# An Application of PTAH to the Voynich Manuscript (U)

BY MARY E. D'IMPERIO

~~*Top Secret Umbra*~~

*(U) This article is the second in a series of studies applying some modern statistical techniques to the problems posed by the Voynich manuscript. This study attempts to discover and demonstrate regularities of patterning in the Voynich text subjectively noted by many earlier students of the manuscript. Three separate PTAH studies are described, attacking the Voynich text at three levels: single symbols, whole "words," and a carefully chosen set of substrings within "words." These analyses are applied to samples of text from the "Biological B" section of the manuscript, in Currier's transcription. A brief general characterization of PTAH is provided, with an explanation of how it is used in the present application.*

*(U) The author draws the following general conclusions: (1) The plain text directly underlying the Voynich text is probably not a natural language written in an alphabet, like English or Latin. (2) The Voynich text probably does not involve any form of simple substitution or alphabetic plain text like English or Latin. (3) The Voynich text probably does not directly represent a variably spelled or "impressionistic" approximation of a natural language like English or Latin, as claimed by Brumbaugh. (4) The words of the Voynich text do not appear to act like code groups in a known code which includes groups for grammatical endings.*

## 1. INTRODUCTION (U)

(U) This article is the second in a series of studies applying some modern statistical techniques to the problems posed by the Voynich manuscript. The first article described an application of cluster analysis and multidimensional scaling [4]. Like that earlier paper, this paper is also intended to serve a tutorial purpose, in explaining how the techniques can be applied to a complex and interesting problem, in the hope of aiding others to apply them in operational contexts. I will not burden the reader with a description of the Voynich manuscript, since I presume most are by now familiar with the general nature of this cryptanalytic puzzle that has come down to us from the late Middle Ages. For any reader desiring more background, I recommend the proceedings of our 1976 Seminar [5], a copy of which may

be obtained from M. D'Imperio, R53/P13. Many readers will also recall the informative and enjoyable presentation by Brigadier John Tiltman on 17 November, 1975.

(U) One of the most frustrating aspects of the Voynich text is its contradictory nature, from the point of view of the analyst. On the one hand, it is highly repetitive, so as to appear at times almost like the "babbling" of many closely similar words in succession (in a manner reminiscent of the refrains of some folk songs or nursery rhymes). This repetitious character has led some to propose that the text might have been generated by some "psychological random" process, as a dummy production to cover some hidden message. Some have even suggested that it may be the product of a mentally disturbed person, who invented the strings of symbols in a form of echolalia, or "speaking in tongues," so that their meaning, if any, is likely to be irrecoverable.

(U) On the other hand, the text has a very clear and consistent structure that is readily apparent to the student as soon as he begins to examine a page. The occurrence of words within lines and symbols within words exhibits the operation of orderly rules, most of which appear to hold throughout the very long and voluminous manuscript, and others of which appear to hold throughout all of certain subsections (as pointed out by Currier, and supported by our cluster analysis results). Certain sequences of symbols recur in similar parts of words consistently; some symbols regularly occupy preferred positions at the beginnings, middles, and ends of words, and at the beginnings and ends of lines; some symbols appear frequently before or after other symbols, and rarely elsewhere. Monographic frequency distributions, regardless of where in the text they are sampled, are very rough. What is more, most symbols retain the same relative frequency of occurrence throughout the lengthy text, with the exception of a few symbols whose frequency seems to vary from subsection to subsection in the "language" contrasts found by Currier. This curious combination of apparently senseless repetition of words with structural regularity of symbols within words poses a very puzzling challenge to the analyst. It is hard indeed to imagine what manner of plain text could be hidden in symbol strings exhibiting these characteristics, if any form of simple substitution is proposed.

(U) William F. Friedman and Brigadier Tiltman have studied the regularity of occurrence of symbols within words in the Voynich text, and have tried to elucidate and exploit the "beginning-middle-end" structure they perceived. A code-like system, with page numbers in sections (all the plant names, parts of the body, star names, etc. together on adjacent pages), might account for the repeated "beginnings" of words. Coded grammatical endings based on Latin, and perhaps including some encipherment of Roman numerals (within their repeated "c" and "i" symbols) might account for the "endings" and "middles." In fact, many code-like systems of this kind were in use by the

Catholic Church during the fifteenth and sixteenth centuries. Early versions of universal or international artificial languages, based on Latin and showing a similar code-like structure, were a favorite preoccupation of scholars in the seventeenth and eighteenth centuries; their ancestry can be traced back to still older mnemonic systems used by the Church and having their ultimate origin in the practices of Roman orators. (For a much more detailed discussion of these topics, see my forthcoming monograph [6]). Friedman and Tiltman hypothesized that an artificial language of this kind might underlie the Voynich text.

~~(C-CCO)~~ I have also found this code or artifical language theory highly attractive as a way of explaining the strange contradictions pointed out above. So far, however, no student has been able to devise a means of confirming or invalidating the theory, or even of clearly demonstrating the intuitively striking regularities of structure in the text. The present study is an attempt to discover and display those regularities, if any, present in a sample of text from one section of the Voynich manuscript, analyzing it at three levels of structure: using single words, and parts of words as units in three separate studies. The statistical tool I chose for the analysis is the PTAH technique of statistical modelling.

## 2. PTAH ~~(C-CCO)~~

~~(C-CCO)~~ PTAH (named for the Egyptian god of wisdom), is a general statistical method developed at IDA (Institute for Defense Analyses), Princeton University. According to Angela Boyter's excellent paper in the *NSA Technical Journal* [2], PTAH got its name when a programmer, Mr. Gerry Mitchell, was listening to the opera "Aida" while working on his program. He was struck by the passage "immenso Ptah noi invociam," and named his program after the Egyptian god. The name was ultimately extended from this program, implementing a particular application of the method, to the method and its mathematical theory as well [2, p 85]. According to [redacted] of R51, the name is pronounced "however you like" [8]. The technique itself and its uses are classified Top Secret Codeword. [redacted] I chose PTAH for the present study for two main reasons: first, because of the applications of PTAH to book codes, and second, because I wished to learn more about PTAH itself [redacted]

(b)(3)-P.L. 86-36

(b)(1)
(b)(3)-50 USC 403
(b)(3)-18 USC 798
(b)(3)-P.L. 86-36

(U) I will make no attempt here to explain "how PTAH works." The documentation seems, with a few exceptions, to fall in two classes: one

