. In the equation $1.2_p=3.4_c$, if 2 and 3 are identical, the letters are all in the same row or column, and in the relative order 124. In the square shown, $AN_p=NK_c$ and the absolute order is ANK. The relative order 124 includes five absolute orders which are cyclic permutations of one another. Thus: ANK.., NK..A, K..AN, ..ANK, and .ANK..

Rule V. In the equation $1.2_p=3.4_c$, if 1 and 4 are identical, the letters are all in the same row or column, and in the relative order 243. In the square shown, $KN_p=RK_c$ and the absolute order is NKR. The relative order 243 includes five absolute orders which are cyclic permutations of one another. Thus NKR.., KR..N, R..NK, ..NKR, and .NKR..

Rule VI. "Analyze the message for group recurrences. Select the groups of greatest recurrence and assume them to be high-frequency digraphs. Substitute the assumed digraphs throughout the message, testing the assumptions in their relation to other groups of the cipher. The reconstruction of the square proceeds simultaneously with the solution of the message and aids in hastening the translation of the cipher."

d. (1) When solutions for the Playfair cipher system were first developed, based upon the fact that the letters were inserted in the cells in keyword-mixed order, cryptographers thought it desirable to place stumbling blocks in the path of such solution by departing from strict, keyword-mixed order. Playfair squares of the latter type are designed as "modified Playfair squares." One of the simplest methods is illustrated in Fig. 25*a*, wherein it will be noted that the last five letters of the keyword proper are inserted in the fourth row of the square instead of the second, where they would naturally fall. Another method is to insert the letters within the cells from left to right and top downward but use a sequence that is a keyword-mixed sequence developed by a columnar transposition based upon the keyword proper. Thus, using the keyword BANKRUPTCY:

```
2 1 5 4 7 9 6 8 3 10
B A N K R U P T C Y
D E F G H I L M O Q
S V W X Z
```

Sequence: A E V B D S C O K G X N F W P L R H Z T M U I Y Q

[9] There is an error in this reasoning. Take, for example, the 24 equations having F as an initial letter:

Case		Case		Case		Case	
1.	$FB_c=DN_p$	2.	FE=ED	2.	FT=NM	1.	FX=GW
2.	FD =EH	1.	FL=EM	2.	FW=NT	1.	FR=HN
1.	FI =DM	1.	FP=ET	1.	FK=GN	2.	FH=EG
1.	FU =DT	1.	FV=EW	2.	FG=EF	1.	FQ=HM
1.	FS =DW	2.	FN=NW	1.	FO=GM	1.	FY=HT
1.	FA =EN	2.	FM=NF	1.	FC=GT	1.	FZ=HW

Here, the initial letter F_c represents the following initial letters of plain-text digraphs:

D E N G H

It is seen that F_c represents D_p, N_p, G_p, H_p 4 times each, and E_p, 8 times. Consequently, supposing that it has been determined that $FA_c=EN_p$, the probability that F_c will represent E_p is not 1 in 5 but 8 in 24, or 1 in 3; but supposing that it has been determined that $FW_c=NT_p$, the probability that F_c will represent N_p is 4 in 24 or 1 in 6. The difference in these probabilities is occasioned by the fact that the first instance, $FA_c=EN_p$ corresponds to a Case 1 encipherment, the second instance, $FW_c=NT_p$, to a Case 2 encipherment. But there is no way of knowing initially, and without other data, whether one is dealing with a Case 1 or Case 2 encipherment. Only as an approximation, therefore, may one say that the probability of F_c representing a given Θ_p is 1 in 5.

The Playfair Square is as follows:

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

Figure 25*b*.

(2) In the foregoing square practically all indications that the square has been developed from a keyword have disappeared. The principal disadvantage of such an arrangement is that it requires more time to locate the letters desired, both in cryptographing and decryptographing, than it usually does when a semblance of normal alphabetic order is preserved in the square.

(3) Note the following three squares:

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

Figure 25*c*.

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

Figure 25*d*.

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

Figure 25*e*.

At first glance they all appear to be different, but closer examination shows them to be *cyclic permutations* of one another and of the square in Fig. 25*b*. They yield identical equivalents in all cases. However, if an attempt be made to reconstruct the original keyword, it would be much easier to do so from Fig. 25*b* than from any of the others, because in Fig. 25*b* the keyword-mixed sequence has not been disturbed as much as in Figs. 25*c*, *d*, *e*. In working with Playfair ciphers, the student should be on the lookout for such instances of cyclic permutation of the original Playfair square, for during the course of solution he will not know whether he is building up the original or an equivalent cyclic permutation of the original square; only after he has completely reconstructed the square will he be able to determine this point.

(4) It can readily be shown that the columns of a Playfair square may be cyclically permuted (see subpar. *d*) to produce a first set of 25 squares all of which, though at first glance apparently different, will yield identical equivalents; likewise, the rows of such a square may be cyclically permuted to produce a second set of 25 squares all of which will also yield identical equivalents. Thus there may be a total of 50 cyclic permutations composed of two sets of 25 each. The cipher equivalents yielded by Case 2 encipherments (letters in the same row or in the same column) will be identical for any two of these 50 different Playfair squares; but the cipher equivalents yielded by Case 1 encipherments (letters at diagonally opposite corners of a rectangle) will be identical only for two squares belonging to the same set of 25 cyclic permutations.

e. (1) The steps in the solution of a typical example of this cipher may be useful. Let the message be as follows:

```
    1 2 3 4 5   6 7 8 9 10  11 12 13 14 15  16 17 18 19 20  21 22 23 24 25  26 27 28 29 30
A.  V T Q E U   H I O F T   C H X S C       A K T V T       R A Z E V       T A G A E
B.  O X T Y M   H C R L Z   Z T Q T D       U M C Y C       X C T G M       T Y C Z U
C.  S N O P D   G X V X S   C A K T V       T P K P U       T Z P T W       Z F N B G
D.  P T R K X   I X B P R   Z O E P U       T O L Z E       K T T C S       N H C Q M
E.  V T R K M   W C F Z U   B H T V Y       A B G I P       R Z K P C       Q F N L V
F.  O X O T U   Z F A C X   X C P Z X       H C Y N O       T Y O L G       X X I I H
G.  T M S M X   C P T O T   C X O T T       C Y A T E       X H F A C       X X C P Z
H.  X H Y C T   X W L Z T   S G P Z T       V Y W C E       T W G C C       M B H M Q
J.  Y X Z P W   G R T I V   U X P U M       Q R K M W       C X T M R       S W G H B
K.  X C P T O   T C X O T   M I P Y D       N F G K I       T C O L X       U E T P X
L.  X F S R S   U Z T D B   H O Z I G       X R K I X       Z P P V Z       I D U H Q
M.  O T K T K   C C H X X
```


(2) Without going through the preliminary tests in detail, with which it will be assumed that the student is now familiar,[10] the conclusion is reached that the cryptogram is digraphic in nature, and a digraphic frequency distribution is made (Fig. 26).

	A	B	C	D	E	F	G	H	I	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A					1		1																		
B							2	2																	
C	2				1	1		2				1				1	1						5	1	
D		1					1						1							2					
E															1				1						
F	2						1						2						1						
G			1																				2		
H		1	1											1		1									
I								1						1	1						1		1		
K			1						1						1				4						
L																					1				1
M			1					1	1							2			1			2			
N														1											
O											3				1				6				2		
P										1							1		4	2	1		1	1	3
Q					1							1							1						
R	1									4								1	1						1
S							1					1	2				1			1					
T			3		1		1					2									2	1	1	2	1
U								1											1				1		1
V																			5						
W							2				1														1
X		1	5			1		3	2									2		1	1				
Y	2		3																			1	1		
Z					2				2					1	2				3	2					

FIGURE 26.

[10] See Par. 44*c*.

(7) It is perhaps high time that the whole list of tentative equivalent values be studied in relation to their consistency with the positions of letters in the Playfair square; moreover, by so doing, additional values may be obtained in the process. The complete list of values is as follows:

Assumed values	*Derived by Rule I*
$AT_p=CX_c$	$TA_p=XC_c$
$LI_p=PZ_c$	$IL_p=ZP_c$
$ON_p=XH_c$	$NO_p=HX_c$
$TH_p=OT_c$	$HT_p=TO_c$
$IR_p=UZ_c$	$RI_p=ZU_c$
$DB_p=FA_c$	$BD_p=AF_c$
$EC_p=TE_c$	$CE_p=ET_c$
$TE_p=PT_c$	$ET_p=TP_c$
$EI_p=TC_c$	$IE_p=CT_c$
$RS_p=YA_c$	$SR_p=AY_c$
$-S_p=SM_c$	$S-_p=MS_c$

(8) By Rule V, the equation $TH_p=OT_c$ means that H, T, and O are all in the same row or column and in the relative order 2–4–3; similarly, C, E, and T are in the same row or column and in the relative order 243. Further E, P, and T are in the same row and column, and their relative order is also 243. That is, these sequences must occur in the square:

(1)	(2)	(3)
H T O . . , or	C E T . . , or	E T P . . , or
T O . . H , or	E T . . C , or	T P . . E , or
O . . H T , or	T . . C E , or	P . . E T , or
. . H T O , or	. . C E T , or	. . E T P , or
. H T O .	. C E T .	. E T P .

(9) Noting the common letters E and T in the second and third sets of relative orders, these may be combined into one sequence of four letters. Only one position remains to be filled and noting, in the list of equivalents that $EI_p=TC_c$, it is obvious that the letter I belongs to the CET sequence. The complete sequence is therefore as follows:

C E T P I , or
E T P I C , or
T P I C E , or
P I C E T , or
I C E T P

(10) Taking up the HTO sequence, it is noted, in the list of equivalents that $ON_p=XH_c$, an equation containing two of the three letters of the HTO sequence. From this it follows that N and X must belong to the same row or column as HTO. The arrangement must be one of the following:

H T O X N
T O X N H
O X N H T
X N H T O
N H T O X

(11) Since the sequence containing HTOXN has a common letter (T) with the sequence CETPI, it follows that if the HTOXN sequence occupies a row, then the CETPI sequence must occupy a column; or, if the HTO sequence occupies a column, then the CETPI sequence must

occupy a row; and they may be combined by means of their common letter, T. According to subpar. *d* (4), the two sequences may be inserted within a Playfair square in 25 different ways by cyclically permuting and shifting the letters of one of these two sequences; and the same two sequences may be again inserted in another set of 25 ways by cyclically permuting and shifting the letters of the other of these two sequences. In Fig. 27 the diagrams labeled (1) to (10), inclusive, show 10 of the possible 25 obtainable by making the HTOXN sequence one of the rows of the square; diagrams (11) and (12) show 2 of the possible 25 obtainable by making the HTOXN sequence one of the columns of the square. The entire complement of 25 arrangements for each set may easily be drawn up by the student; space forbids their being completely set forth and it is really unnecessary to do so.

(1)

C				
E				
T	O	X	N	H
P				
I				

(2)

	C			
	E			
H	T	O	X	N
	P			
	I			

(3)

		C		
		E		
N	H	T	O	X
		P		
		I		

(4)

			C	
			E	
X	N	H	T	O
			P	
			I	

(5)

				C
				E
O	X	N	H	T
				P
				I

(6)

P				
I				
C				
E				
T	O	X	N	H

(7)

	P			
	I			
	C			
	E			
H	T	O	X	N

(8)

		P		
		I		
		C		
		E		
N	H	T	O	X

(9)

			P	
			I	
			C	
			E	
X	N	H	T	O

(10)

				P
				I
				C
				E
O	X	N	H	T

(11)

				N
				H
P	I	C	E	T
				O
				X

(12)

N				
H				
T	P	I	C	E
O				
X				

FIGURE 27.

(12) Before trying to discover means whereby the actual or absolute arrangement may be detected from among the full set of 50 possible arrangements, the question may be raised: is it necessary? So far as concerns Case 2 encipherments, since any one of the 50 arrangements will yield the same equivalents as any of the remaining 49, perhaps a relative arrangement will do.

(13) Let arrangement 8 be arbitrarily selected for trial.

		P		
		I		
		C		
		E		
N	H	T	O	X

FIGURE 28a.

(14) What additional letters can be inserted, using as a guide the list of equivalents in subparagraph (7)? There is $AT_p = CX_c$, for example. It contains only one letter, A, not in the arrangement selected for trial, and this letter may immediately be placed, as shown:[10]

		P		
		I		
		C		A
		E		
N	H	T	O	X

FIGURE 28b.

