Word Structure in the Voynich Manuscript

Jorge Stolfi
Institute of Computing, Univ. of Campinas
13083-970 Campinas, SP - Brazil
stolfi@dcc.unicamp.br

Abstract

We give here a paradigm (combinatorial description) of 'typical' words from the Voynich Manuscript (VMS), namely a fairly restrictive grammar whose language contains 95% the word occurrences of the manuscript (■% of all distinct words). We also give frequency counts for the various components of the typical word, as defined by the model. The paradigm is shown to hold, with similar component frequencies, not only for words from all sections, but also for the figure labels.

1 Introduction

The Voynich manuscript (VMS) is an ancient medical/astrological treatise, written in an unknown script or code which has resisted decipherment for nearly four centuries. This baffling manuscript has became a vexing challenge for cryptologists and paleographers, amateur and professional alike. The analysis of its bizarre text raises several interesting problems in statistics and computational linguistics as well — such as, how can we tell whether there is a meaningful message to be decoded?

The text of the VMS is composed of discrete symbols, and is clearly divided into word-like symbol groups by fairly distinct spaces. It has long been known that those *Voynichese words* have a non-trivial internal structure, manifested by constraints on the sequence and position of different symbols within each word. This note describes new structural paradigm for Voynichese words, that is significantly more detailed and comprehensive than previous models.

The nature and complexity of the new paradigm, and its fairly uniform fit over all sections of the manuscript (including the labels on illustrations), are further evidence that the text has significant contents of some sort. Moreover, the paradigm imposes severe contraints on possible decipherment theories. In particular, it seems highly unlikely that the text is a Vigenère-style cipher, or was generated by a random process, or is a simple transliteration of an Indo-European language. On the other hand, the paradigm may be compatible with a codebook-based cipher (like Kircher's universal language), an invented language with systematic lexicon (like Dalgarno's), or a non-European language with largely monosyllabic words.

In section 2 we summarize the history of the manuscript; in sections 3–?? we describe the known features of the book and its script. In section ?? we look more closely at the structure of words, and, in section ??, we describe the new word model which is the main topic of this paper. ⋆[Confirm.]

2 A brief history

The manuscript is named after the Russian-American antiquarian W. Voynich, who acquired it in 1912 from from the library of a Jesuit college near Rome. The book now resides in Yale's Beinecke Library, under catalog number MS 408 [?, ?]. Nothing definite is known about its author and place of origin. Based on stylistic and material evidence, the book is believed to have been written in the late 15th or early 16th century, within the European cultural sphere; but even these meagre conclusions cannot be trusted, since the book may well be an European copy of an older and more exotic original.

The documented history of the manuscript has now been traced back to Prague in the 17th century [?]. Its earliest confirmed owner was Georg Baresch — an otherwise obscure alchemist, to whom the book was already a baffling mystery [?, ?]. We also have a faint scribble in the margin of the cover page which is believed to be the signature of Jakub Horčicky de Tepeneč (1575-1622), in Latin Jacobus Sinapius, chief physician of Emperor Rudolf II of Bohemia (1552-1612). What we know of Jacobus's life and background makes him an unlikely author, but a plausible owner of the manuscript prior to Baresch [?].

For the book's history before Jacobus, our only clue is a cover letter found attached to the manuscript, from Charles University's rector J. M. Marci to the Jesuit scholar A. Kircher in Rome [?, ?]. That letter, dated 1665, does not mention Jacobus, but quotes a claim by Marci's friend R. Mnishovsky that the manuscript once belonged to Rudolf, who believed it to be a Roger Bacon original.

Although Marci himself declared that he was "suspending his judgement" on the matter, the *Bacon hypothesis* was taken quite seriously by Voynich. Working under that assumption, he identified the English scholar John Dee (1527-1608), as the person most likely to have carried the VMs to Prague [?]. This hypothesis had some strong arguments in its favor: Dee himself was a foremost collector of Bacon manuscripts, was extremely interested in cryptography, alchemy, and occult sciences, owned several books written in mysterious alphabets, and lived in Bohemia from 1584 to 1588 and made friends with several members of Rudolph's court.