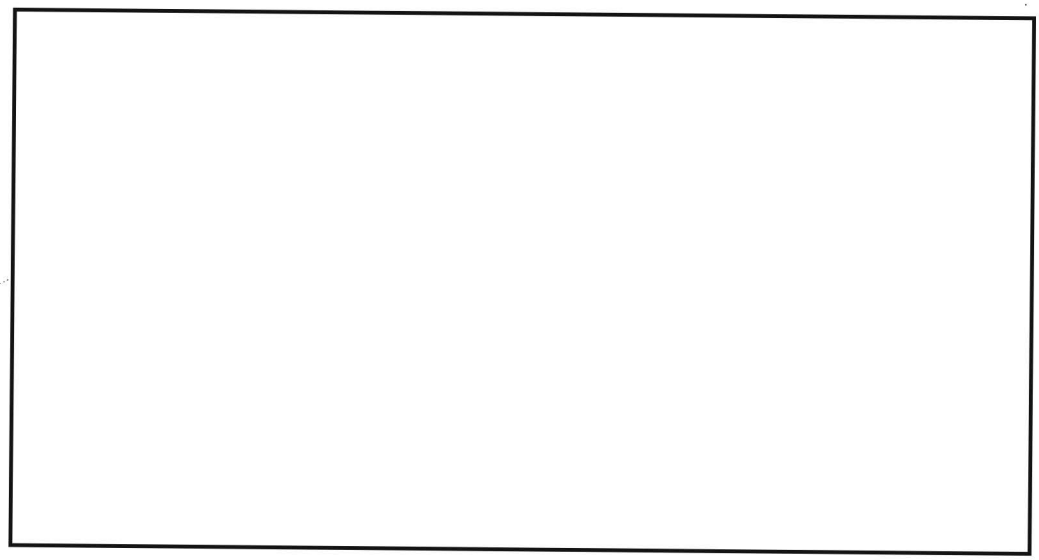
clearly oriented toward mathematicians, and presenting very heavy going indeed for others; and another describing a specific application and providing little or no insight into PTAH itself or the rationale of its use in the given case. As a nonmathematician, I cannot hope to understand the first class of papers on PTAH, let alone attempt to explain their concepts in simple words meaningful to prospective users with an application in mind. Since this article is aimed at such prospective users, I will restrict my remarks on PTAH to a general attempt to characterize the machine runs and analyses that were made in this study, and to provide some flavor or the approach a user might take to his problem and his data in order to prepare input to the PTAH computer programs and interpret their output. I strongly recommend the paper by [redacted] [2], which is a notable exception to my plaint above concerning documentation. The following paragraphs of explanation are based entirely on her clear and helpful exposition. I wish also to express my sincere appreciation for the aid of [redacted] of PI, who made the computer runs in support of this study and assisted me in planning the analyses and interpreting the findings.

(b)(3)-P.L. 86-36

~~(C CCO)~~ The explanations of PTAH provided in the papers for nonmathematicians employ examples involving urns filled with slips of paper on which letters, or some other observable events, are recorded. The PTAH "model" is like a conceptual "machine" whose behavior is adjusted to simulate the observed behavior (as expressed in a long sequence of letters or other unitary events of interest) produced [redacted]

~~(C CCO)~~ [redacted]

(b)(1)
(b)(3)-50 USC 403
(b)(3)-18 USC 798
(b)(3)-P.L. 86-36



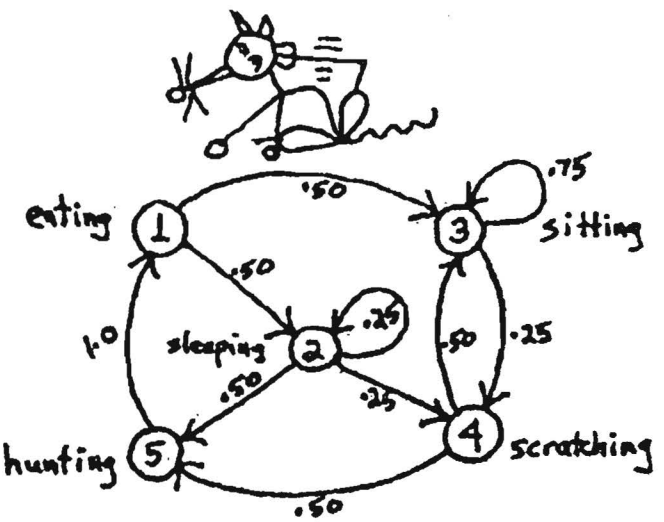
(b)(1)
(b)(3)-50 USC 403
(b)(3)-18 USC 798
(b)(3)-P.L. 86-36

(U) PTAH is iterative, i.e., it cycles repetitively through its calculations until it achieves the best approximation to the events it is simulating, as judged by statistical tests. At the start, we guess at a number of states to try in the model, and arrive at the best number by trying several models of varying sizes and seeing which appears to fit the data best. We provide the programs with a string of text (which may need to be edited to get at arbitrary units other than single letters, *n*-graphs, or words). We prime the program with initial probabilities to start off the two sets of "urns"—the "transition matrix" for the states, and the "output matrix" for the outputs. These initial probabilities can be chosen at random, so long as they add up to 1 for each row of the transition matrix and each column of the output matrix. As the program runs, it changes the probabilities a little on each cycle until the results seem to be converging on the set of values most likely to have produced the input text sample. Having created this statistical "machine" in the form of the two matrices, we could now demonstrate it if we wished by another program which causes it to manufacture new text according to its probabilities. [redacted] provides an example of artificial "English" text produced in this way by a 12-state PTAH model of single letters [2, p. 93]. PTAH is different from other models, such as the digraphic probability model, in that PTAH provides the best model of the data along the entire stream, not just digraph by digraph; it "remembers" more about the system it is simulating.

(U) All of the PTAH program runs made by [redacted] for all three phases of this study contained the following displays: 1) an initial transition matrix representing the starting probabilities for the iterative process; 2) frequency counts of the units (letters, words, word-parts) being studied, ranked by descending frequency; 3) a set of scores for each iteration to aid the PTAH expert in assessing how well the process is converging on an optimum result; 4) transition matrix after a given number of iterations; and

(b)(3)-P.L. 86-36

5) output matrix after that number of iterations. Several intermediate matrices are provided; the results probably of most interest to the user are 6) the final transition matrix, 7) the final output matrix, and 8) several listings of "clusters" at thresholds of decreasing restrictiveness, which show smaller sets of relatively similar elements within those associated with the states.




**State Transition Diagram (U)**

| State | Label | Successor State 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| 1 | Eating | — | .50 | .50 | — | — |
| 2 | Sleeping | — | .25 | — | .25 | .50 |
| 3 | Sitting | — | — | .75 | .25 | — |
| 4 | Scratching | — | — | .50 | — | .50 |
| 5 | Hunting | 1.0 | — | — | — | — |

**Transition Matrix (U)**

# UNCLASSIFIED

**Fig. 1—Behavior of a Mythical Animal (U)**

3. APPLICATIONS TO THE VOYNICH MANUSCRIPT (U)

(U) Researchers have seen apparent regularities in the Voynich text on various levels of structure; patterns have been seen in sequences of single "letters," sequences of words, and sequences of parts within words. Accordingly, with [redacted] aid, I decided upon a three-pronged attack on the Voynich text on these three levels. Each of the resulting three separate studies will be described separately in the paragraphs below, and the findings of each presented.



(U) 3.1 *Analysis of Single Voynich Symbols.* A sample of 3313 consecutive "letters" was chosen from the "Biological B" Section of the manuscript, converted to machinable form by means of Capt. Prescott Currier's transcription. This transcription, as may be seen from Fig. 2a, already includes some combinations of from one to four smaller elements (e.g. "M" = ıı٢, "U" = ıı٧,) which Currier found to be almost invariably linked to form an apparent unit. I chose the "Biological B" pages for my sample because they have been shown (originally by Currier and also in my cluster analysis study) to be more homogeneous and to display a stronger statistical identity than any other section of the manuscript. The extreme roughness of the monographic frequencies is apparent in Fig. 2a. Since Currier and others have found that certain symbols occur more often at the beginnings and ends of words and lines, I included an arbitrary symbol for "end of word" and another for "end of line" in the analysis. Including these, a total of 28 different symbols occurred in the sample, comprising 554 "words" in 67 lines of text. The text sample was fed into the PTAH programs, which generated the frequency counts of symbols ranked in order of descending frequency as an initial step in the analysis.