Scanning the list for additional cases of this type, none are found. But seeing that several high-frequency letters have already been inserted in the square, perhaps reference to the cryptogram itself in connection with values derived from these inserted letters may yield further clues. For example, the vowels A, E, I, and O are all in position, as are the very frequent consonants N and T. The following combinations may be studied:

$AN_p = \Theta X_c$	$AT_p = CX_c$	$NA_p = X\Theta_c$	$TA_p = XC_c$
$EN_p = \Theta T_c$	$ET_p = TP_c$	$NE_p = T\Theta_c$	$TE_p = PT_c$
$IN_p = \Theta T_c$	$IT_p = CP_c$	$NI_p = T\Theta_c$	$TI_p = PC_c$
$ON_p = XH_c$	$OT_p = XO_c$	$NO_p = HX_c$	$TO_p = OX_c$

$AT_p (= CX_c)$, $TA_p (= XC_c)$, $ON_p (= XH_c)$, $TE_p (= PT_c)$ and $ET_p (= TP_c)$ have already been inserted in the text. Of the others, only $OX_c (= TO_p)$ occurs two times, and this value can be at once inserted in the text. But can the equivalents of AN, EN, or IN be found from frequency considerations?

[10] The fact that the placement of A yields $AT_p = CX_c$ means that the outline selected for experiment really belongs to the correct set of 25 possible cyclic permutations, and that the letters of the NHTOX sequence belong in a row, the letters of the PICET sequence belong in a column of the original Playfair square. If the reverse were the case, one could not obtain $AT_p = CX_c$ but would obtain $AT_p = XC_c$.

Take EN_p, for example; it is represented by ΘT_c. What combination of ΘT is most likely to represent EN_p among the following candidates:

KT_c (4 times); by Rule I, NE_p would $= TK_c$ (no occurrences)

VT_c (5 times); by Rule I, NE_p would $= TV_c$ (2 times)

ZT_c (3 times); by Rule I, NE_p would $= TZ_c$ (1 time)

VT_c certainly looks good: it begins the message, suggesting the word ENEMY; in line H, in the sequence PZTV would become LINE. Let this be assumed to be correct, and let the word ENEMY also be assumed to be correct. Then $EM_p = QE_c$ and the square then becomes as shown herewith:

		P		
		I		
		C		A
V	M	E	Q	
N	H	T	O	X

FIGURE 28c.

(15) In line E is seen the following sequence:

Line E: . . . . VT RK MW CF ZU BH TV YA BG IP RZ KP CQ FN LV
EN RI NE RS PT E

The sequence . . . RI . . NERS . . PT . . . suggests PRISONERS CAPTURED, as follows:

MW CF ZU BH TV YA BG IP RZ KP
P RI SO NE RS CA PT UR ED

This gives the following new values: $\Theta P_p = CF_c$; $SO_p = BH_c$; $CA_p = BG_c$; $UR_p = RZ_c$; $ED_p = KP_c$.

The letters B and G can be placed in position at once, since the positions of C and A are already known. The insertion of the letter B immediately permits the placement of the letter S, from the equation $SO_p = BH_c$. Of the remaining equations only $ED_p = KP_c$ can be used. Since E and P are fixed and are in the same column, D and K must be in the same column, and moreover the K must be in the same row as E. There is only one possible position for K, viz, immediately after Q. This automatically fixes the position of D. The square is now as shown herewith:

		P		D
		I		
G	S	C	B	A
V	M	E	Q	K
N	H	T	O	X

FIGURE 28d.

(16) A review of all equations, including the very first ones established, gives the following which may now be used: $DB_p = FA_c$; $RS_p = YA_c$. The first permits the immediate placement of F; the second, by elimination of possible positions, permits the placement of both R and Y. The square is now as shown herewith:

		P	F	D
	Y	I		R
G	S	C	B	A
V	M	E	Q	K
N	H	T	O	X

FIGURE 28e.

Once more a review is made of all remaining thus far unused equations. $LI_p = PZ_c$ now permits the placement of L and Z. $IR_p = UZ_c$ now permits the placement of U, which is confirmed by the equation $UR_p = RZ_c$ from the word CAPTURED.

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

FIGURE 28f.

There is then only one cell vacant, and it must be occupied by the only letter left unplaced, viz, W. Thus the whole square has been reconstructed, and the message can now be decryptographed.

(17) Is the square just reconstructed identical with the original, or is it a cyclic permutation of a keyword-mixed Playfair square of the type illustrated in Fig. 25b? Even though the message can be read with ease, this point is still of interest. Let the sequence be written in five ways, each composed of five partial sequences made by cyclicly permuting each of the horizontal rows of the reconstructed square. Thus:

	Row 1	Row 2	Row 3	Row 4	Row 5
(a)	L W P F D	Z Y I U R	G S C B A	V M E Q K	N H T O X
(b)	W P F D L	Y I U R Z	S C B A G	M E Q K V	H T O X N
(c)	P F D L W	I U R Z Y	C B A G S	E Q K V M	T O X N H
(d)	F D L W P	U R Z Y I	B A G S C	Q K V M E	O X N H T
(e)	D L W P F	R Z Y I U	A G S C B	K V M E Q	X N H T O

By experimenting with these five sequences, in an endeavor to reconstruct a transposition rectangle conformable to a keyword sequence, the last sequence yields the following:

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

By shifting the O from the last position to the first, and rearranging the columns, the following is obtained:

2	5	3	6	1	4	7
C	O	M	P	A	N	Y
B	D	E	F	G	H	I
K	L	Q	R	S	T	U
V	W	X	Z			

The original square must have been this:

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

FIGURE 28g

f. Continued practice in the solution of Playfair ciphers will make the student quite expert in the matter and will enable him to solve shorter and shorter messages.[11] Also, with practice it will become a matter of indifference to him as to whether the letters are inserted in the square with any sort of regularity, such as simple keyword-mixed order, columnar transposed keyword-mixed order, or in a purely random order.

g. It may perhaps seem to the student that the foregoing steps are somewhat too artificial, a bit too "cut and dried" in their accuracy to portray the process of analysis, as it is applied in practice. For example, the critical student may well object to some of the assumptions and the reasoning in step (5) above, in which the words THREE and ONE (1st hypothesis) were rejected in favor of the words THIRD and SECOND (2nd hypothesis). This rested largely upon the rejection of RE_p and ER_p as the equivalents of UZ_c and ZU_c, and the adoption of IR_p and RI_p as their equivalents. Indeed, if the student will examine the final message with a critical eye he will find that while the bit of reasoning in step (5) is perfectly logical, the assumption upon which it is based is in fact wrong, for it happens that in this case ER_p occurs only once and RE_p does not occur at all. Consequently, although most of the reasoning which led to the rejection of the 1st hypothesis and the adoption of the 2nd was logical, it was in fact based upon erroneous assump-

[11] The author once had a student who "specialized" in Playfair ciphers and became so adept that he could solve messages containing as few as 50–60 letters within 30 minutes.

tion. In other words, despite the fact that the assumption was incorrect, a correct deduction was made. *The student should take note that in cryptanalysis situations of this sort are not at all unusual.* Indeed they are to be expected and a few words of explanation at this point may be useful.

h. Cryptanalysis is a science in which deduction, based upon observational data, plays a very large role. But it is also true that in this science most of the deductions usually rest upon assumptions. It is most often the case that the cryptanalyst is forced to make his assumptions upon a quite limited amount of text. It cannot be expected that assumptions based upon statistical generalizations will always hold true when applied to data comparatively very much smaller in quantity than the total data used to derive the generalized rules. Consequently, as regards assumptions made in specific messages, *most of the time* they will be correct, but *occasionally* they will be incorrect. In cryptanalysis it is often found that among the correct deductions there will be cases in which subsequently discovered facts do not bear out the assumptions on which the deduction was based. Indeed, it is sometimes true that if the *facts* had been known *before* the deduction was made, this knowledge would have prevented making the correct deduction. For example, suppose the cryptanalyst had somehow or other divined that the message under consideration contained no RE, only one ER, one IR, and two RI's (as is actually the case). He would certainly not have been able to choose between the words THREE and ONE (1st hypothesis) as against THIRD and SECOND (2d hypothesis). But because he assumes that there should be more ER_p's and RE_p's than IR's and RI's in the message, he deduces that UZ_c cannot be RE_p, rejects the 1st hypothesis and takes the 2d. It later turns out, after the problem has been solved, that the deduction was correct, although the assumption on which it was based (expectation of more frequent appearance of RE_p and ER_p) was, in fact, *not* true in this particular case. The cryptanalyst can only hope that the number of times when his deductions are correct, even though based upon assumptions which later turn out to be erroneous, will abundantly exceed the number of times when his deductions are wrong, even though based upon assumptions which later prove to be correct. If he is lucky, the making of an assumption which is really not true will make no difference in the end and will not delay solution; but if he is specially favored with luck, it may actually help him solve the message—as was the case in this particular example.

i. Another comment of a general nature may be made in connection with this specific example. The student may ask what would have been the procedure in this case if the message had not contained such a tell-tale repetition as the word BATTALION, which formed the point of departure for the solution, or, as it is often said, permitted an "entering wedge" to be driven into the message. The answer to his query is that if the word BATTALION had not been repeated, there would probably have been some other repetition which would have permitted the same sort of attack. If the student is looking for cut and dried, straight-forward, unvarying methods of attack, he should remember that cryptanalysis, while it may be considered a branch of mathematics, is not a science which has many "general solutions" such as are found and expected in mathematics proper. It is inherent in the very nature of cryptanalytics that, *as a rule,* only general principles can be established; their practical application must take advantage of peculiarities and particular situations which are noted in specific messages. This is especially true in a text on the subject. The illustration of a general principle requires a specific example, and the latter must of necessity manifest characteristics which make it different from any other example. The word BATTALION was not purposely repeated in this example in order to make the demonstration of solution easy; "it just happened that way." In another example, some other entering wedge would have been found. The student can be expected to learn only the *general principles* which will enable him to take advantage of the *specific characteristics* manifested in *specific cases.* Here it is desired to illustrate the general principles of solving Playfair ciphers and to point out the fact that entering wedges must and can be found. The specific nature of the entering wedge varies with specific examples.

SECTION X

CONCLUDING REMARKS

47. Special remarks concerning the initial classification of cryptograms.—*a.* The student should by this time have a good conception of the basic nature of monoalphabetic substitution and of the many "changes" which may be rung upon this simple tune. The first step of all, naturally, is to be able to classify a cryptogram properly and place it in either the transposition or the substitution class. The tests for this classification have been given and as a rule the student will encounter no difficulty in this respect.

b. There are, however, certain kinds of cryptograms whose class cannot be determined in the usual manner, as outlined in Par. 13 of this text. First of all there is the type of code message which employs bona-fide dictionary words as code groups.[1] Naturally, a frequency distribution of such a message will approximate that for normal plain text. The appearance of the message, however, gives clear indications of what is involved. The study of such cases will be taken up in its proper place. At the moment it is only necessary to point out that these are *code* messages and not *cipher*, and it is for this reason that in Pars. 12 and 13 the words "cipher" and "cipher messages" are used, the word "cryptogram" being used only where technically correct.

c. Secondly, there come the unusual and borderline cases, including cryptograms whose nature and type can *not* be ascertained from frequency distributions. Here, the cryptograms are technically not ciphers but special forms of disguised secret writings which are rarely susceptible of being classed as transposition or substitution. These include a large share of the cases wherein the cryptographic messages are disguised and carried under an external, innocuous text which is innocent and seemingly without cryptographic content—for instance, in a message wherein specific letters are indicated in a way not open to suspicion under censorship, these letters being intended to constitute the letters of the cryptographic message and the other letters constituting "dummies." Obviously, no amount of frequency tabulations will avail a competent, expert cryptanalyst in demonstrating or disclosing the presence of a cryptographic message, written and secreted within the "open" message, which serves but as an envelop and disguise for its authentic or real import. Certainly, such frequency tabulations can disclose the existence *neither* of substitution *nor* transposition in these cases, since both forms are absent. Another very popular method that resembles the method mentioned above has for its basis a simple grille. The whole words forming the secret text are inserted within perforations cut in the paper and the remaining space filled carefully, using "nulls" and "dummies", making a seemingly innocuous, ordinary message. There are other methods of this general type which can obviously neither be detected nor cryptanalyzed, using the principles of frequency of recurrences and repetition. These can not be further discussed herein, but at a subsequent date a special text may be written for their handling.[2]

[1] See Sec. XV, *Elementary Military Cryptography.*

[2] The subparagraph which the student has just read (47*c*) contains a hidden cryptographic message. With the hints given in Par. 35*e* let the student see if he can find it.