Voynich's Bacon/Dee hypothesis was widely accepted until a few years ago, and led many would-be decipherers to assume that the undelying language of the VMS was Latin, or possibly medieval English [?]. Unfortunately, experts in Bacon's work flatly reject the possibility that he was the VMS author [?]; and no mention of the VMS has been found in Dee's quite detailed diaries. Thus, although Rudolf (who was indeed an avid collector of arcana) may well have owned the manuscript, and may have believed it to be Bacon's, there is no significant evidence that the manuscript came from England, or that John Dee had anything to do with it. The Bacon/Dee hypothesis having thus been discredited, we

are now left any clue about the origin and language of the manuscript.

Over the last 80 years, several people have claimed to have deciphered the VMS, and found it to contain all sorts of material — from Khazar diplomatic correspondence in early Ukranian [?], to Cathar death rituals in a French-German pastiche [?]. Unfortunately, all these "solutions" leave so much freedom to the reader (by assuming a lossy encoding scheme, and/or a lost dialect, and/or highly variable spelling) that they could be used to extract equally (im)plausible contents from any random string of symbols. Most serious students of the manuscript reject those solutions, and still regard the VMS "code" as a complete mystery.

Good (if somewhat dated) introductions to the VMS puzzle and its history can be found in the books by M. D'Imperio [?] and D. Kahn [3], and in several magazine and newspaper articles [?, ?, ?, ?]. A more detailed and up-to-date account, available through the Internet, is being maintained by R.Zandbergen [?]. James Reeds has collected an extensive bibliography [?], that already lists several books and over a hundred articles devoted to the VMS. Reproductions of the manuscript can be bought from Beinecke Library, and selected page images are available at their internet site [?] as well as in many of the publications cited above.

Interest in the manuscript has grown considerably over the last decade, after digital transcriptions of the text became freely available [?, ?, ?]. At present, most of the known VMS research efforts are being carried out by an informal study group, scattered over the globe, commnicating through an electronic mailing list created and maintained by J. Gillogly [?].

3 The book

The Voynich manuscript measures about 16 by 23cm when closed. It consists of about 58 sheets of prepared calfskin (*vellum*), of various sizes, folded into 116 leaves (*folios*). Some of the leaves are oversize, and fold out to display 2, 3, 4, or 6 physical pages (*panels*) on each side. All together, the book contains 265 panels. The vellum sheets are gathered into 20 nested sets (*quires*) containing from one to 6 sheets. A detailed description of the folio sequence and quire structure was compiled by J. Reeds [?].

We know that the book was re-bound at least once after it left the hands of its author; and it is quite obvious that some of the sheets were bound in the wrong order. The quires and folios are numbered — but the numbers must be apocryphal, since they agree with the current (wrong) physical order. Gaps in the numbering do reveal, however, that at least 14 folios have been lost. In fact, some of those missing folios appear to have been cut away from the already bound book.

The standard VMS page numbering scheme, which we follow in this report, is based on the folio numbers penned on the manuscript itself, suffixed with 'r' for recto and 'v' for verso. The multiple panels of fold-out pages are identified by an additional digit suffix, starting with 1 at the panel next to the binding gutter and increasing outwards. Thus, for example, page `f70v2` is a part of the back side of folio 70, which is a fold-out leaf — specifically, the second panel away from the bound edge.

3.1 Handwriting style

Almost every page contains some text, and most pages are illustrated with freehand pen drawings or diagrams, some of them quite complex. Sometimes the contents of a logical page extends across a fold, spanning two or more adjacent panels.

Magnification of the text shows that the writing ink was applied with a split pen or quill, with a squarish nib, held with the right hand and somewhat tilted relative to the page's vertical edges — all very typical of documents from that epoch. The book was examined in 1942 by A. H. Carter, a handwriting expert, who stated quite confidently that the entire text was the work of a single person, who probably also penned the figure outlines.