(U) Figure 2b shows the "Final Transition Matrix" for five states produced by the PTAH programs after 70 iterations. The programs can be set up to produce other numbers of states, depending on the guesses the researcher may have about the structure of his text. In this case, since we knew nothing to start with about the Voynich script and its alphabet, five states were chosen because that number has often proven useful in other studies. Each "state" is associated with one of five subsets of the Voynich script symbols. The decimal numbers in the cells of the matrix are probabilities that the state for that row, and its associated set of symbols, will be followed by each of the states (and sets of symbols) in the columns. The characters assigned by the analysis to each of the states may be seen in Fig. 2b below the matrix. I have also suggested an intuitive verbal label for each state. Each state represents a set of Voynich symbols that seem to act alike in their contacts with other symbols within the text. Figure 2c shows a "state transition diagram" — a graphic representation of the information in the transition matrix. Arrows lead from each state to the other states most likely to follow, and are labelled with the respective probabilities.

| Transcribed symbol | Voynich symbol | Frequency | Rank |
|---|---|---|---|
| l | space | 554 | 1 |
| c |  | 378 | 2 |
| 9 |  | 365 | 3 |
| o |  | 355 | 4 |
| 8 |  | 273 | 5 |
| f |  | 216 | 6 |
| e |  | 191 | 7 |
| a |  | 186 | 8 |
| 4 |  | 181 | 9 |
| z |  | 110 | 10 |
| s |  | 105 | 11 |
| r |  | 98 | 12 |
| ] | line end | 67 | 13 |
| n |  | 64 | 14 |
| p |  | 60 | 15 |
| 2 |  | 32 | 16 |
| x |  | 17 | 17 |
| b |  | 16 | 18 |
| j |  | 12 | 19 |
| m |  | 11 | 20 |
| q |  | 6 | 21 |
| t |  | 5 | 22 |
| d |  | 4 | 23 |
| u |  | 3 | 24 |
| l |  | 1 | 25 |
| v |  | 1 | 26 |
| y |  | 1 | 27 |
| 0 |  | 1 | 28 |

**UNCLASSIFIED**

**Fig. 2a—Monographic Frequencies and Ranks (U)**

(U) With due apologies to any purists, mathematical or otherwise, who may be reading this paper, I will present a frivolous and over-simplified example in an attempt to get across the flavor of the PTAH model, and the import of the matrices and other displays in Fig. 2. Let us imagine an animal that can exhibit five major kinds of activities (or most of whose life can be adequately described in terms of five sets of behaviors). He can eat, sleep, hunt for food, sit still, and scratch for fleas. By counting a long sequence of

actions in the animal's life, we can arrive at an idea of which sets of actions he is likely to do, in which order. If we see him hunting, we know he is most likely to be eating next; after eating, he will either sleep or sit still; after sleeping he will either scratch fleas or start hunting again, and so forth, like an automatic washer going through its cycle. We presume that, underlying these five major sets of common behaviors, the animal has five internal states: an eating, sleeping, hunting, sitting, and scratching state. (Since all we see are his actions, and we cannot get "inside his head," the best we can do in labelling the states is to call them after the strongest or commonest action or characteristic of the event-set associated with each state.) Figure 1 shows a state transition diagram for the "five-state model" of the animal and the "transition matrix" on which the diagram was based.

**State Transition Matrix (U)**

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | .1176077 | .0000000 | .0097794 | .8610576 | .0115553 |
| 2 | .9731800 | .0000000 | .0049215 | .0218984 | .0000000 |
| 3 | .0525056 | .9474944 | .0000000 | .0000000 | .0000000 |
| 4 | .6234602 | .0358985 | .0260699 | .0228168 | 2917546 |
| 5 | .0297188 | .0745607 | .6002830 | .0000000 | .2954375 |
| | Static State Probabilities | | | | |
| | .3520302 | .1124551 | .0921134 | .3098980 | .1335033 |

**State Output Characters and Suggested Labels (U)**

| State | Label | Associated Voynich Symbols |
|---|---|---|
| 1 | "beginners-1," "separators" | word-space, o, a, line-end, v |
| 2 | "enders" | 9, m, t, l, u, 0 |
| 3 | "pre-enders" | 8, x, q |
| 4 | "beginners-2," "post-beg.-1" | f, e, 4, r, n, p, 2, b, j, d |
| 5 | "middles" | c, z, s, y |

**UNCLASSIFIED**

**Figure 2b (U)**

(U) Our way of looking at the "letters" of the Voynich text, in Fig. 2, is similar to our view of the mythical five-state animal. The state transition diagram in 2c shows the probabilities of the different states associated with the letters in 2b. I have shown only those arrows (likely movements from one state to another) with the highest probabilities, leaving out all those under .10 (representing changes to be expected in less than one tenth of the cases). We can summarize the import of Fig. 2 somewhat as follows:

*a.* State 1 is a "beginning" state, including my arbitrary characters for word-end and line-end and certain others that often follow immediately to start a word. It has a high probability of being followed by state 4 (secondary beginners) and most of the remaining time is followed by itself (i.e., a state 1 character following another state 1 character: typically end-of-line or end-of-word, then a word-beginning symbol.)

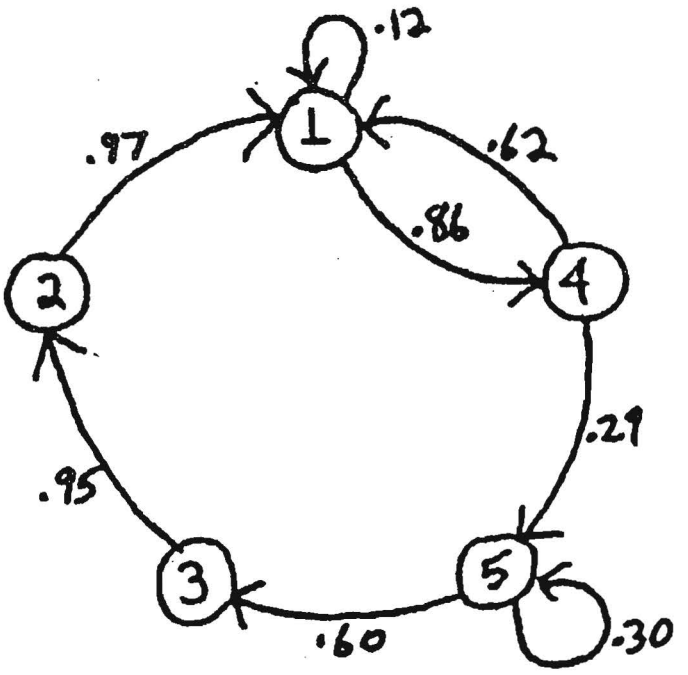
*b.* State 4, a secondary beginning state, is very likely to be followed by state 1 again (representing many common short "words"); otherwise, it is followed by state 5, a "middle" state. The two states 1 and 4 together account for most word-initial patterns of letters throughout the text.

*c.* State 5, the "middle" state, is most likely to be followed by state 3, which I call a "pre-ending" state for reasons to be seen below; otherwise, it is followed by another state 5. This state accounts for the "middles" of words—primarily the sequences "c, cc, ccc, sc, zc" which commonly come between the intial patterns and the "endings" proper.

*d.* State 3 seems to be a "pre-ending", or penultimate state, because it has a .95 probability of being followed by state 2, the "ending" state. (The small number of remaining cases of changes out of state 3 are to state 1, the beginning of a new word, with .05 probability, probably occasioned by the symbol "8" for the most part, which often precedes the ending "9" but sometimes occurs alone at the end of a word.)

*e.* State 2 is clearly an "ending" state, for word-final patterns; it is followed by State 1, the word-initial state, with a .97 probability. It is interesting to note that most of the few remaining cases are transitions to state 4, the secondary word-initial state; I would hazard a guess that these are cases where the text "words" were incorrectly separated in the transcription, so that the word-separator symbol was ommited.