48. Ciphers employing characters other than letters or figures.—*a*. In view of the foregoing remarks, when so-called symbol ciphers, that is, ciphers employing peculiar symbols, signs of punctuation, diacritical marks, figures of "dancing men", and so on are encountered in practical work nowadays, they are almost certain to be simple, monoalphabetic ciphers. They are adequately described in romantic tales,[3] in popular books on cryptography, and in the more common types of magazine articles. No further space need be given ciphers of this type in this text, not only because of their simplicity but also because they are encountered in military cryptography only in sporadic instances, principally in censorship activities. Even in the latter cases, it is usually found that such ciphers are employed in "intimate" correspondence for the exchange of sentiments that appear less decorous when set forth in plain language. They are very seldom used by authentic enemy agents. When such a cipher is encountered nowadays it may practically always be regarded as the work of the veriest tyro, when it is not that of a "crank" or a mentally-deranged person.

b. The usual preliminary procedure in handling such cases, where the symbols may be somewhat confusing to the mind because of their unfamiliar appearance to the eye, is to substitute letters for them consistently throughout the message and then treat the resulting text as an ordinary cryptogram composed of letters is treated. This procedure also facilitates the construction of the necessary frequency distributions, which would be tedious to construct by using symbols.

c. A final word must be said on the subject of symbol ciphers by way of caution. When symbols are used to replace letters, syllables, and entire words, then the systems approach code methods in principle, and can become difficult of solution.[4] The logical extension of the use of symbols in such a form of writing is the employment of arbitrary characters for a specially developed "shorthand" system bearing little or no resemblance to well-known, and therefore nonsecret, systems of shorthand, such as Gregg, Pitman, etc. Unless a considerable amount of text is available for analysis, a privately-devised shorthand may be very difficult to solve. Fortunately, such systems are rarely encountered in military cryptography. They fall under the heading of cryptographic curiosities, of interest to the cryptanalyst in his leisure moments.[5]

d. In practical cryptography today, as has been stated above, the use of characters other than the 26 letters of the English alphabet is comparatively rare. It is true that there are a few governments which still adhere to systems yielding cryptograms in groups of figures. These are almost in every case code systems and will be treated in their proper place. In some cases cipher systems, or systems of enciphering code are used which are basically mathematical in character and operation, and therefore use numbers instead of letters. Some persons are inclined toward the use of numbers rather than letters because numbers lend themselves much more readily to certain arithmetical operations such as addition, subtraction, and so on, than do letters.[6] But there is usually added some final process whereby the figure groups are converted into letter groups, for the sake of economy in transmission.

[3] The most famous: Poe's *The Gold Bug;* Arthur Conan Doyle's *The Sign of Four.*

[4] The use of symbols for abbreviation and speed in writing goes back to the days of antiquity. Cicero is reported to have drawn up "a book like a dictionary, in which he placed before each word the notation (symbol) which should represent it, and so great was the number of notations and words that whatever could be written in Latin could be expressed in his notations."

[5] An example is found in the famous Pepys Diary, which was written in shorthand, purely for his own eyes by Samuel Pepys (1633–1703). "He wrote it in Shelton's system of tachygraphy (1641), which he complicated by using foreign languages or by varieties of his own invention whenever he had to record passages least fit to be seen by his servants, or by 'all the world.' "

[6] But, this of course, is because we are taught arithmetic by using numbers, based upon the decimal system as a rule. By special training one *could* learn to perform the usual "arithmetical" operations using letters. For example, using our English alphabet of 26 letters, where A=1, B=2, C=3, etc., it is obvious that A+B=C, just as 1+2=3; $(A+B)^2=I$, etc. This sort of cryptographic arithmetic could be learned by rote, just as multiplication tables are learned.

e. The only notable exceptions to the statement contained in the first sentence of the preceding subparagraph are those of Russian messages transmitted in the Russian Morse alphabet and Japanese messages transmitted in the Kata Kana Morse alphabet. As regards Chinese, which is not an alphabetical language and comprises some 40,000 ideographs, since the Morse telegraph code comprises only some 40 combinations, telegrams in Chinese are usually prepared by means of codes which permit of substituting arbitrarily-assigned code groups for the characters. Usually the code groups consist of figures. One such code known as the *Official Chinese Telegraph Code*, has about 10,000 4-figure groups, beginning with 0001, and these are arranged so that there are 100 characters on each page. Sometimes, for purposes of secrecy or economy, these figure groups are enciphered and converted in letter groups.

49. Concluding remarks concerning monoalphabetic substitution.—*a*. The alert student will have by this time gathered that the solution of monoalphabetic substitution ciphers of the simple or fixed type are particularly easy to solve, once the underlying principles are thoroughly understood. As in other arts, continued practice with examples leads to facility and skill in solution, especially where the student concentrates his attention upon traffic all of the same general nature, so that the type of text which he is continually encountering becomes familiar to him and its peculiarities or characteristics of construction give clues for short cuts to solution. It is true that a knowledge of the general phraseology of messages, the kind of words used, their sequences, and so on, is of very great assistance in practical work in all fields of cryptanalysis. The student is urged to note particularly these finer details in the course of his study.

b. Another thing which the student should be on the lookout for in simple monoalphabetic substitution is the consecutive use of several different mixed cipher alphabets in a single long message. Obviously, a single, composite frequency distribution for the whole message will not show the characteristic crest and trough appearance of a simple monoalphabetic cipher, since a given cipher letter will represent different plain-text letters in different parts of the message. But if the cryptanalyst will carefully observe the distribution *as it is being compiled*, he will note that at first it presents the characteristic crest and trough appearance of monoalphabeticity, and that after a time it begins to lose this appearance. If possible he should be on the lookout for some peculiarity of grouping of letters which serves as an indicator for the shift from one cipher alphabet to the next. If he finds such an indicator he should begin a second distribution from that point on, and proceed until another shift or indicator is encountered. By thus isolating the different portions of the text, and restricting the frequency distributions to the separate monoalphabets, the problem may be treated then as an ordinary simple monoalphabetic substitution. Consideration of these remarks in connection with instances of this kind leads to the comment that it is often more advisable for the cryptanalyst to compile his own data, than to have the latter prepared by clerks, especially when studying a system *de novo*. For observations which will certainly escape an untrained clerk can be most useful and may indeed facilitate solution. For example, in the case under consideration, if a clerk should merely hand the uniliteral distribution to the cryptanalyst, the latter might be led astray; the appearance of the composite distribution might convince him that the cryptogram is a good deal more complicated than it really is.

c. Monoalphabetic substitution with variants represents an extension of the basic principle, with the intention of masking the characteristic frequencies resulting from a strict monoalphabeticity, by means of which solutions are rather readily obtained. Some of the subterfuges applied on the establishment of variant or multiple values are simple and more or less fail to serve the purpose for which they are intended; others, on the contrary, may interpose serious difficulties to a straightforward solution. But in no case may the problem be considered of more than ordinary difficulty. Furthermore, it should be recognized that where these subterfuges are really adequate to the purpose, the complications introduced are such that the practical manipulation of the system becomes as difficult for the cryptographer as for the cryptanalyst.

d. As already mentioned in monoalphabetic substitution with variants it is most common to employ figures or groups of figures. The reason for this is that the use of numerical groups seems more natural or easier to the uninitiated than does the use of varying combinations of letters. Moreover, it is easy to draw up cipher alphabets in which some of the letters are represented by single digits, others by pairs of digits. Thus, the decomposition of the cipher text which is an irregular intermixture of uniliteral and multiliteral equivalents, is made more complicated and correspondingly difficult for the cryptanalyst, who does not known which digits are to be used separately, which in pairs.

e. A few words may be added here in regard to a method which often suggests itself to laymen. This consists in using a book possessed by all the correspondents and indicating the letters of the message by means of numbers referring to specific letters in the book. One way consists in selecting a certain page and then giving the line number and position of the letter in the line, the page number being shown by a single initial indicator. Another way is to use the entire book, giving the cipher equivalents in groups of three numbers representing page, line, and number of letter. (Ex.: 75–8–10 means page 75, 8th line, 10th letter in the line.) Such systems are, however, extremely cumbersome to use and, when the cryptographing is done carelessly, can be solved. The basis for solution in such cases rests upon the use of adjacent letters on the same line, the accidental repetitions of certain letters, and the occurrence of unenciphered words in the messages, when laziness or fatigue intervenes in the cryptographing.[7]

f. It may also be indicated that human nature and the fallibility of cipher clerks is such that it is rather rare for an encipherer to make full use of the complement of variants placed at his disposal. The result is that in most cases certain of the equivalents will be used so much more often than others that diversities in frequencies will soon manifest themselves, affording important data for attack by the cryptanalyst.

g. In the World War the cases where monoalphabetic substitution ciphers were employed in actual operations on the Western Front were exceedingly rare because the majority of the belligerents had a fair knowledge of cryptography. On the Eastern Front, however, the extensive use, by the poorly prepared Russian Army, of monoalphabetic ciphers in the fall of 1914 was an important, if not the most important, factor in the success of the German operations during the Battle of Tannenberg.[8] It seems that a somewhat more secure cipher system was authorized, but proved too difficult for the untrained Russian cryptographic and radio personnel. Consequently, recourse was had to simple substitution ciphers, somewhat interspersed with plain text, and sometimes to messages completely in plain language. The damage which this faulty use of cryptography did to the Russian Army and thus to the Allied cause is incalculabl

h. Many of the messages found by censors in letters sent by mail during the World War were cases of monoalphabetic substitution, disguised in various ways.

[7] In 1915 the German Government conspired with a group of Hindu revolutionaries to stir up a rebellion in India, the purpose being to cause the withdrawal of British troops from the Western Front. Hindu conspirators in the United States were given money to purchase arms and ammunition and to transport them to India. For communication with their superiors in Berlin the conspirators used, among others, the system described in this paragraph. A 7-page typewritten letter, built up from page, line, and letter-number references to a book known only to the communicants, was intercepted by the British and turned over to the United States Government for use in connection with the prosecution of the Hindus for violating our neutrality. The author solved this message without the book in question, by taking full advantage of the clues referred to.

[8] Gyldén, Yves. *Chifferbyåernas Insatser I Världskriget Till Lands*, Stockholm, 1931. A translation under the title *The Contribution of the Cryptographic Bureaus in the World War*, appeared in the Signal Corps Bulletin in seven successive installments, from November–December 1933 to November–December 1934, inclusive.

Nikolaieff, A. M. *Secret Causes of German success on the Eastern Front.* Coast Artillery Journal, September–October, 1935.