On the other hand, US Navy cryptographer P. Currier discovered in 1960 that large sets of pages with apparently similar contents could be partitioned into two sets with very different word distributions, which he named "language A" and "language B." Currier further claimed that each set was in a visibly different handwriting, but this subjective claim does not seem to be widely shared among VMS investigators.

A possible resolution for these conflicting views, which seems to be supported by later statistical analyses [**?**, **?**], is that the two subsets in question were written by the same person but on two separate occasions. The book was almost certainly composed over a period of several months or years (the text and ink drawings alone must have required several hundred man-hours of work, exclusive of research and planning); so it is quite conceivable that the the author's vocabulary, style, and handwriting evolved through the project, enough to explain the differences seen by Currier.

Recently, S. Toresella — an expert in medieval herbals — observed a strong resemblance between the Voynichese script and the *humanistic hand*: a rounded, upright writing style, that was popular in Europe for a few decades around 1500, before being displaced by the slanted and compact italic hand [**?**]. This rather tenuous connection is actually the best clue we have as to the date of the manuscript.

3.2 Colors

The only instances of colored writing are two oversize symbols on the first page (f1r), and small amount of text (a single line, and a single ring around the diagram) on page f67r2 — both in red ink.

On the other hand, most figures have been colored, with a wide variety of paints and instruments. The colors often seem to have been chosen rather casually, either for their decorative value, or according to simple conventions. On page f16v, for example, we see a plant which had its star-shaped leaves painted red, and its leafy flower painted green. Moreover, the paint was often applied rather crudely, with little regard to the penned outlines.

The a sloppiness of the fill-in painting stands in contrast to the care that was obviously invested in the text and penned figure outlines. It is quite possible, therefore, that some of the fill-in paints (if not all of them) were applied by later owners; and we should be wary of any intrepretations of the figures that are based on their colors. These doubts could

perhaps be resolved by a careful exhamination of the original; and a scientific analysis of the paints, inks, and stains may be able to provide some useful clues.

3.3 The sections

Although the illustrations are quite unusual and difficult to interpret, they allow us to assign almost every page to one of six quite distinct classes, according to its contents:

- *herbal*: a plant drawing, and a couple of paragraphs of text.
- *cosmological*: a diagram — usually circular and divided into sectors, often showing stars, the sun, or the moon — surrounded by rings of text.
- *zodiacal*: a circular diagram, having at its center a pictorial symbol from the zodiac, surrounded by two or three rings of text and bands of stars (either 15 or 30 per page), each with a short label and flanked by a tiny female figure.
- *pharmaceutical*: two or three short paragraphs, alternating with rows of pictures of leaves and roots, some of them labeled.
- *biological*: a long text, apparently continuous across page boundaries, flowing around one or more illustrations. These show many small female figures bathing in bizarre assemblies of tubs and conduits, some of them resembling body organs.
- *starred-items* (or *recipes*): several dense paragraphs of text, each marked with a star-like "bullet" in the left margin, without any illustrations.

These page classes are conventionally called *sections*. It must be stressed that the section names above are merely conventional labels for superficially homogeneous but dissimilar subsets of the pages. In particular, the true contents of the pharmaceutical, biological, and starred-items sections is essentially unknown.

Some VMS investigators distinguish a separate *astronomical* section, consisting of those cosmological pages that contain obvious depictions of the sun, moon, and stars. In addition, there are a few isolated pages without illustrations, usually at section boundaries, whose classification is uncertain; we have chosen to bundle them together into the *unknown* pseudo-section.

Table 1 lists the pages traditionally assigned to the major sections. As the table shows, some sections — in particular, herbal and pharmaceutical — actually consist of two or more blocks of consecutive pages, separated by material belonging in other sections. Moreover, while most sections seem to be fairly homogeenous with respect to Currier's language classification, the herbal pages can be split into two subsets on that basis, which are labeled `hea` and `heb` in the table. (Although the two subsets are presently interleaved and scattered all over the manuscript, it turns out that the four pages in the same vellum sheet are always in the same language. Therefore, the scrambling may well be the result of improper binding by a later owner.)