(U) I will not attempt to describe here the output matrix or all the cluster displays for single Voynich symbols. My main use for the output matrix was to identify the letters associated with each of the five states. Involved in the interpretation is the frequency rank of each letter, as well as the probability it has in each state column. A letter which occurred only 5 times in 3313 characters of text, but which had a 1.0 probability of being seen in a given state, may or may not be significant (the letter "Q" might be a somewhat similar case in English, being rare and almost always beginning words). On the other hand, a letter which occurred 500 times in the same text, and had a probability of .8 or .9 for one state is interesting in quite another way.




UNCLASSIFIED

Fig. 2c—State Transition Diagram

A (enders)
l, u, 0, 9

B (beg.-1)
a, o, v, line-end

C (beg.-2)
d, e, j, n, r, 4

UNCLASSIFIED

Fig. 2d—Clusters (threshold = .0010) (U)

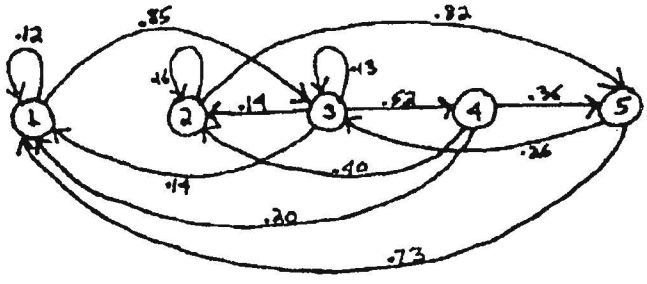
These interpretations are embodied in the list of characters associated with the states in Fig. 2b. The "clusters" were generated by the program by analysis of the values in the final output matrix. Figure 2d shows these for the most restrictive threshold (.001). Their meaning is problematical, and I venture no interpretation. Half of the letters involved are very low in frequency, the other half very high.

(U) *Conclusions for the Single Letter Analysis.* It seems quite clear to me that the view exposed by Friedman and Tiltman concerning the positional structure within Voynich text words is strongly supported by these results. The Voynich symbols do indeed fall into well-defined classes associated with beginnings, middles, and ends of words. In addition, there is a mechanical, regimented quality about the picture we see here—an appearance of surprising orderliness, a highly limited and regular behavior and a resultant degree of predictability. All this seems to me most unlike what one would expect in a simple substitution on any natural language alphabet in running plain text.

(U) Let us compare the situation in Fig. 2 for Voynich symbols to that found by Cave and Neuwirth in a 5-state model for a very large sample of single letters in English text [3]. Figure 3a shows the transition matrix, a state diagram, and the letters assigned to each state. First we note that the diagram contains far more arrows, and has a cluttered look compared to our diagram for the Voynich symbols, made in exactly the same way (leaving out probabilities below .10). The diagram for Voynich symbols shows only eigth significant transitions, while that for English letter shows thirteen. Then we may see that it is much harder to characterize the sets of letters for each state; state 3 seems to concern vowels and "H," and state 5 is for the word spacer alone. The other states are hard to label, and do not relate in any clear and unequivocal way to position within words, except for state 4, which is followed most often by word space and seems to be a word-final state. State 1 contains most of the consonants, and is most often followed by state 3 for vowels and h. Nowhere do we see the positional regularity of beginners to middles to enders to new beginners that is so striking in Fig. 2. For a very complete and interesting analysis of various PTAH models of English, the reader is urged to consult the referenced paper, which is quite readable for the nonmathematician.

(U) The reader may well raise an objection here, pointing out that English is not an inflected language. It makes little use of grammatical affixes (prefixes, endings, etc.) in forming words, as do inflectional languages such as Latin or Russian. Even though the positional structure we have seen in the Voynich symbols looks nothing like that in English letters, might it not look more like the structure in Latin (which is considered by many students to be a likely language to seek in the Voynich text because of its universal use by medieval scholars)? With this reasonable question in mind, I asked [redacted] to make a PTAH run for a five-state model of some Latin text,

**Five-State Model for English Single Letters (U)**
(adapted from Reference 3, p. 10)




UNCLASSIFIED

**State Transition Diagram (U)**

| State | Associated English Letters |
|---|---|
| 1 | t b c j m k p v z w q |
| 2 | s y e d g |
| 3 | a o h i u |
| 4 | n r f l x |
| 5 | word space |

UNCLASSIFIED

**Figure 3a (U)**

4700 characters in length, from *Magia Naturalis*, by Giovanni Battista Porta, 1644 (a work concerned with materia medica, medical spells, natural "wonders," and such matters which seem related to the apparent content of the Voynich manuscript as evidenced by the drawings). In fact, the text I chose contained a series of prescriptions and instructions for preparing and administering herbal recipes to cure various diseases, and so should be closely comparable to the "Biological B" section of the Voynich text.

(U) Figure 3b shows the results of this analysis. While not quite as complex as that for English, the Latin diagram still has a lot more arrows and a much more intricate set of interconnections than that for Voynich symbols (eleven arrows as compared to eight). State 3 is the word separator; state 4 seems to contain many word-final letters which are last letters of common endings (-um, -us, -is, -ur, etc.); and state 1 contains some vowels that form these common endings. State 2 seems to show many word-beginning consonants. State 5 is an odd mixture of vowels "a" and "e," which also enter into common endings, and a conglomeration of odd consonants. While we can see reflections of the grammatical structure of Latin in the state diagram, we can find nothing like the clear positional structure evident in the Voynich symbol diagram of Fig. 2c. We can find little support for an attempt to explain the positional orderliness so clearly apparent in Voynich symbols within text "words" by referring them to Latin prefixes or endings in monographic plain text.

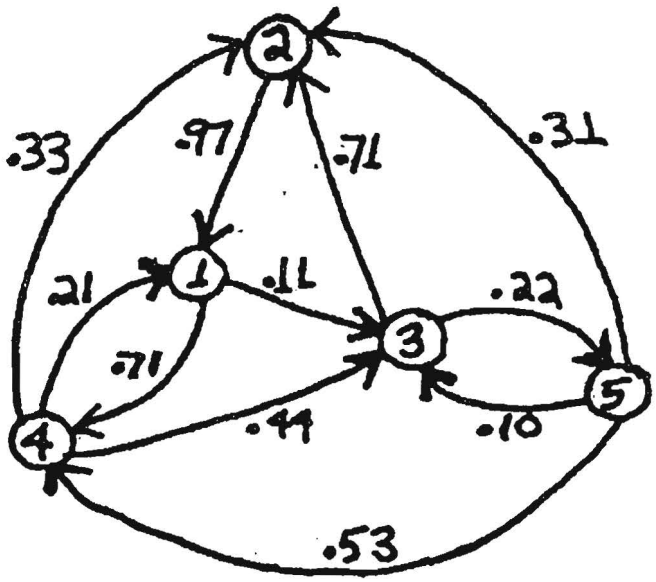
(U) I find the above comparisions quite convincing support for a view that the Voynich text, regarded as a string of single letters, does not "act like" natural language. Instead, it exhibits a clear positional regularity of characters within words. I believe that these findings strengthen the theory of Friedman and Tiltman that an artificial language may underlie the Voynich text.

(U) 3.2 *Analysis of Voynich Text Words.* My second study examines the behavior of whole words in the text, using the presence of spaces and end of line as indications of word separations. (It should be pointed out that the determination of "word" boundaries is often difficult in the Voynich manuscript, and some students have questioned the reliability of spacing as an indication of separate words. The transcription of our text sample, although made with great care by Currier, may have been mistaken as regards word separation in some unknown proportion of cases. The strength of the "beginning" and "ending" states in the first study just described may serve to reassure us that the space is indeed meaningful in separating units of structure, whatever they may be, and that the transcription was accurate for the most part in recognizing the boundries.) A sample consisting of 5567 apparent "words" in 764 consecutive lines was chosen, again from the "Biological B" section of the manuscript in Currier's transcription. A five-state PTAH analysis was run by [redacted] of PI, using "words" as units.