50. Analytical key for cryptanalysis.—*a.* It may be of assistance to indicate, by means of an outline, the relationships existing among the various cryptographic systems thus far considered. This graphic outline will be augmented from time to time as the different cipher systems are examined, and will constitute what has already been alluded to in Par. 6*d* and there termed an analytical key for cryptanalysis.[9] Fundamentally its nature is that of a schematic classification of the different systems examined. The analytical key forms an insert at the end of the book.

b. Note, in the analytical key, the rather clear-cut, dichotomous method of treatment; that is, classification by subdivision into pairs. For example, in the very first step there are only two alternatives: the cryptogram is either (1) cipher, or (2) code. If it is cipher, it is either (1) substitution, (2) transposition. If it is a substitution cipher, it is either (1) monographic, or (2) polygraphic—and so on. If the student will study the analytical key attentively, it will assist him in fixing in mind the manner in which the various systems covered thus far are related to one another, and this will be of benefit in clearing away some of the mental fog or haziness from which he is at first apt to suffer.

c. The numbers in parentheses refer to specific paragraphs in this text, so that the student may readily turn to the text for detailed information or for purposes of refreshing his memory as to procedure.

d. In addition to these reference numbers there have been affixed to the successive steps in the dichotomy, numbers that mark the "routes" on the cryptanalytic map (the analytical key) which the student cryptanalyst should follow if he wishes to facilitate his travels along the rather complicated and difficult road to success in cryptanalysis, in somewhat the same way in which an intelligent motorist follows the routes indicated on a geographical map if he wishes to facilitate his travels along unfamiliar roads. The analogy is only partially valid, however. The motorist usually knows in advance the distant point which he desires to reach and he proceeds thereto by the best and shortest route, which he finds by observing the route indications on a map and following the route markers on the road. Occasionally he encounters a detour but these are unexpected difficulties as a rule. Least of all does he anticipate any necessity for journeys down what may soon turn out to be blind alleys and "dead-end" streets, forcing him to double back on his way. Now the cryptanalyst also has a distant goal in mind—the solution of the cryptogram at hand—but he does not know at the outset of his journey the exact spot where it is located on the cryptanalytic map. The map contains many routes and he proceeds

[9] This analytical key is quite analogous to the analytical keys usually found in the handbooks biologists commonly employ in the classification and identification of living organisms. In fact, there are several points of resemblance between, for example, that branch of biology called taxonomic botany and cryptanalysis. In the former the first steps in the classificatory process are based upon observation of externally quite marked differences; as the process continues, the observational details become finer and finer, involving more and more difficulties as the work progresses. Towards the end of the work the botanical taxonomist may have to dissect the specimen and study internal characteristics. The whole process is largely a matter of painstaking, accurate observation of data and drawing proper conclusions therefrom. Except for the fact that the botanical taxonomist depends almost entirely upon ocular observation of characteristics while the cryptanalyst in addition to observation must use some statistics, the steps taken by the former are quite similar to those taken by the latter. It is only at the very end of the work that a significant dissimilarity between the two sciences arises. If the botanist makes a mistake in observation or deduction, he merely fails to identify the specimen correctly; he has an "answer"—but the answer is wrong. He may not be cognizant of the error; however, other more skillful botanists will find him out. But if the cryptanalyst makes a mistake in observation or deduction, he fails to get any "answer" at all; he needs nobody to tell him he has failed. Further, there is one additional important point of difference. The botanist is studying a bit of Nature—and she does not consciously interpose obstacles, pitfalls, and dissimulations in the path of those trying to solve her mysteries. The cryptanalyst, on the other hand, is studying a piece of writing prepared with the express purpose of preventing its being read by any persons for whom it is not intended. The obstacles, pitfalls, and dissimulations are here consciously interposed by the one who cryptographed the message. These, of course, are what make cryptanalysis different and difficult.

to test them one by one, in a successive chain. He encounters many blind alleys and dead-end streets, which force him to retrace his steps; he makes many detours and jumps many hurdles. Some of these retracings of steps, doubling back on his tracks, jumping of hurdles, and detours are unavoidable, but a few are avoidable. If properly employed, the analytical key will help the careful student to avoid those which should and can be avoided; if it does that much it will serve the principal purpose for which it is intended.

e. The analytical key may, however, serve another purpose of a somewhat different nature. When a multitude of cryptographic systems of diverse types must be filed in some systematic manner apart from the names of the correspondents or other reference data, or if in conducting instructional activities classificatory designations are desirable, the reference numbers on the analytical key may be made to serve as "type numbers." Thus, instead of stating that a given cryptogram is a keyword-systematically-mixed-uniliteral-monoalphabetic-monographic substitution cipher one may say that it is a "Type 901 cryptogram."

f. The method of assigning type numbers is quite simple. If the student will examine the numbers he will note that successive levels in the dichotomy are designated by successive hundreds. Thus, the first level, the classification into cipher and code is assigned the numbers 101 and 102. On the second level, under cipher, the classification into monographic and polygraphic systems is assigned the numbers 201 and 202, etc. Numbers in the same hundreds apply therefore to systems at the same level in the classification. There is no particular virtue in this scheme of assigning type numbers except that it provides for a considerable degree of expansion in future studies.

APPENDIX 1

(105)

APPENDIX

Table 1–A.—*Absolute frequencies of letters appearing in five sets of Governmental plain-text telegrams, each set containing 10,000 letters, arranged alphabetically*

Set No. 1		Set No. 2		Set No. 3		Set No. 4		Set No. 5	
Letter	Absolute Frequency	Letter	Absolute Frequency	Letter	Absolute Frequency	Letter	Absolute Frequency	Letter	Absolute Frequency
A	738	A	783	A	681	A	740	A	741
B	104	B	103	B	98	B	83	B	99
C	319	C	300	C	288	C	326	C	301
D	387	D	413	D	423	D	451	D	448
E	1, 367	E	1, 294	E	1, 292	E	1, 270	E	1, 275
F	253	F	287	F	308	F	287	F	281
G	166	G	175	G	161	G	167	G	150
H	310	H	351	H	335	H	349	H	349
I	742	I	750	I	787	I	700	I	697
J	18	J	17	J	10	J	21	J	16
K	36	K	38	K	22	K	21	K	31
L	365	L	393	L	333	L	386	L	344
M	242	M	240	M	238	M	249	M	268
N	786	N	794	N	815	N	800	N	780
O	685	O	770	O	791	O	756	O	762
P	241	P	272	P	317	P	245	P	260
Q	40	Q	22	Q	45	Q	38	Q	30
R	760	R	745	R	762	R	735	R	786
S	658	S	583	S	585	S	628	S	604
T	936	T	879	T	894	T	958	T	928
U	270	U	233	U	312	U	247	U	238
V	163	V	173	V	142	V	133	V	155
W	166	W	163	W	136	W	133	W	182
X	43	X	50	X	44	X	53	X	41
Y	191	Y	155	Y	179	Y	213	Y	229
Z	14	Z	17	Z	2	Z	11	Z	5
Total	10, 000		10, 000		10, 000		10, 000		10, 000

Table 2–A.—*Absolute frequencies of letters appearing in the combined five sets of messages totalling 50,000 letters, arranged alphabetically*

A	3, 683	G	819	L	1, 821	Q	175	V	766
B	487	H	1, 694	M	1, 237	R	3, 788	W	780
C	1, 534	I	3, 676	N	3, 975	S	3, 058	X	231
D	2, 122	J	82	O	3, 764	T	4, 595	Y	967
E	6, 498	K	148	P	1, 335	U	1, 300	Z	49
F	1, 416								

Table 1–B.—*Absolute frequencies of letters appearing in five sets of Government plain-text telegrams, each set containing 10,000 letters, arranged according to frequency*

Set No. 1		Set No. 2		Set No. 3		Set No. 4		Set No. 5	
Letter	Absolute Frequency	Letter	Absolute Frequency	Letter	Absolute Frequency	Letter	Absolute Frequency	Letter	Absolute Frequency
E	1, 367	E	1, 294	E	1, 292	E	1, 270	E	1, 275
T	936	T	879	T	894	T	958	T	928
N	786	N	794	N	815	N	800	R	786
R	760	A	783	O	791	O	756	N	780
I	742	O	770	I	787	A	740	O	762
A	738	I	750	R	762	R	735	A	741
O	685	R	745	A	681	I	700	I	697
S	658	S	583	S	585	S	628	S	604
D	387	D	413	D	423	D	451	D	448
L	365	L	393	H	335	L	386	H	349
C	319	H	351	L	333	H	349	L	344
H	310	C	300	P	317	C	326	C	301
U	270	F	287	U	312	F	287	F	281
F	253	P	272	F	308	M	249	M	268
M	242	M	240	C	288	U	247	P	260
P	241	U	233	M	238	P	245	U	238
Y	191	G	175	Y	179	Y	213	Y	229
G	166	V	173	G	161	G	167	W	182
W	166	W	163	V	142	V	133	V	155
V	163	Y	155	W	136	W	133	G	150
B	104	B	103	B	98	B	83	B	99
X	43	X	50	Q	45	X	53	X	41
Q	40	K	38	X	44	Q	38	K	31
K	36	Q	22	K	22	K	21	Q	30
J	18	J	17	J	10	J	21	J	16
Z	14	Z	17	Z	2	Z	11	Z	5
Total	10, 000		10, 000		10, 000		10, 000		10, 000

148274—38——8

TABLE 1–C.—*Absolute frequencies of vowels, high frequency consonants, medium frequency consonants, and low frequency consonants appearing in five sets of Government plain-text telegrams, each set containing 10,000 letters*

Set No.	Vowels	High Frequency Consonants	Medium Frequency Consonants	Low Frequency Consonants
1	3, 993	3, 527	2, 329	151
2	3, 985	3, 414	2, 457	144
3	4, 042	3, 479	2, 356	123
4	3, 926	3, 572	2, 358	144
5	3, 942	3, 546	2, 389	123
Total [1]	19, 888	17, 538	11, 889	685

[1] Grand total, 50,000.

TABLE 2–B.—*Absolute frequencies of letters appearing in the combined five sets of messages totalling 50,000 letters arranged according to frequencies*

E	6, 498	I	3, 676	C	1, 534	Y	967	X	231
T	4, 595	S	3, 058	F	1, 416	G	819	Q	175
N	3, 975	D	2, 122	P	1, 335	W	780	K	148
R	3, 788	L	1, 821	U	1, 300	V	766	J	82
O	3, 764	H	1, 694	M	1, 237	B	487	Z	49
A	3, 683								

TABLE 2–C.—*Absolute frequencies of vowels, high frequency consonants, medium frequency consonants, and low frequency consonants appearing in the combined five sets of messages totalling 50,000 letters*

Vowels	19, 888
High Frequency Consonants (D, N, R, S, and T)	17, 538
Medium Frequency Consonants (B, C, F, G, H, L, M, P, V, and W)	11, 889
Low Frequency Consonants (J, K, Q, X, and Z)	685
Total	50, 000

TABLE 2–D.—*Absolute frequencies of letters as initial letters of 10,000 words found in Government plain-text telegrams*

(1) ARRANGED ALPHABETICALLY

A	905	G	109	L	196	Q	30	V	77
B	287	H	272	M	384	R	611	W	320
C	664	I	344	N	441	S	965	X	4
D	525	J	44	O	646	T	1, 253	Y	88
E	390	K	23	P	433	U	122	Z	12
F	855								

Total 10, 000

(2) ARRANGED ACCORDING TO ABSOLUTE FREQUENCIES

T	1, 253	R	611	M	384	L	196	J	44
S	965	D	525	I	344	U	122	Q	30
A	905	N	441	W	320	G	109	K	23
F	855	P	433	B	287	Y	88	Z	12
C	664	E	390	H	272	V	77	X	4
O	646								

Total 10, 000

TABLE 2–E.—*Absolute frequencies of letters as final letters of 10,000 words found in Government plain-text telegrams*

(1) ARRANGED ALPHABETICALLY

A	269	G	225	L	354	Q	8	V	4
B	22	H	450	M	154	R	769	W	45
C	86	I	22	N	872	S	962	X	116
D	1, 002	J	6	O	575	T	1, 007	Y	866
E	1, 628	K	53	P	213	U	31	Z	9
F	252								

Total 10,000

(2) ARRANGED ACCORDING TO ABSOLUTE FREQUENCIES

E	1, 628	R	769	F	252	C	86	I	22
T	1, 007	O	575	G	225	K	53	Z	9
D	1, 002	H	450	P	213	W	45	Q	8
S	962	L	354	M	154	U	31	J	6
N	872	A	269	X	116	B	22	V	4
Y	866								

Total 10,000

TABLE 3.—*Relative frequencies of letters appearing in 1,000 letters based upon Table 2-B*

(1) ARRANGED ALPHABETICALLY

A	73.66	G	16.38	L	36.42	Q	3.50	V	15.32
B	9.74	H	33.88	M	24.74	R	75.76	W	15.60
C	30.68	I	73.52	N	79.50	S	61.16	X	4.62
D	42.44	J	1.64	O	75.28	T	91.90	Y	19.34
E	129.96	K	2.96	P	26.70	U	26.00	Z	.98
F	28.32								
								Total	1,000.00

(2) ARRANGED ACCORDING TO FREQUENCY

E	129.96	I	73.52	C	30.68	Y	19.34	X	4.62
T	91.90	S	61.16	F	28.32	G	16.38	Q	3.50
N	79.50	D	42.44	P	26.70	W	15.60	K	2.96
R	75.76	L	36.42	U	26.00	V	15.32	J	1.64
O	75.28	H	33.88	M	24.74	B	9.74	Z	.98
A	73.66								
								Total	1,000.00

(3) VOWELS

A	73.66
E	129.96
I	73.52
O	75.28
U	26.00
Y	19.34
Total	397.76

(4) HIGH-FREQUENCY CONSONANTS

D	42.44
N	79.50
R	75.76
S	61.16
T	91.90
Total	350.76

(5) MEDIUM-FREQUENCY CONSONANTS

B	9.74
C	30.68
F	28.32
G	16.38
H	33.88
L	36.42
M	24.74
P	26.70
V	15.32
W	15.60
Total	237.78