Section		Sbsec.	Size		Page list
			Pages	Symbs.	
herbal (A)	hea	hea.1	84	27931	f1v(21)f11v, f13r(26)f25v, f27r(8)f30v, f32r+v, f35r(8)f38v, f42r+v, f44r(4)f45v, f47r+v, f49r, f51r(8)f54v, f56r+v.
		hea.2	10	3783	f87r+v, f90r1(4)f90v1, f93r+v, f96r+v.
herbal (B)	heb	heb.1	26	12755	f26r+v, f31r+v, f33r(4)f34v, f39r(6)f41v, f43r+v, f46r+v, f48r+v, f50r+v, f55r+v, f57r, f66v.
		heb.2	6	2471	f94r(6)f95v1.
cosmological	cos	cos.1	1	454	f57v.
		cos.2	14	7966	f67r1(14)f70r2.
		cos.3	4	4597	f85r2, f86v4, f85v2, f86v3.
zodiacal	zod	zod.1	12	6562	f70v2(12)f73v.
biological	bio	bio.1	20	31415	f75r(20)f84v.
pharmaceutical	pha	pha.1	6	4581	f88r(6)f89v1.
		pha.2	10	7189	f99r(10)f102v1.
starred-items	str	str.1	2	3438	f58r+v
		str.2	23	52179	f103r(12)f108v, f111r(11)f116r
unknown	unk	unk.1	1	833	f1r.
		unk.2	1	623	f49v.
		unk.3	1	195	f65r+v.
		unk.4	1	1471	f66r.
		unk.5	1	1621	f85r1.
		unk.6	1	2261	f86v6.
		unk.7	1	1707	f86v5.
		unk.8	1	8	f116v.
missing	xxx	xxx.1	2	–	f12r+v.
		xxx.2	14	–	f59r(12)f64v, f74r+v.
		xxx.3	4	–	f91r(4)f92v.
		xxx.4	4	–	f97r(4)f98v.
		xxx.5	4	–	f109r(4)f110v.

Table 1: The main sections of the Voynich manuscript. The notation 'f1v(21)f11v' means '21 consecutive logical pages, from leaf 1(verso) to leaf 11(verso), inclusive'. Section `xxx` comprises those pages that are known to have been lost. The symbol counts are approximate (see section **??**).

4 The Voynichese script

The most striking feature of the book is its script, which bears no visible relation to any known writing system in the world, living or extinct — and must therefore be an original invention of the author. See figure 1.

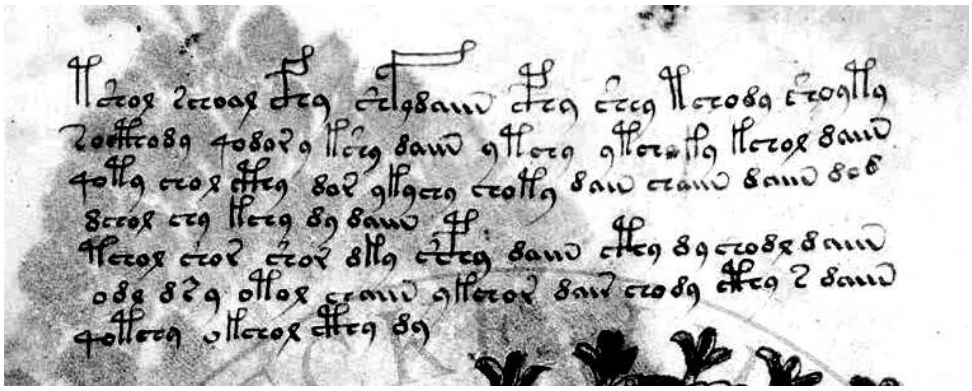


Figure 1: A sample of the VMS script (page f11r). Courtesy of Yale's Beinecke Library ■(get permission!).