(U) Figure 4a shows the final transition matrix after 100 iterations, and provides a list of the strongest words for each of the five states, with a suggested label or characterization of each state, in terms of the composition of the words and the apparent relationships among the states. Only those words were included which had both frequencies of 10 or higher, and also probabilities greater than .6 of occurring in their assigned state. A state diagram may be seen in Fig. 4b. It is apparent that there are many transitions (13 in all as compared to 8 in Fig. 2c). There are three reciprocal

(b)(3)-P.L. 86-36

Five-State Model for Latin Single Letters (U)




UNCLASSIFIED

State Transition Diagram (U)

| State | Label | Associated Latin Letters |
|---|---|---|
| 1 | Pre-ender vowels? | y u i o |
| 2 | Beginner consonants? | z g v h t c d j |
| 3 | Word Space | word space |
| 4 | Ender consonants? | x m n r l s |
| 5 | ? | f p e b g a |

UNCLASSIFIED

Figure 3b (U)

transitions (state pairs for which state A can lead to state B, but B can also lead back to A again to form a little loop); there is only one such pair in the diagram for single letters. Thus, the diagram for words seems much more complex that that for letters, which is not really too surprising.

**Voynich Text Words: Final Transition Matrix, Iteration 100 (U)**

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | .001312 | .326610 | .445735 | .226324 | .000018 |
| 2 | .000013 | .601243 | .078920 | .319821 | .000003 |
| 3 | .596741 | .000001 | .222830 | .180226 | .000202 |
| 4 | .000000 | .104757 | .002946 | .000000 | .892297 |
| 5 | .550690 | .177433 | .264313 | .000372 | .007192 |

Static State Probabilities (U)

.204763 .274850 .198561 .168628 .153198

**Summary of Outputs and Major Features of States (U)**

| State | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Features | z/s final 9 | f/p, final 89 | final a/o ending | end of line | line initial words |
| Output Words | am | oefcc89 | r | / / / | 8sc89 |
| | zcx9 | 4opc89 | fan | (end | 2or |
| | zq9 | ofc89 | oefan | of | 8zc89 |
| | zcc89 | opc9 | oeor | line | eoe |
| | zcc9 | 4ofc89 | or | symbol) | psc89 |
| | sq9 | opc89 | 4ofar | | zx9 |
| | scf9 | 89 | ofan | | bsc89 |
| | zc89 | ezc89 | ar | | 4ofs89 |
| | zc9 | oefc89 | opam | | 2an |
| | scx9 | 4ofcc89 | oe | | 8an |
| | sccf9 | 4opae | 4ofan | | 2ae |
| | zcf9 | oezc89 | | | |
| | scc9 | | | | |
| | oesc9 | | | | |

# UNCLASSIFIED

**Figure 4a**

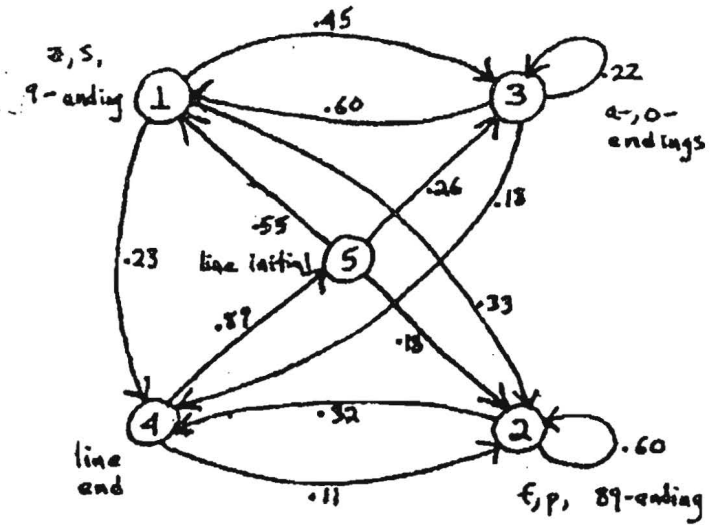





**Fig. 4b—Voynich Text Words: State Diagram (U)**

(U) We may sum up the information in Fig. 4 as follows:

*a.* State 4 is associated with the line-ending symbol. It leads with a probability of .89 to state 5, which seems to consist of line-initial words, and with a probability of .11 to state 2.

*b.* State 5, as we have just seen, appears to be for line-initial words. Half of its high-frequency, high-probability words start with "2" or "8," a feature not seen in the word-lists for any other state. It is followed by state 1 with a .55 probability, state 3 with .22, and state 2 with .18.

*c.* State 3 exhibits a large number of words with "a" and "o" endings (AR, AM, AN, OR, OE); these are rare in the lists for any other state except state 5. It leads to state 1 with a .60 probability, to itself with .22, and to state 4 with .18.

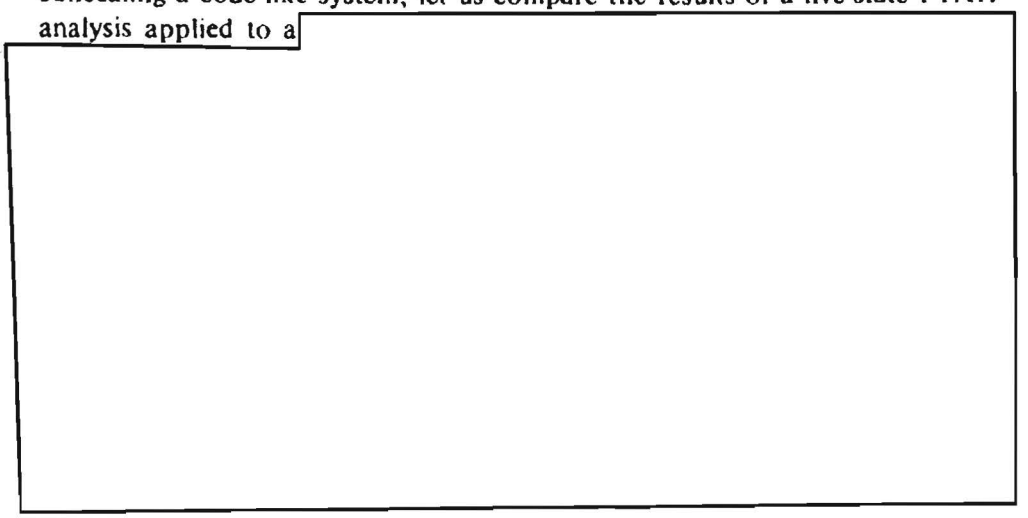
*d.* State 1 shows many words ending in "9," with an initial or central "z" or "s," and a medial "c" or "cc." It leads to state 3 with probability .45, state 2 with .33, and state 4 with .23.

*e.* State 2 appears to involve many words ending in "89," having a central "f" or "p," and a medial "c" or "cc." It is followed by itself with probability .60, and by state 4 with .32.

*f.* The situation for line-final words is not as clear-cut as that for beginnings of lines. State 2 leads to the line-ending state 4 with probability .32, state 1 with .23, and state 3 with .18.