(6) LOW-FREQUENCY CONSONANTS

X	4.62
Q	3.50
K	2.96
J	1.64
Z	.98
Total	13.70
Total (3), (4), (5), (6)	1,000.00

TABLE 6.—*Frequency distribution of digraphs—Based on 50,000 letters of Government plain-text telegrams; reduced to 5,000 digraphs*

SECOND LETTER

FIRST LETTER	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	Total	Blanks
A	3	6	14	27	1	4	6	2	17	1	2	32	14	64	2	12		44	41	47	13	7	3		12		374	3
B	4				18				2	1		6	1		4			2	1	1	2				7		49	14
C	20		3	1	32	1		14	7		4	5	1	1	41			4	1	14	4		1		1		155	8
D	32	4	4	8	33	8	2	2	27	1		3	5	4	16	5	2	12	13	15	5	3	4		1		209	3
E	35	4	32	60	42	18	4	7	27	1		29	14	111	12	20	12	87	54	37	3	20	7	7	4	1	648	1
F	5		2	1	10	11	1		39			2	1		40	1		9	3	11	3		1		1		141	9
G	7		2	1	14	2	1	20	5	1		2	1	3	6	2		5	3	4	2		1				82	7
H	20	1	3	2	20	5			33			1	2	3	20	1	1	17	4	28	8		1		1		171	7
I	8	2	22	6	13	10	19				2	23	9	75	41	7		27	35	27		25		15		2	368	7
J	1				2										2						2						7	22
K	1		1		6				2			1		1					1								13	19
L	28	3	3	9	37	3	1	1	20			27	2	1	13	3		2	6	8	2	2	2		10		183	5
M	36	6	3	1	26	1		1	9				13		10	8		2	4	2	2				2		126	10
N	26	2	19	52	57	9	27	4	30	1	2	5	5	8	18	3	1	4	24	82	7	3	3		5		397	2
O	7	4	8	12	3	25	2	3	5	1	2	19	25	77	6	25		64	14	19	37	7	8	1	2		376	2
P	14	1	1	1	23	2		3	6			13	4	1	17	11		18	6	8	3	1	1		1		135	6
Q													1					1			15						17	23
R	39	2	9	17	98	6	7	3	30	1	1	5	9	7	28	13		11	31	42	5	5	4		9		382	3
S	24	3	13	5	49	12	2	26	34		1	2	3	4	15	10		5	19	63	11	1	4		1		307	4
T	28	3	6	6	71	7	1	78	45			5	6	7	50	2	1	17	19	19	5		36		41	1	454	4
U	5	3	3	3	11	1	8		5			6	5	21	1	2		31	12	12		1					130	9
V	6				57				12						1					1							77	21
W	12				22			4	13			1		2	19			1	1						1		76	16
X	2		2	1	1	1		1	2					1	1	2		1	1	7							23	13
Y	6	2	4	4	9	11	1	1	3			2	2	6	10	3		4	11	15	1		1				96	7
Z	1				2				1																		4	23
Total	370	46	154	217	657	137	82	170	374	8	14	189	123	397	373	130	17	368	304	462	130	75	77	23	99	4	5,000	
Blanks	1	11	6	7	1	7	12	10	3	18	19	6	6	7	3	8	21	4	4	5	7	15	11	23	10	23		248

TABLE 4.—*Frequency distribution for 10,000 letters of literary English, as compiled by Hitt* [1]

(1) ALPHABETICALLY ARRANGED

A	778	G	174	L	372	Q	8	V	112
B	141	H	595	M	288	R	651	W	176
C	296	I	667	N	686	S	622	X	27
D	402	J	51	O	807	T	855	Y	196
E	1,277	K	74	P	223	U	308	Z	17
F	197								

(2) ARRANGED ACCORDING TO FREQUENCY

E	1,277	R	651	U	308	Y	196	K	74
T	855	S	622	C	296	W	176	J	51
O	807	H	595	M	288	G	174	X	27
A	778	D	402	P	223	B	141	Z	17
N	686	L	372	F	197	V	112	Q	8
I	667								

TABLE 5.—*Frequency distribution for 10,000 letters of telegraphic English as compiled by Hitt*

(1) ALPHABETICALLY ARRANGED

A	813	G	201	L	392	Q	38	V	136
B	149	H	386	M	273	R	677	W	166
C	306	I	711	N	718	S	656	X	51
D	417	J	42	O	844	T	634	Y	208
E	1,319	K	88	P	243	U	321	Z	6
F	205								

(2) ARRANGED ACCORDING TO FREQUENCY

E	1,319	S	656	U	321	F	205	K	88
O	844	T	634	C	306	G	201	X	51
A	813	D	417	M	273	W	166	J	42
N	718	L	392	P	243	B	149	Q	38
I	711	H	386	Y	208	V	136	Z	6
R	677								

[1] Hitt, Capt. Parker. *Manual for the Solution of Military Ciphers.* Army Service Schools Press, Fort Leavenworth, Kansas, 1916.

TABLE 7-A.—*The 438 different digraphs of table 6 arranged according to their absolute frequencies*

EN	111	EC	32	OL	19	US	12
RE	98	RS	31	OT	19	UT	12
ER	87	UR	31	TS	19	VI	12
NT	82	NI	30	WO	19	WA	12
TH	78	RI	30	BE	18	FF	11
ON	77	EL	29	EF	18	PP	11
IN	75	HT	28	NO	18	RR	11
TE	71	LA	28	PR	18	UE	11
AN	64	RO	28	AI	17	FT	11
OR	64	TA	28	HR	17	SU	11
ST	63			PO	17	YF	11
ED	60		[2]2,495	RD	17	YS	11
NE	57	LL	27	TR	17	YO	10
VE	57	AD	27	DO	16	FE	10
ES	54	DI	27	DT	15	IF	10
ND	52	EI	27	IX	15	LY	10
TO	50	IR	27	QU	15	MO	10
SE	49	IT	27	SO	15	SP	10
		NG	27	YT	15	YE	9
	[1]1,249	ME	26	AC	14	FR	9
AT	47	NA	26	AM	14	IM	9
TI	45	SH	26	CH	14	LD	9
AR	44	IV	25	CT	14	MI	9
EE	42	OF	25	EM	14	NF	9
RT	42	OM	25	GE	14	RC	9
AS	41	OP	25	OS	14	RM	9
CO	41	NS	24	PA	14	RY	9
IO	41	SA	24	PL	13	DD	8
TY	41	IL	23	RP	13	NN	8
FO	40	PE	23	SC	13	DF	8
FI	39	IC	22	WI	13	IA	8
RA	39	WE	22	MM	13	HU	8
ET	37	UN	21	DS	13	LT	8
OU	37	CA	20	AU	13	MP	8
LE	37	EP	20	IE	13	OC	8
MA	36	EV	20	LO	13	OW	8
TW	36	GH	20			PT	8
EA	35	HA	20		[3]3,745	UG	8
IS	35	HE	20	AP	12	AV	7
SI	34	HO	20	DR	12	BY	7
DE	33	LI	20	EQ	12	CI	7
HI	33	SS	19	AY	12	EH	7
AL	32	TT	19	EO	12	OA	7
CE	32	IG	19	OD	12	EW	7
DA	32	NC	19	SF	12	EX	7

[1] The 18 digraphs above this line compose 25% of the total.
[2] The 53 digraphs above this line compose 50% of the total.
[3] The 117 digraphs above this line compose 75% of the total.

TABLE 7-A.—*The 438 different digraphs of table 6 arranged according to their absolute frequencies*—Continued

GA	7	SD	5	DV	3	KI	2
IP	7	SR	5	AA	3	LM	2
NU	7	TL	5	EU	3	LR	2
OV	7	TU	5	OE	3	LU	2
RG	7	UM	5	YI	3	LV	2
RN	7	AF	4	FS	3	LW	2
TE	7	BA	4	FU	3	MR	2
TN	7	BO	4	GN	3	MT	2
XT	7	CK	4	GS	3	MU	2
AB	6	CR	4	HC	3	MY	2
AG	6	CU	4	HN	3	NB	2
BL	6	DB	4	LB	3	NK	2
OO	6	DC	4	LC	3	OG	2
YA	6	DN	4	LF	3	OK	2
GO	6	DW	4	LP	3	PF	2
ID	6	EB	4	MC	3	RB	2
KE	6	EG	4	NP	3	SG	2
LS	6	EY	4	NV	3	SL	2
MB	6	GT	4	NW	3	TP	2
PI	6	HS	4	OH	3	UP	2
PS	6	MS	4	AH	2	WN	2
RF	6	NH	4	AK	2	XA	2
TC	6	NR	4	BI	2	XC	2
TD	6	OB	4	BR	2	XI	2
TM	6	PM	4	BU	2	XP	2
UL	6	RW	4	DG	2	YB	2
VA	6	SN	4	DH	2	YL	2
YN	6	SW	4	DO	2	YM	2
CL	5	WH	4	AO	2	ZE	2
DM	5	YC	4	OY	2	GG	1
DP	5	YD	4	FC	2	AJ	1
DU	5	YR	4	FL	2	BJ	1
OI	5	PH	3	GC	2	BM	1
UA	5	PU	3	GF	2	BS	1
UI	5	RH	3	GL	2	BT	1
FA	5	SB	3	GP	2	CD	1
GI	5	SM	3	GU	2	CF	1
GR	5	TB	3	HD	2	CM	1
HF	5	UB	3	HM	2	CN	1
NL	5	UC	3	IB	2	CS	1
NM	5	UD	3	IK	2	CW	1
NY	5	YP	3	IZ	2	CY	1
RL	5	CC	3	JE	2	DJ	1
RU	5	AW	3	JO	2	DY	1
RV	5	DL	3	JU	2	EJ	1

TABLE 7-A.—*The 138 different digraphs of table 6 arranged according to their absolute frequencies*—Continued

AE	1	HY	1	PD	1	WL	1
UO	1	JA	1	PN	1	WR	1
YU	1	KA	1	PV	1	WS	1
EZ	1	KC	1	PW	1	WY	1
FD	1	KL	1	PY	1	XD	1
FG	1	KN	1	QM	1	XE	1
FM	1	KS	1	QR	1	XF	1
FP	1	LG	1	RJ	1	XH	1
FW	1	LH	1	RK	1	XN	1
FY	1	LN	1	SK	1	XO	1
GD	1	MD	1	SV	1	XR	1
GJ	1	MF	1	SY	1	XS	1
GM	1	MH	1	TG	1	YG	1
GW	1	NJ	1	TQ	1	YH	1
HB	1	NQ	1	TZ	1	YW	1
HL	1	OJ	1	UF	1	ZA	1
HP	1	OX	1	UV	1	ZI	1
HQ	1	PB	1	VO	1		
HW	1	PC	1	VT	1	Total	5,000

TABLE 7-B.—*The 18 digraphs composing 25% of the digraphs in Table 6 arranged alphabetically according to their initial letters*

(1) AND ACCORDING TO THEIR FINAL LETTERS

AN	64	ON	77
		OR	64
ED	60	RE	98
EN	111		
ER	87	SE	49
ES	54	ST	63
		TE	71
IN	75	TH	78
		TO	50
ND	52	VE	57
NE	57		
NT	82	Total	1,249

(2) AND ACCORDING TO THEIR ABSOLUTE FREQUENCIES

AN	64	ON	77
		OR	64
EN	111	RE	98
ER	87		
ED	60	SE	49
ES	54	ST	63
		TH	78
IN	75	TE	71
		TO	50
NT	82	VE	57
NE	57		
ND	52	Total	1,249

TABLE 7-C.—*The 53 digraphs composing 50% of the 5,000 digraphs of Table 6, arranged alphabetically according to their initial letters*

(1) AND ACCORDING TO THEIR FINAL LETTERS

AL	32	MA	36
AN	64		
AR	44	ND	52
AS	41	NE	57
AT	47	NI	30
		NT	82
CE	32		
CO	41	ON	77
		OR	64
DA	32	OU	37
DE	38		
		RA	39
EA	35	RE	98
EC	32	RI	30
ED	60	RO	28
EE	42	RS	31
EL	29	RT	42
EN	111		
ER	87	SE	49
ES	54	SI	34
ET	37	ST	63
		TA	28
FI	39	TE	71
FO	40	TH	78
		TI	45
HI	33	TO	50
HT	28	TW	36
		TY	41
IN	75		
IO	41	UR	31
IS	35		
		VE	57
LA	28		
LE	37	Total	2,495