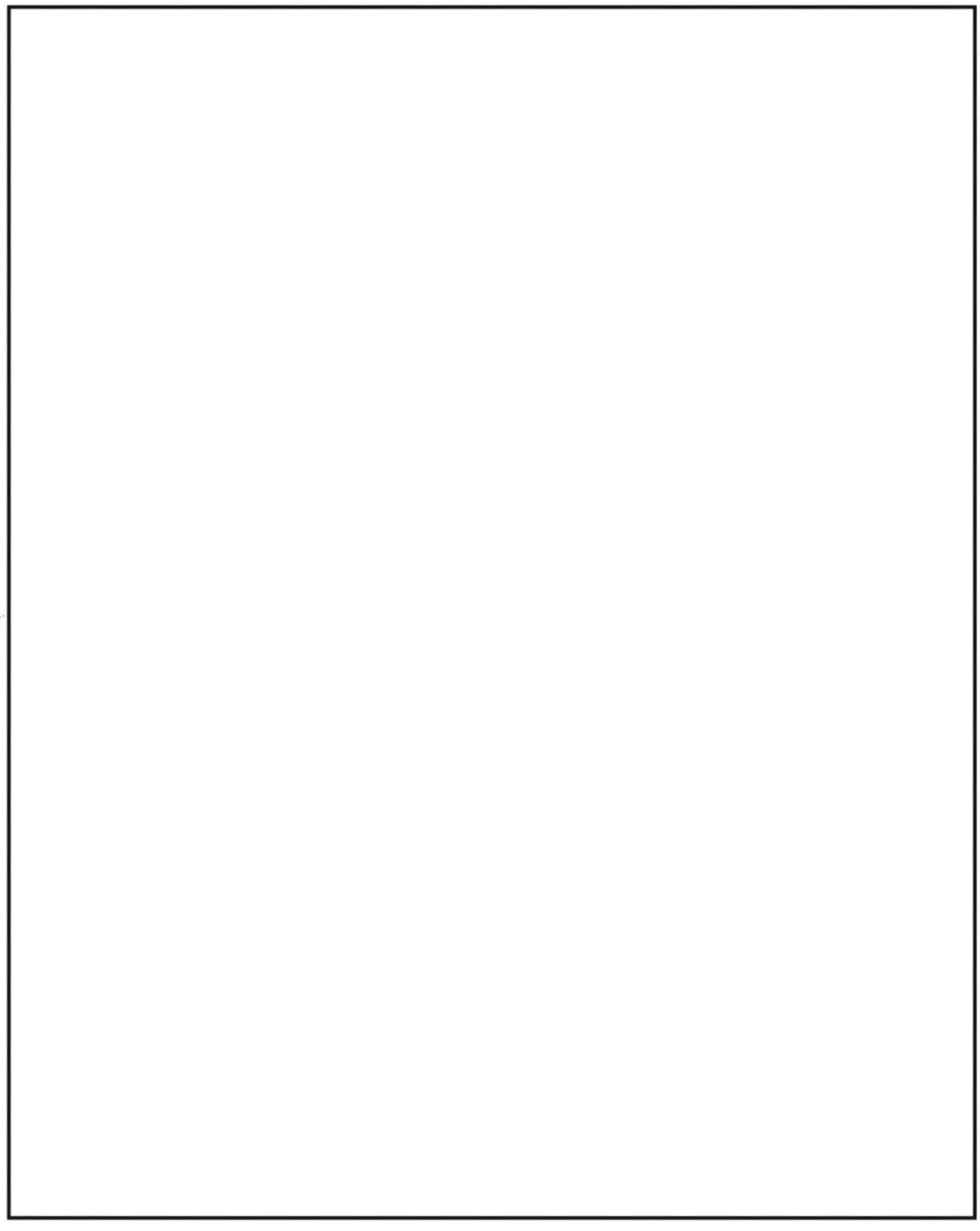
(U) *Conclusions from Analysis of Whole Words.* It seems strikingly clear that there is a positional structure of words within lines in the Voynich text, and that certain sets of words, with characertistic beginnings, middles and endings, are most likely to follow or precede certain other sets of words, with different beginnings, middles and endings. Currier has pointed out these two features of the text [5, pp 65-66]. Our analysis clearly supports both his view of the lines as functional entities, and his finding that words with certain endings were more likely to be followed by words with certain beginnings within a line. This is strange behavior indeed for any running plain text, unless it represents lists of parallel phrases (incantations? instructions? recipes?) in highly stereotyped form. Alternatively, the plaintext units underlying the "words" may not be natural language words but instead numbers or code groups of some sort, subject to some positional constraints. In any case, this curious characteristic of the Voynich text remains to be explained by any would-be decipherer; it does not appear to have been addressed by any of the claims known to me.

~~(TSC)~~ Since one of the theories about the Voynich text views it as possibly concealing a code-like system, let us compare the results of a five-state PTAH analysis applied to a



(b)(1)
(b)(3)-50 USC 403
(b)(3)-18 USC 798
(b)(3)-P.L. 86-36

(U) 3.3 *Analysis of Word parts.* The analysis of repeating patterns of letters within words appears to me to provide the strongest and most interesting results of the three studies. The word parts, which I will call simply "strings" in what follows, were chosen by me on the (admittedly subjective) basis of my own experience on working with large volumes of text over several years, and in accordance with Tiltman's theories on "beginning" and "ending" patterns in words. I made up an initial list of about 50 strings, (shown in Appendix 1), which was used in the first of two PTAH analyses applied to word-parts. In this list I tried to include pairs of symbols that seemed related or similar in form and behavior ("s" and "z," "p" and "f," etc.). I also tried to anticipate and avoid conflicts in the resolution of letter



(b)(1)
(b)(3)-50 USC 403
(b)(3)-18 USC 798
(b)(3)-P.L. 86-36

sequences wherever possible. [redacted] very kindly ran a pre-editing program on the input text to find and isolate all the strings on my list, leaving other character sequences as "left overs" that were also counted as elements in the analysis. To make this clearer, let us imagine that we were "parsing" the English phrase "now/is/the/time/" using strings "no," "is," "/," "the," and "me": the result would be "no w / is / the / ti me /," with ten product strings, two of which ("w" and "ti") are leftovers. As in

(b)(3)-P.L. 86-36

the other studies, word-space and line-end were represented by arbitrary symbols, and I included them explicitly in my string list. A very large volume of text was entered, comprising 13,464 strings, in 3680 words, on 490 consecutive lines of "Biological B" data. A five-state PTAH model was used.

(U) Figure 6 shows the transition diagram and states for the first string list. 225 different elements were isolated in all by the pre-editing program; 99 of these, having frequencies of five or higher, were included in the analysis. The diagram in Fig. 6 shows a surprisingly simple structure, having eight transitions with probabilities of .10 or more. My interpretation can be summed up as follows:

*a.* State 1 is for word-separator and line-ending. It is followed by state 2 with probability .73, and by state 3 with .25.

*b.* State 5 is for word-endings. It leads to the separator state 1 with probability .96.

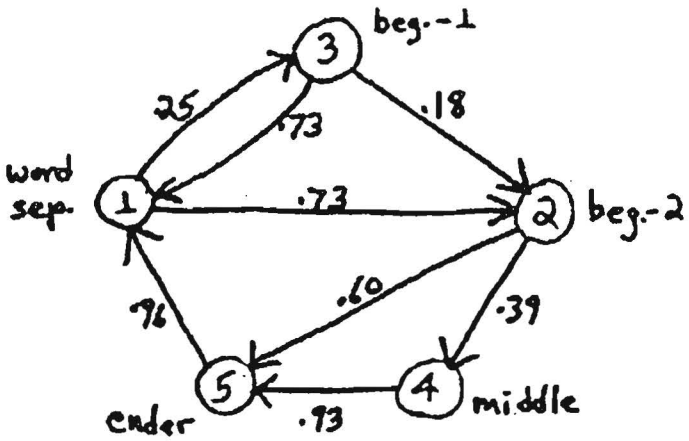
*c.* State 4 is the "middle" state. It exhibits only the special sequences of one, two, or three "c's" in a row, and the related sequence "c8." (It should be remembered that these "c8's" are only those not involved in a sequence of "c" followed by "89," which would have been split in that way.) State 4 is followed by the ending state 5 with probability .93.

*d.* State 3 I call the "beginners-1" state. It shows a special set of beginner strings, many associated with very common short words. It leads to the separator state 1 with probability .73, and to state 2 with .18.

*e.* State 2 is the "beginners-2" state. It produces a large list of strings starting longer words, some few of which can also follow certain of the "beginners-1" strings. It is followed by the ending state 5 with probability .60 and by the middle state 4 with .39.

*f.* Beginnings of words are shown by the successors of the separator state 1. They are, predictably, state 2 with probability .73 and state 3 with .25 (the two "beginner" states).

(U) In examining the complete list of strings produced by the pre-editing program and the "cluster" lists found in the PTAH analysis, I was struck by the recurrence of certain strings in the "leftovers." I collected an additional list of possible word parts to be added to the list, and also included all single letters, to force complete decomposition of "leftovers" in a new study. Appendix 2 shows the new characters and sequences. Since the pre-editing program looked for the longest matches first, the additions should have the effect of greatly shortening the list of elements in the study. With the new list, the elements found in the text should comprise only those sequences specified plus single symbols, producing a much more complete analysis. The same text was input to the pre-editing step with this new list of word parts, and a new set of PTAH runs was made. Input text consisted of 13,410 strings, in 3152 words, on 421 consecutive lines. It may be seen that a smaller volume of text was required to produce about the same number of




| State | 1 | 2 | | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Label | word sep. | begin-ner-2 | | begin-ner-1 | middle | ender |
| Output Strings | / (word space) | p | sc | 4oe | cc | c9 |
| | | 8z | of | zc9 | c | 89 |
| | | 9z | zc | o | ccc | 9 |
| | | bz | z | sc9 | c8 | an |
| | | 4of | f | 4o | | am |
| | | op | bs | b | | aj |
| | | 4op | sx | oe | | |
| | | es | ef | | | |
| | | ez | 8s | | | |
| | | ps | s | | | |
| | | sq | 4 | | | |
| | | 8 | | | | |

## UNCLASSIFIED

**Fig. 6—State Diagram of Voynich Text Strings: First List (U)**

strings as in the first study, due to the more complete decomposition into shorter elements. Also, in contrast to the 225 different elements found in the first word-part analysis, only 81 unique elements were produced, with only 72 having frequencies of 5 or higher.

(U) Figure 7 shows the state diagram, list of states, and some "clusters" of similar elements found by the program at its most restrictive threshold (.005 for this run). The diagram is a bit more complicated, and some of the

states have been renumbered, but the five states are basically similar with respect to the associated output strings. There are ten major transitions, compared to eight in the previous study, and two pairs of states are linked by reciprocal transitions. The state diagram in Fig. 7 was deliberately constructed so as to facilitate comparison with Fig. 6; for the most part, there is surprisingly little essential change. The main differences are the following:

*a.* The strings associated with the states have been slightly altered, in ways that seem to me to improve their consistency and to bring them even closer in line with what I expected, based on my subjective "feel" for recurrent units in the text. More elements that I guessed might be similar are together in the same state, and few if any that seemed well placed in the first word-part results have been lost in the second.

*b.* There is a new cycle of reciprocating transitions between "beginners-2" and "middles," reflecting the curious linking behavior of the common "c" sequences.

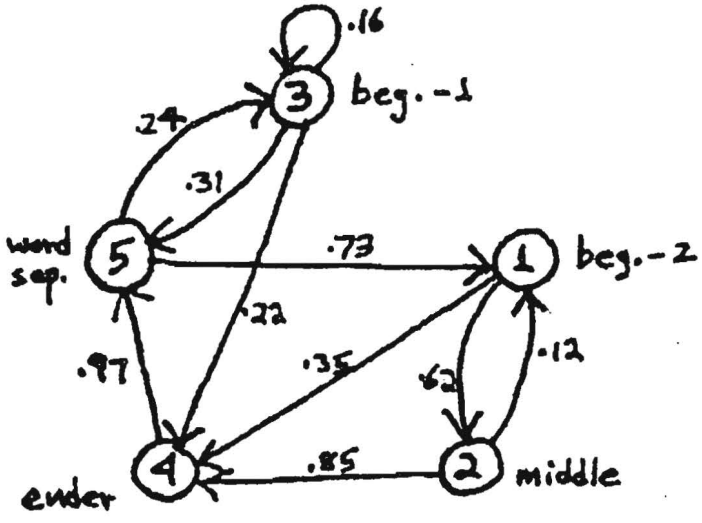
*c.* There is a new transition from "beginners-1" directly to "enders" and the arrow from "beginners-1" to "beginners-2" has disappeared. This appears to reflect the better separation of common short words from longer words.

*d.* The "beginners-1" state has a new, relatively low-probability transition to itself, probably occasioned in part by the inclusion of the single symbols "r" and "e" in its output set.

*e.* In general, far more of the information in the text has been utilized, and the "noise" from the many "leftovers" in the first analysis has been removed (at the possible risk of adding a different source of "noise" in the single symbols).

*f.* The "clusters" in Fig. 7 are smaller sets of word parts which the program found to be especially similar. They were generated by the program through an analysis of the final output matrix and comparison of the probabilities there. This list of twelve tight clusters is striking in contrast to a list of twenty-eight much more diffuse and multiply-intersecting clusters produced by the first word-part study at the same threshold value (.005). Striking parallelisms between symbols that look alike will be apparent to anyone familiar with the Voynich text (e.g., b, f, and p all followed by s and z; "4of" and "4op"; "of" and "op," etc.).

(U) *Conclusions from the Word-Part Studies.* I find these analyses even more convincing in confirming the highly regular positional structure of elements in Voynich text "words." In addition, these results suggest that the meaningful elements are not words as wholes, or single letters, but larger, variable-length sequences of symbols. Early codes and ciphers in use by the Catholic Church show many instances of such mixed-length elements (single symbols and two- or three-letter units intermixed, some standing for plaintext letters and some for common words and phrases). It is interesting to note




| State | 5 | 1 | | 3 | 2 | 4 |
|---|---|---|---|---|---|---|
| Label | word sep. | begin-ner-2 | | begin-ner-1 | middle | ender |
| Output Strings | / (word space) | 4of | op | 4oe | c | aj |
| | | z | of | 2o | cc | 89 |
| | | 4op | f | e | a | an |
| | | es | x | r | ccc | 9 |
| | | bs | q | o | | am |
| | | ez | 8 | 4o | | ar |
| | | sx | 8s | b | | 3 |
| | | ef | 9z | oe | | m |
| | | 8z | bz | or | | ad |
| | | s | sq | o8 | | |
| | | p | vs | z9 | | |
| | | oef | 4oef | | | |
| | | 9f | ps | | | |
| | | 4 | 2z | | | |