(2) AND ACCORDING TO THEIR ABSOLUTE FREQUENCIES

AN	64	MA	36
AT	47		
AR	44	NT	82
AS	41	NE	57
AL	32	ND	52
		NI	30
CO	41		
CE	32	ON	77
		OR	64
DE	33	OU	37
DA	32		
		RE	98
EN	111	RT	42
ER	87	RA	39
ED	60	RS	31
ES	54	RI	30
EE	42	RO	28
ET	37		
EA	35	ST	63
EC	32	SE	49
EL	29	SI	34
FO	40	TH	78
FI	39	TE	71
		TO	50
HI	33	TI	45
HT	28	TY	41
		TW	36
		TA	28
IN	75		
IO	41	UR	31
IS	35		
		VE	57
LE	37		
LA	28	Total	2,495

Table 7–D.—*The 117 digraphs composing 75% of the 5,000 digraphs of Table 6, arranged alphabetically according to their initial letters——*

(1) AND ACCORDING TO THEIR FINAL LETTERS

AC	14	EP	20	LO	13	RI	30
AD	27	ER	87			RO	28
AI	17	ES	54	MA	36	RS	31
AL	32	ET	37	ME	26	RT	42
AM	14	EV	20				
AN	64			NA	26	SA	24
AR	44	FI	39	NC	19	SE	49
AS	41	FO	40	ND	52	SH	26
AT	47			NE	57	SI	34
AU	13	GE	14	NG	27	SO	15
		GH	20	NI	30	SS	19
BE	18			NO	18	ST	63
		HA	20	NS	24		
CA	20	HE	20	NT	82	TA	28
CE	32	HI	33			TE	71
CH	14	HO	20	OF	25	TH	78
CO	41	HR	17	OL	19	TI	45
CT	14	HT	28	OM	25	TO	50
				ON	77	TR	17
DA	32	IC	22	OP	25	TS	19
DE	33	IE	13	OR	64	TT	19
DI	27	IG	19	OS	14	TW	36
DO	16	IL	23	OT	19	TY	41
DS	13	IN	75	OU	37		
DT	15	IO	41			UN	21
		IR	27	PA	14	UR	31
EA	35	IS	35	PE	23		
EC	32	IT	27	PO	17	VE	57
ED	60	IV	25	PR	18		
EE	42	IX	15			WE	22
EF	18			QU	15	WO	19
EI	27	LA	28				
EL	29	LE	37	RA	39	YT	15
EM	14	LI	20	RD	17		
EN	111	LL	27	RE	98	Total	3,745

Table 7–D, Concluded.—*The 117 digraphs comprising 75% of the 5,000 digraphs of Table 6, arranged alphabetically according to their initial letters——*

(2) AND ACCORDING TO THEIR ABSOLUTE FREQUENCIES

AN	64	EI	27	MA	36	RI	30
AT	47	EP	20	ME	26	RO	28
AR	44	EV	20			RD	17
AS	41	EF	18	NT	82		
AL	32	EM	14	NE	57	ST	63
AD	27			ND	52	SE	49
AI	17	FO	40	NI	30	SI	34
AC	14	FI	39	NG	27	SH	26
AM	14			NA	26	SA	24
AU	13	GH	20	NS	24	SS	19
		GE	14	NC	19	SO	15
BE	18			NO	18		
		HI	33			TH	78
CO	41	HT	28	ON	77	TE	71
CE	32	HA	20	OR	64	TO	50
CA	20	HE	20	OU	37	TI	45
CH	14	HO	20	OF	25	TY	41
CT	14	HR	17	OM	25	TW	36
		IN	75	OP	25	TA	28
DE	33	IO	41	OL	19	TS	19
DA	32	IS	35	OT	19	TT	19
DI	27	IR	27	OS	14	TR	17
DO	16	IT	27				
DT	15	IV	25	PE	23	UR	31
DS	13	IL	23	PR	18	UN	21
		IC	22	PO	17		
EN	111	IG	19	PA	14	VE	57
ER	87	IX	15				
ED	60	IE	13	QU	15	WE	22
ES	54					WO	19
EE	42	LE	37				
ET	37	LA	28	RE	98		
EA	35	LL	27	RT	42	YT	15
EC	32	LI	20	RA	39		
EL	29	LO	13	RS	31	Total	3,745

Table 7–E.—*All the 438 digraphs of Table 6, arranged first alphabetically according to their initial letters and then alphabetically according to their final letters.*

(SEE TABLE 6.—READ ACROSS THE ROWS)

TABLE 8.—*The 438 different digraphs of Table 6, arranged first alphabetically according to their initial letters, and then according to their absolute frequencies under each initial letter* [1]

AN	64	CT	14	ED	60	GH	20
AT	47	CI	7	ES	54	GE	14
AR	44	CL	5	EE	42	GA	7
AS	41	CK	4	ET	37	GO	6
AL	32	CR	4	EA	35	GI	5
AD	27	CU	4	EC	32	GR	5
AI	17	CC	3	EL	29	GT	4
AC	14	CD	1	EI	27	GN	3
AM	14	CF	1	EP	20	GS	3
AU	13	CM	1	EV	20	GC	2
AP	12	CN	1	EF	18	GF	2
AY	12	CS	1	EM	14	GL	2
AV	7	CW	1	EO	12	GP	2
AB	6	CY	1	EQ	12	GU	2
AG	6			EH	7	GD	1
AF	4	DE	33	EW	7	GG	1
AA	3	DA	32	EX	7	GJ	1
AW	3	DI	27	EB	4	GM	1
AH	2	DO	16	EG	4	GW	1
AK	2	DT	15	EY	4		
AO	2	DS	13	EU	3		
AE	1	DR	12	EJ	1		
AJ	1	DD	8	EZ	1	HI	33
		DF	8			HT	28
BE	18	DM	5	FO	40	HA	20
BY	7	DP	5	FI	39	HE	20
BL	6	DU	5	FF	11	HO	20
BA	4	DB	4	FT	11	HR	17
BO	4	DC	4	FE	10	HU	8
BI	2	DN	4	FR	9	HF	5
BR	2	DW	4	FA	5	HS	4
BU	2	DL	3	FS	3	HC	3
BJ	1	DV	3	FU	3	HN	3
BM	1	DG	2	FC	2	HD	2
BS	1	DH	2	FL	2	HM	2
BT	1	DQ	2	FD	1	HB	1
		DJ	1	FG	1	HL	1
CO	41	DY	1	FM	1	HP	1
CE	32			FP	1	HQ	1
CA	20	EN	111	FW	1	HW	1
CH	14	ER	87	FY	1	HY	1

[1] For arrangement alphabetically first under intial letters and then under final letters, see Table 6.

TABLE 8, Contd.—*The 438 different digraphs of Table 6, arranged first alphabetically according to their initial letters, and then according to their absolute frequencies under each initial letter* [1]

IN	75	LI	20	NE	57	OA	7
IO	41	LO	13	ND	52	OV	7
IS	35	LY	10	NI	30	OO	6
IR	27	LD	9	NG	27	OI	5
IT	27	LT	8	NA	26	OB	4
IV	25	LS	6	NS	24	OE	3
IL	23	LB	3	NC	19	OH	3
IC	22	LC	3	NO	18	OG	2
IG	19	LF	3	NF	9	OK	2
IX	15	LP	3	NN	8	OY	2
IE	13	LM	2	NU	7	OJ	1
IF	10	LR	2	NL	5	OX	1
IM	9	LU	2	NM	5		
IA	8	LV	2	NY	5	PE	23
IP	7	LW	2	NH	4	PR	18
ID	6	LG	1	NR	4	PO	17
		LH	1	NP	3	PA	14
IB	2	LN	1	NV	3	PL	13
IK	2			NW	3	PP	11
IZ	2	MA	36	NB	2	PT	8
		ME	26	NK	2	PI	6
JE	2	MM	13	NJ	1	PS	6
JO	2	MO	10	NQ	1	PM	4
JU	2	MI	9			PH	3
JA	1	MP	8	ON	77	PU	3
		MB	6	OR	64	PF	2
KE	6	MS	4	OU	37	PB	1
KI	2	MC	3	OF	25	PC	1
KA	1	MR	2	OM	25	PD	1
KC	1	MT	2	OP	25	PN	1
KL	1	MU	2	OL	19	PV	1
KN	1	MY	2	OT	19	PW	1
KS	1	MD	1	OS	14	PY	1
		MF	1				
LE	37	MH	1	OD	12	QU	15
LA	28			OC	8	QM	1
LL	27	NT	82	OW	8	QR	1

[1] For arrangement alphabetically first under initial letters and then under final letters, see Table 6.

TABLE 8, Concluded.—*The 438 different digraphs of Table 6, arranged first alphabetically according to their initial letters, and then according to their absolute frequencies under each initial letter* [1]

RE	98	SR	5	US	12	XI	2
RT	42	SN	4	UT	12	XP	2
RA	39	SW	4	UE	11	XD	1
RS	31	SB	3	UG	8	XE	1
RI	30	SM	3	UL	6	XF	1
RO	28	SG	2	UA	5	XH	1
RD	17	SL	2	UI	5	XN	1
RP	13	SK	1	UM	5	XO	1
RR	11	SV	1	UB	3	XR	1
RC	9	SY	1	UC	3	XS	1
RM	9			UD	3		
RY	9	TH	78	UP	2	YT	15
RG	7	TE	71	UF	1	YF	11
RN	7	TO	50	UO	1	YS	11
RF	6	TI	45	UV	1	YO	10
RL	5	TY	41			YE	9
RU	5	TW	36	VE	57	YA	6
RV	5	TA	28	VI	12	YN	6
RW	4	TS	19	VA	6	YC	4
RH	3	TT	19	VO	1	YD	4
RB	2	TR	17	VT	1	YR	4
RJ	1	TF	7			YI	3
RK	1	TN	7	WE	22	YP	3
		TC	6	WO	19	YB	2
ST	63	TD	6	WI	13	YL	2
SE	49	TM	6	WA	12	YM	2
SI	34	TL	5	WH	4	YG	1
SH	26	TU	5	WN	2	YH	1
SA	24	TB	3	WL	1	YU	1
SS	19	TP	2	WR	1	YW	1
SO	15	TG	1	WS	1		
SC	13	TQ	1	WY	1	ZE	2
SF	12	TZ	1			ZA	1
SU	11			XT	7	ZI	1
SP	10	UR	31	XA	2		
SD	5	UN	21	XC	2	Total	5,000

[1] For arrangement alphabetically first under initial letters and then under final letters, see Table 6.

TABLE 9–A.—*The 438 different digraphs of Table 6, arranged first alphabetically according to their final letters, and then according to their absolute frequencies*

RA	39	EC	32	RE	98	GF	2
MA	36	IC	22	TE	71	PF	1
EA	35	NC	19	NE	57	CF	2
DA	32	AC	14	VE	57	MF	1
LA	28	SC	13	SE	49	UF	1
TA	28	RC	9	EE	42	XF	1
NA	26	OC	8	LE	37		
SA	24	TC	6	DE	33		
CA	20	DC	4	CE	32	NG	27
HA	20	YC	4	ME	26	IG	19
PA	14	CC	3	PE	23	UG	8
WA	12	HC	3	WE	22	RG	7
IA	8	LC	3	HE	20	AG	6
GA	7	MC	3	BE	18	EG	4
OA	7	UC	3	GE	14	DG	2
VA	6	FC	2	IE	13	OG	2
YA	6	GC	2	UE	11	SG	2
FA	5	XC	2	FE	10	FG	1
UA	5	KC	1	YE	9	GG	1
BA	4	PC	1	KE	6	LG	1
AA	3			OE	3	TG	1
XA	2			JE	2	YG	1
JA	1	ED	60	ZE	2		
KA	1	ND	52	AE	1		
ZA	1	AD	27	XE	1		
		RD	17			TH	78
AB	6	OD	12			SH	26
MB	6	LD	9			GH	20
DB	4	DD	8	OF	25	CH	14
EB	4	ID	6	EF	18	EH	7
OB	4	TD	6	SF	12	NH	4
LB	3	SD	5	FF	11	WH	4
SB	3	YD	4	YF	11	OH	3
TB	3	UD	3	IF	10	PH	3
UB	3	HD	2	NF	9	RH	3
IB	2	CD	1	DF	8	AH	2
NB	2	FD	1	TF	7	DH	2
RB	2	GD	1	RF	6	LH	1
YB	2	MD	1	HF	5	MH	1
HB	1	PD	1	AF	4	XH	1
PB	1	XD	1	LF	3	YH	1