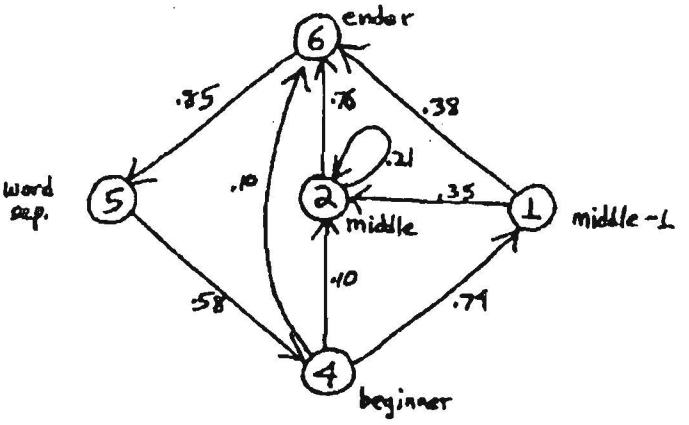
Fig. 7—State Diagram of Voynich Text Strings: Second List (U)

that the number of unique word parts found in the second study (81) is very close to that required if Voynich text elements were assumed to stand for plaintext consonant-vowel syllables after the fashion of a syllabary. A 16-consonant list appropriate for Latin (b, c, d, f, g, j, l, m, n, p, q, r, s, t, v, x) in combination with the five vowels (a, e, i, o, u) would provide 80 syllabic symbols. Of course, some convention would be required for the representation of closed syllables and consonant clusters, but this problem is readily solved in many known syllabaries (Japanese phonetic representations for foreign words, for example). It is interesting to speculate that the "ligatured" symbols in the Voynich script might stand for Latin consonant clusters; a similar ligaturing approach to clusters is used in the Devanagari syllabary of India.

(U) We are fortunate in having a PTAH study made by [redacted] applying PTAH to the symbols of a known syllabary: the "Linear B" syllabary used in Greece and the Aegean Islands around the middle of the second millenium B.C. [7]. This writing system, originally thought by many to embody the records of the Minoan civilization, was deciphered in 1953 by Michael Ventris and John Chadwick to reveal an early form of Greek, similar to that of the Homeric epics. Thus, it provides us with a very interesting parallel to the situation I have hypothesized above: a language involving consonant clusters and closed syllables, written down in a syllabary designed for a language having only open (VC) syllables. Figure 8 shows a state diagram adapted from the phonetic portions of the seven-state PTAH model on page 35 of the reference. (I urge the interested reader to examine this highly readable and informative paper in its entirety.) I have omitted the two states for numeral signs and ideographic signs, leaving a set of five states for word-divider and vowel-consonant syllables which may be compared to our five word-part states for the Voynich text.

(b)(3)-P.L. 86-36

(U) The diagram for "Linear B" phonetic signs shows nine transitions, with a clear positional structure very like what we have seen in Figs. 6 and 7. Word-separator is followed by "beginners"; these are followed by "middles-1" or "middles-2"; either of the two "middle" states can lead to the other or to the "ender" state, which in turn leads to word separator. While I will not attempt to make too much of this comparison, and offer it only for its suggestive value, it is still quite striking. When we recall how different the English and Latin five-state models for single letters and the five-state model for code groups appeared, the similarity between Figs. 6, 7 and 8 seems to support a guess that short plaintext word parts may underlie the Voynich script. The distribution of word lengths in the text provides additional support: few words are as long as seven or eight symbols (and these often contain the medial "c" sequences), while many common words are only three, four, or five symbols in length. This picture is quite unlike that in Latin or English written in an alphabet of single letters, where the




UNCLASSIFIED

**Fig. 8—State Diagram of Five-State Model for "Linear B" Syllabary (U)**
(adapted from Reference 7, p. 35)

range of word lengths includes many of ten to fifteen characters or more, and there are a great number and variety of seven- and eight-letter words.

## 4. SOME GENERAL CONCLUSIONS (U)

(U) In closing, I will state some conclusions that I have drawn from these analyses. At the risk of appearing overly positive, and alienating some other students who are convinced that they have found the secret of the Voynich manuscript, I will couch these statements in a relatively unequivocal form.

(U) 4.1 The plain text directly underlying the Voynich text is probably not a natural language represented by an alphabet of single letters like the English alphabet. A PTAH five-state model for single letters of an agglutinative language such as Turkish would provide an additional interesting test.

(U) 4.2 As a corollary, the encryption or concealment system in the Voynich text probably is not any form of simple substitution on an alphabet of single letters like the English alphabet.

(U) 4.3 The Voynich text probably does not represent a natural language, written in an "impressionistic" way (to recall a statement by Dr. Robert Brumbaugh, who claims to have deciphered it as a misspelled, distorted form of Latin), nor can its characteristics be explained by hypothesizing many

variant spellings of the same words in an alphabetic writing system (cf. older forms of English). Its structure seems far too ruly and regular to accord with these views. Rather than a distorted or degraded form of English or Latin monographic structure, it seems to exhibit a DIFFERENT structure of its own.

~~(TSC)~~ 4.4 If the Voynich text conceals a code, it is not very like the example examined above in section 3.2 (a code involving a partially inflected Romance language comparable in some ways to and descended from Latin, and a code in which grammatical endings were represented by code groups: a situation I had considered to be quite close to that called for by Friedman's and others' guesses about artificial languages underlying the Voynich text.)

(U) My intention here is not to attack other students, or to "put down" their opinions; rather, it is to stimulate new research. I have no thought of "clearing the field" for some cherished claim of my own about the Voynich text; I wish to emphasize the fact that I have no single "pet" theory about the manuscript. As others also have said, it is hard to imagine any directly underlying natural language plain text whose characteristics can explain the phenomena adequately. My hope is that this paper, if it has no other impact, will at least provoke some others to approach the puzzle of the Voynich manuscript with some of the modern scientific tools at our disposal, in addition to the intuitive and subjective methods chosen so predominantly by earlier researchers.

5. REFERENCES (U)

[1] [redacted]

[2] [redacted]

[3] R. L. Cave and L. P. Neuwirth, "Hidden Markov Models of English," *IDA-CRD Working Paper 239* (January 1969). (U)

[4] M. E. D'Imperio, "An Application of Cluster Analysis to the Question of 'Hands' and 'Languages' in the Voynich Manuscript," *PI Informal No. 3.* (June 1978, S-216,867) and *NSA Technical Journal*, Vol. XXIII, No. 3 (Summer 1978), pp. 59-75. (U)

[5] M. E. D'Imperio, "New Research on the Voynich Manuscript: Proceedings of a Seminar" (Washington, D.C. 30 November 1976). (U)

[6] M. E. D'Imperio, "The Voynich Manuscript: An Elegant Enigma" (National Security Agency/Central Security Service, 1978). (U)

[7] J. Ferguson and H. E. Kulsrud, "Statistical Studies on Linear B," *IDA-CRD Working Paper 441* (January 1975). (U)

[8] [redacted]

**Appendix 1: First List of Voynich Text Strings (U)**

| | |
|---|---|
| word-sep. | cc |
| line-end | ccc |
| 2o | ef |
| 2oef | ep |
| 2of | es |
| 2z | ez |
| 4o | fs |
| 4o8 | fz |
| 4oe | o8 |
| 4oef | oe |
| 4of | oef |
| 4op | oep |
| 89 | of |
| 8s | oj |
| 8z | op |
| 98 | or |
| 9f | ps |
| 9z | pz |
| ad | rz |
| ae | sq |
| aj | sx |
| am | vs |
| an | z9 |
| ar | z9f |
| at | |
| bs | |
| bz | |

**UNCLASSIFIED**

**Appendix 2: Additions for Second String List (U)**

| | | | |
|---|---|---|---|
| a3 | sc | zcp | zf |
| a6 | scf | zcq | zp |
| au | scp | zcx | zv |
| 92 | scq | zcb | zb |
| 29 | scx | zcv | sf |
| 9p | scb | zq | sp |
| 9q | scv | sq | sv |
| 9x | zc | zx | sb |
| 9s | zcf | sx | rs |

**UNCLASSIFIED**