Table 9–A, Contd.—*The 438 different digraphs of Table 6, arranged first alphabetically according to their final letters, and then according to their absolute frequencies*

TI	45
FI	39
SI	34
HI	33
NI	30
RI	30
DI	27
EI	27
LI	20
AI	17
WI	13
VI	12
MI	9
CI	7
PI	6
GI	5
OI	5
UI	5
YI	3
BI	2
KI	2
XI	2
ZI	1
AJ	1
BJ	1
DJ	1
EJ	1
GJ	1
NJ	1
OJ	1
RJ	1
CK	4
AK	2
IK	2
NK	2
OK	2
RK	1
SK	1
AL	32
EL	29

LL	27
IL	23
OL	19
PL	13
BL	6
UL	6
CL	5
NL	5
RL	5
TL	5
DL	3
FL	2
GL	2
SL	2
YL	2
HL	1
KL	1
WL	1
OM	25
AM	14
EM	14
MM	13
IM	9
RM	9
TM	6
DM	5
NM	5
UM	5
PM	4
SM	3
HM	2
LM	2
YM	2
BM	1
CM	1
FM	1
GM	1
QM	1
EN	111
ON	77
IN	75

AN	64
UN	21
NN	8
RN	7
TN	7
YN	6
DN	4
SN	4
GN	3
HN	3
WN	2
CN	1
KN	1
LN	1
PN	1
XN	1
TO	50
CO	41
IO	41
FO	40
RO	28
HO	20
WO	19
NO	18
PO	17
DO	16
SO	15
LO	13
EO	12
MO	10
YO	10
GO	6
OO	6
BO	4
AO	2
JO	2
UO	1
VO	1
XO	1
OP	25
EP	20

RP	13
AP	12
PP	11
SP	10
MP	8
IP	7
DP	5
LP	3
NP	3
YP	3
GP	2
TP	2
UP	2
XP	2
FP	1
HP	1
EQ	12
DQ	2
HQ	1
NQ	1
TQ	1
ER	87
OR	64
AR	44
UR	31
IR	27
PR	18
HR	17
TR	17
DR	12
RR	11
FR	9
GR	5
SR	5
CR	4
NR	4
YR	4
BR	2
LR	2
MR	2
QR	1
WR	1
XR	1

Table 9–A, Concluded.—*The 438 different digraphs of Table 6, arranged first alphabetically according to their final letters, and then according to their absolute frequencies*

ES	54
AS	41
IS	35
RS	31
NS	24
SS	19
TS	19
OS	14
DS	13
US	12
YS	11
LS	6
PS	6
HS	4
MS	4
FS	3
GS	3
BS	1
CS	1
KS	1
WS	1
XS	1
NT	82
ST	63
AT	47
RT	42
ET	37
HT	28
IT	27

OT	19
TT	19
DT	15
YT	15
CT	14
UT	12
FT	11
LT	8
PT	8
XT	7
GT	4
MT	2
BT	1
VT	1
OU	37
QU	15
AU	13
SU	11
HU	8
NU	7
DU	5
RU	5
TU	5
CU	4
EU	3
FU	3
PU	3
BU	2
GU	2

JU	2
LU	2
MU	2
YU	1
IV	25
EV	20
AV	7
OV	7
RV	5
DV	3
NV	3
LV	2
PV	1
SV	1
UV	1
TW	36
OW	8
EW	7
DW	4
RW	4
SW	4
AW	3
NW	3
LW	2
CW	1
FW	1
GW	1
HW	1

PW	1
YW	1
IX	15
EX	7
OX	1
TY	41
AY	12
LY	10
RY	9
BY	7
NY	5
EY	4
MY	2
OY	2
CY	1
DY	1
FY	1
HY	1
PY	1
SY	1
WY	1
IZ	2
EZ	1
TZ	1
Total	5,000

148274—38——9

Table 9–B.—*The 18 digraphs composing 25% of the 5,000 digraphs of Table 6, arranged alphabetically according to their final letters*——

(1) AND ACCORDING TO THEIR INITIAL LETTERS

ED	60	IN	75
ND	52	ON	77
NE	57	TO	50
RE	98	ER	87
SE	49	OR	64
TE	71	ES	54
VE	57	NT	82
TH	78	ST	63
AN	64		
EN	111	Total	1, 249

(2) AND ACCORDING TO THEIR ABSOLUTE FREQUENCIES

ED	60	IN	75
ND	52	AN	64
RE	98	TO	50
TE	71	ER	87
NE	57	OR	64
VE	57	ES	54
SE	49	NT	82
TH	78	ST	63
EN	111		
ON	77	Total	1, 249

Table 9–C.—*The 53 digraphs composing 50% of the 5,000 digraphs of Table 6, arranged alphabetically according to their final letters*——

(1) AND ACCORDING TO THEIR INITIAL LETTERS

DA	32	RE	98	EN	111	IS	35
EA	35	SE	49	IN	75	RS	31
LA	28	TE	71	ON	77	AT	47
MA	36	VE	57	CO	41	ET	37
RA	39	TH	78	FO	40	HT	28
TA	28	FI	39	IO	41	NT	82
EC	32	HI	33	RO	28	RT	42
ED	60	NI	30	TO	50	ST	63
ND	52	RI	30	AR	44	OU	37
CE	32	SI	34	ER	87	TW	36
DE	33	TI	45	OR	64	TY	41
EE	42	AL	32	UR	31		
LE	37	EL	29	AS	41		
NE	57	AN	64	ES	54	Total	2, 495

Table 9–C, Concluded.—*The 53 digraphs composing 50% of the 5,000 digraphs of Table 6, arranged alphabetically according to their final letters*——

(2) AND ACCORDING TO THEIR ABSOLUTE FREQUENCIES

RA	39	LE	37	ON	77	IS	35
MA	36	DE	33	IN	75	RS	31
EA	35	CE	32	AN	64	NT	82
DA	32	TH	78	TO	50	ST	63
LA	28	TI	45	CO	41	AT	47
TA	28	FI	39	IO	41	RT	42
EC	32	SI	34	FO	40	ET	37
ED	60	HI	33	RO	28	HT	28
ND	52	NI	30	ER	87	OU	37
RE	98	RI	30	OR	64	TW	36
TE	71	AL	32	AR	44	TY	41
NE	57	EL	29	UR	31		
VE	57	EN	111	ES	54		
SE	49			AS	41	Total	2, 495
EE	42						

Table 9–D.—*The 117 digraphs composing 75% of the 5,000 digraphs of Table 6, arranged alphabetically according to their final letters*——

(1) AND ACCORDING TO THEIR INITIAL LETTERS

CA	20	ND	52	EF	18	SI	34
DA	32	RD	17	OF	25	TI	45
EA	35	BE	18	IG	19	AL	32
HA	20	CE	32	NG	27	EL	29
LA	28	DE	33	CH	14	IL	23
MA	36	EE	42	GH	20	LL	27
NA	26	GE	14	SH	26	OL	19
PA	14	HE	20	TH	78	AM	14
RA	39	IE	13	AI	17	EM	14
SA	24	LE	37	DI	27	OM	25
TA	28	ME	26	EI	27	AN	64
AC	14	NE	57	FI	39	EN	111
EC	32	PE	23	HI	33	IN	75
IC	22	RE	98	LI	20	ON	77
NC	19	SE	49	NI	30	UN	21
AD	27	TE	71	RI	30		
ED	60	VE	57				
		WE	22				

Table 9–D, Contd.—*The 117 digraphs composing 75% of the 5,000 digraphs of Table 6, arranged alphabetically according to their final letters*—

(1) AND ACCORDING TO THEIR INITIAL LETTERS—Continued

CO	41	AR	44	OS	14	YT	15
DO	16	TR	17	IS	35		
FO	40	UR	31	RS	31	AU	13
HO	20	ER	87			OU	37
IO	41	OR	64	AT	47	QU	15
LO	18	PR	18	CT	14		
NO	18	HR	17	DT	15	EV	20
PO	17	IR	27	ET	37	IV	25
RO	28			HT	28	TW	36
SO	15	AS	41	IT	27		
TO	50	SS	19	NT	82	IX	15
WO	19	TS	19	OT	19		
		DS	13	RT	42	TY	41
EP	20	ES	54	ST	63		
OP	25	NS	24	TT	19	Total	3,745

(2) AND ACCORDING TO THEIR ABSOLUTE FREQUENCIES

RA	39	TE	71	TH	78	AM	14
MA	36	NE	57	SH	26	EM	14
EA	35	VE	57	GH	20		
DA	32	SE	49	CH	14		
LA	28	EE	42			EN	111
TA	28	LE	37	TI	45	ON	77
NA	26	DE	33	FI	39	IN	75
SA	24	CE	32	SI	34	AN	64
CA	20	ME	26	HI	33	UN	21
HA	20	PE	23	NI	30		
PA	14	WE	22	RI	30	TO	50
		HE	20	DI	27	CO	41
EC	32			EI	27	IO	41
IC	22	BE	18	LI	20	FO	40
NC	19	GE	14	AI	17	RO	28
AC	14	IE	13	AL	32	HO	20
ED	60			EL	29	WO	19
ND	52	OF	25	LL	27	NO	18
AD	27	EF	18	IL	23	PO	17
RD	17			OL	19	DO	16
		NG	27			SO	15
RE	98	IG	19	OM	25	LO	13

Table 9–D, Concluded.—*The 117 digraphs composing 75% of the 5,000 digraphs of Table 6, arranged alphabetically according to their final letters*

(2) AND ACCORDING TO THEIR ABSOLUTE FREQUENCIES—Continued

OP	25	ES	54	AT	47	QU	15
EP	20	AS	41	RT	42	AU	13
		IS	35	ET	37		
		RS	31	HT	28	IV	25
ER	87	NS	24	IT	27	EV	20
OR	64	SS	19	OT	19		
AR	44	TS	19	TT	19	TW	36
UR	31	OS	14	DT	15		
IR	27	DS	13	YT	15	IX	15
PR	18			CT	14	TY	41
HR	17	NT	82				
TR	17	ST	63	OU	37	Total	3,745

Table 9–E.—*All the 438 different digraphs of Table 6 arranged alphabetically first according to their final letters and then according to their initial letters*

(SEE TABLE 6.—READ DOWN THE COLUMNS)

Table 10–A.—*The 56 trigraphs appearing 100 or more times in the 50,000 letters of Government plain-text telegrams arranged according to their absolute frequencies*

ENT	569	TOP	174	EIG	135
ION	260	NTH	171	FIV	135
AND	228	TWE	170	MEN	131
ING	226	TWO	163	SEV	131
IVE	225	ATI	160	ERS	126
TIO	221	THR	158	UND	125
FOR	218	NTY	157	NET	118
OUR	211	HRE	153	PER	115
THI	211	WEN	153	STA	115
ONE	210	FOU	152	TER	115
NIN	207	ORT	146	EQU	114
STO	202	REE	146	RED	113
EEN	196	SIX	146	TED	112
GHT	196	ASH	143	ERI	109
INE	192	DAS	140	HIR	106
VEN	190	IGH	140	IRT	105
EVE	177	ERE	138	DER	101
EST	176	COM	136	DRE	100
TEE	174	ATE	135		

TABLE 10–B. *The 56 trigraphs appearing 100 or more times in the 50,000 letters of Government plain-text telegrams arranged first alphabetically according to their initial letters and then according to their absolute frequencies*

AND	228	GHT	196	REE	146
ATI	160			RED	113
ASH	143	HRE	153		
ATE	135	HIR	106	STO	202
				SIX	146
COM	136	ION	260	SEV	131
		ING	226	STA	115
DAS	140	IVE	225		
DER	101	INE	192	TIO	221
DRE	100	IGH	140	THI	211
		IRT	105	TEE	174
ENT	569			TOP	174
EEN	196	MEN	131	TWE	170
EVE	177			TWO	163
EST	176	NIN	207	THR	158
ERE	138	NTH	171	TER	115
EIG	135	NTY	157	TED	112
ERS	126	NET	118		
EQU	114			UND	125
ERI	109	OUR	211		
		ONE	210	VEN	190
FOR	218	ORT	146		
FOU	152			WEN	153
FIV	135	PER	115		

TABLE 10–C.—*The 56 trigraphs appearing 100 or more times in the 50,000 letters of Government plain-text telegrams arranged first alphabetically according to their central letters and then according to their absolute frequencies*

DAS	140	DER	101	HIR	106
EEN	196	IGH	140	ENT	569
VEN	190			AND	228
TEE	174	THI	211	ING	226
WEN	153	GHT	196	ONE	210
REE	146	THR	158	INE	192
MEN	131			UND	125
SEV	131	TIO	221		
NET	118			ION	260
PER	115	NIN	207	FOR	218
TER	115	SIX	146	TOP	174
RED	113	EIG	135	FOU	152
TED	112	FIV	135	COM	136

TABLE 10–C, Concluded.—*The 56 trigrams appearing 100 or more times in the 50,000 letters of Government plain-text telegrams arranged first alphabetically according to their central letters and then according to their absolute frequencies*

EQU	114	DRE	100	STA	115
HRE	153	EST	176	OUR	211
ORT	146	ASH	143		
				IVE	225
ERE	138	STO	202	EVE	177
ERS	126	NTH	171		
ERI	109	ATI	160	TWE	170
IRT	105	NTY	157	TWO	163
		ATE	135		

TABLE 10–D.—*The 56 trigraphs appearing 100 or more times in the 50,000 letters of Government plain-text telegrams arranged first alphabetically according to their final letters and then according to their absolute frequencies*

STA	115	IGH	140	TER	115
				HIR	106
AND	228	THI	211	DER	101
UND	125	ATI	160		
RED	113	ERI	109	DAS	140
TED	112			ERS	126
		COM	136		
IVE	225			ENT	569
ONE	210	ION	260	GHT	196
INE	192	NIN	207	EST	176
EVE	177	EEN	196	ORT	146
TEE	174	VEN	190	NET	118
TWE	170	WEN	153	IRT	105
HRE	153	MEN	131		
REE	146	TIO	221	FOU	152
ERE	138	STO	202	EQU	114
ATE	135	TWO	163		
DRE	100			FIV	135
		TOP	174	SEV	131
ING	226				
EIG	135	FOR	218	SIX	146
		OUR	211		
NTH	171	THR	158	NTY	157
ASH	143	PER	115		

Table 11-A.—*The 54 tetragraphs appearing 50 or more times in the 50,000 letters of Government plain-text telegrams arranged according to their absolute frequencies*

TION	218	THIR	104	ASHT	64
EVEN	168	EENT	102	HUND	64
TEEN	163	REQU	98	DRED	63
ENTY	161	HIRT	97	RIOD	63
STOP	154	COMM	93	IVED	62
WENT	153	QUES	87	ENTS	62
NINE	153	UEST	87	FFIC	62
TWEN	152	EQUE	86	FROM	59
THRE	149	NDRE	77	IRTY	59
FOUR	144	OMMA	71	RTEE	59
IGHT	140	LLAR	71	UNDR	59
FIVE	135	OLLA	70	NAUG	56
HREE	134	VENT	70	OURT	56
EIGH	132	DOLL	68	UGHT	56
DASH	132	LARS	68	STAT	54
SEVE	121	THIS	68	AUGH	52
ENTH	114	PERI	67	CENT	52
MENT	111	ERIO	66	FICE	50

Table 11-B.—*The 54 tetragraphs appearing 50 or more times in the 50,000 letters of Government plain-text telegrams, arranged first alphabetically according to their initial letters, and then according to their absolute frequencies*

ASHT	64	HREE	134	REQU	98
AUGH	52	HIRT	97	RIOD	63
		HUND	64	RTEE	59
COMM	93				
CENT	52	IGHT	140	STOP	154
		IVED	62	SEVE	121
DASH	132	IRTY	59	STAT	54
DOLL	68				
DRED	63	LLAR	71		
		LARS	68	TION	218
EVEN	168			TEEN	163
ENTY	161	MENT	111	TWEN	152
EIGH	132			THRE	149
ENTH	114	NINE	153	THIR	104
EENT	102	NDRE	77	THIS	68
EQUE	86	NAUG	56		
ERIO	66			UEST	87
ENTS	62	OMMA	71	UNDR	59
		OLLA	70	UGHT	56
FOUR	144	OURT	56		
FIVE	135				
FFIC	62	PERI	67	VENT	70
FROM	59				
FICE	50	QUES	87	WENT	153

Table 11-C.—*The 54 tetragraphs appearing 50 or more times in the 50,000 letters of Government plain-text telegrams arranged first alphabetically according to their second letters and then according to their absolute frequencies*

DASH	132	THIS	68	EQUE	86
LARS	68				
NAUG	56	TION	218	HREE	134
		NINE	153	ERIO	66
NDRE	77	FIVE	135	DRED	63
		EIGH	132	FROM	59
TEEN	163	HIRT	97	IRTY	59
WENT	153	RIOD	63		
SEVE	121	FICE	50	ASHT	64
MENT	111				
EENT	102	LLAR	71	STOP	154
REQU	98	OLLA	70	RTEE	59
UEST	87			STAT	54
VENT	70	OMMA	71		
PERI	67			QUES	87
CENT	52	ENTY	161	HUND	64
		ENTH	114	OURT	56
FFIC	62	ENTS	62	AUGH	52
IGHT	140	UNDR	59		
UGHT	56			EVEN	168
		FOUR	144	IVED	62
THRE	149	COMM	93		
THIR	104	DOLL	68	TWEN	152

Table 11-D.—*The 54 tetragraphs appearing 50 or more times in the 50,000 letters of Government plain-text telegrams arranged first alphabetically according to their third letters and then according to their absolute frequencies*

LLAR	71	EIGH	132	COMM	93
STAT	54	AUGH	52	OMMA	71
FICE	50	IGHT	140	WENT	153
		ASHT	64	NINE	153
UNDR	59	UGHT	56	MENT	111
				EENT	102
EVEN	168	THIR	104	VENT	70
TEEN	163	THIS	68	HUND	64
TWEN	152	ERIO	66	CENT	52
HREE	134	FFIC	62		
QUES	87			TION	218
DRED	63			STOP	154
IVED	62	OLLA	70	RIOD	63
RTEE	59	DOLL	68	FROM	59

TABLE 11-D, Concluded.—*The 54 tetragraphs appearing 50 or more times in the 50,000 letters of Government plain-text telegrams arranged first alphabetically according to their third letters and then according to their absolute frequencies*

REQU	98	OURT	56	IRTY	59
		DASH	132	FOUR	144
THRE	149	UEST	87	EQUE	86
HIRT	97			NAUG	56
NDRE	77	ENTY	161		
LARS	68	ENTH	114	FIVE	135
PERI	67	ENTS	62	SEVE	121

TABLE 11-E.—*The 54 tetragraphs appearing 50 or more times in the 50,000 letters of Government plain-text telegrams arranged first alphabetically according to their final letters and then according to their absolute frequencies*

OMMA	71	DASH	132	QUES	87
OLLA	70	EIGH	132	THIS	68
		ENTH	114	LARS	68
		AUGH	52	ENTS	62
FFIC	62				
		PERI	67		
				WENT	153
HUND	64	DOLL	68	IGHT	140
DRED	63			MENT	111
RIOD	63	COMM	93	EENT	102
IVED	62	FROM	59	HIRT	97
				UEST	87
		TION	218	VENT	70
NINE	153	EVEN	168	ASHT	64
THRE	149	TEEN	163	UGHT	56
FIVE	135	TWEN	152	OURT	56
HREE	134			STAT	54
SEVE	121	ERIO	66	CENT	52
EQUE	86				
NDRE	77	STOP	154		
RTEE	59			REQU	98
FICE	50	FOUR	144		
		THIR	104		
		LLAR	71	ENTY	161
NAUG	56	UNDR	59	IRTY	59

TABLE 12.—*Average and mean lengths of words*

Number of letters in word	Number of times word appears	Number of letters
1	378	378
2	973	1, 946
3	1, 307	3, 921
4	1, 635	6, 540
5	1, 410	7, 050
6	1, 143	6, 858
7	1, 009	7, 063
8	717	5, 736
9	476	4, 284
10	274	2, 740
11	161	1, 771
12	86	1, 032
13	23	299
14	23	322
15	4	60
120	9, 619	50, 000

(1) Mean length of ~~messages~~ words .... 5.2 Letters.
(2) Average length of messages .... 217 Letters.
(3) Mean length of messages .... 191 Letters.
(4) Mode (most frequent) length .... 105–114 Letters.
(5) It is extremely unusual to find 5 consecutive letters without at least one vowel.
(6) The average number of letters between vowels is 2.

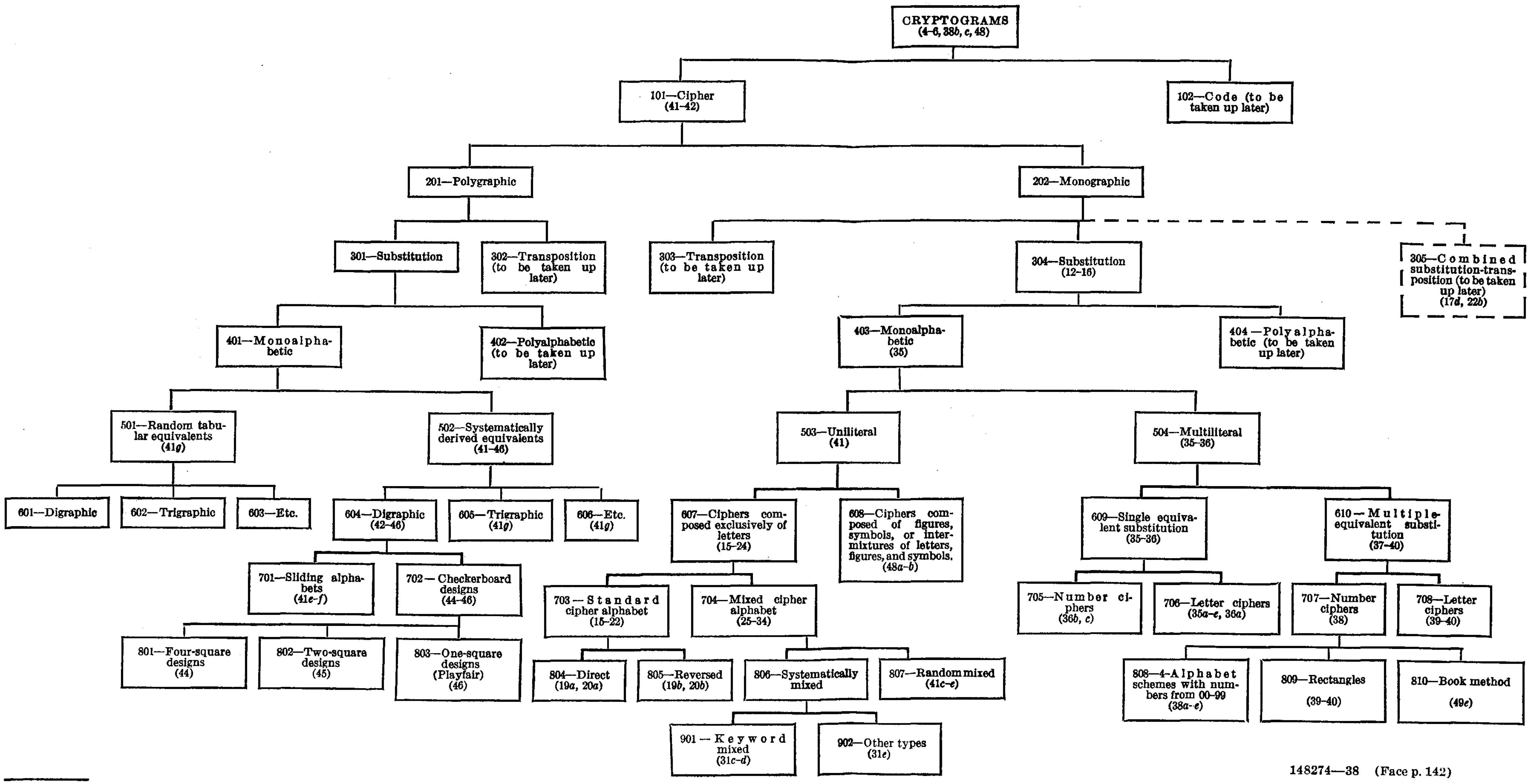
INDEX

INDEX

148274—38——10

ANALYTICAL KEY FOR CRYPTANALYSIS [1]

(Numbers in parentheses refer to paragraph numbers in this text)



[1] For explanation, see Par. 50.

148274—38 (Face p. 142)

U. S. GOVERNMENT PRINTING OFFICE: 1938