Most of the continuous text consists of paragraphs, like those shown in figure 1, spanning the usable width of the page — with a fairly even margin on the left, a more ragged one on the right, and a left-justified partial line at the bottom. Some text is incorporated into diagrams, either in circular bands (almost always clockwise, usually starting near the 10 o'clock position), or along radial lines (outwards or inwards). Many figures have short *labels* written next to them.

The layout of the main text strongly implies that it was written in lines from left to right, top to bottom; a conclusion that is confirmed by observing how the ink density varies along a line, and how the spacing between charactes varies next to figures or vellum defects. It is obvious also that, in most cases, the text was written after the illustrations had already been drawn — or at least sketched.

4.1 Glyphs, tokens, and words

The pen strokes are fairly clear and deliberate — i.e. "printed" rather than cursive. The strokes are obviously organized into *glyphs*, where each glyph consists of a few connected pen strokes, usually separated from adjacent glyphs. The glyphs are laid out horizontally on top of an imaginary baseline, with occasional ascenders and descenders — much in the way of modern Roman letters. The vast majority of the glyphs seem to be instances of a fairly small repertoire of discrete symbols.

The glyphs in turn are clearly grouped into word-like segments by *interword spaces* that are noticeably wider than the normal gaps between consecutive glyphs. Following standard parsing nomenclature, we will call those text segments *tokens*, and use *word* to mean an

abstract sequence of symbols, independently of its occurence in the text. Thus the sentence "the man can open the can" contains six tokens but only four words.

The glyph statistics of line breaks are fairly similar to those of inter-word spaces, suggesting that lines were generally broken at word boundaries. The same observation applies to gaps in the text due to intruding illustrations. Like most medieval manuscripts, the VMS contains no obvious punctuation marks; thus, even though each paragraph is a single sequence of words, we cannot assume that it is a single sentence.

4.2 The basic glyphs

Most of the text symbols seem to be instances of the 22 glyphs listed in table 2.

glyph	*in tokens*		*in words*		*glyph*	*in tokens*		*in words*	
	18799	.1168	4823	.1204		10779	.0670	1993	.0498
	23689	.1472	6176	.1542		13538	.0841	3438	.0858
	16837	.1046	3745	.0935		5133	.0319	739	.0185
	10057	.0625	2815	.0703		12467	.0775	3002	.0750
	7105	.0442	1934	.0483		2405	.0149	987	.0246
	5577	.0347	900	.0225		1053	.0065	399	.0100
	10433	.0648	2820	.0704		4335	.0269	1133	.0283
	9371	.0582	2092	.0522		5560	.0346	1485	.0371
	883	.0055	227	.0057		918	.0057	231	.0058
	365	.0023	277	.0069		1317	.0082	673	.0168
	73	.0004	55	.0014		205	.0013	107	.0027

Table 2: The 22 basic glyphs of the Voynichese script, with their occurrence counts and relative frequencies in the text and in the lexicon.

Many of these symbols occur isolated, in contexts that seem to be letter enumerations, or labels in list items. On the basis of these and other clues, it seems safe to assume that the glyphs listed in table 2 are indeed the primary 'combinatorial elements' of the script.

4.3 Major glyph classes

The basic glyphs of table 2 are traditionally classified by their shape into a few classes. The glyphs  , , , and are traditionally called *gallows*, and the corresponding forms ,

 ,  , and  are said to stand on *platforms*. We will refer to  and  as the *benches* (respectively with and without *plume*), and to  ,  ,  , and as the *leaders* (because of their codes in the EVA alphabet, `l d r s`; see appendix [?]). We'll also call the *initial* glyph,  the *final* glyphs, and  the *circles*. Finally, we'll refer to and as the *stick* and *crescent* glyphs.

As we shall see later on, this classification is strongly correlated with the occurrence patterns of those glyphs in the text. Therefore, is almost certain that the symbols were not assigned at random, but according to some system; and that the morphological classes above have some linguistic value.

4.4 Rare glyphs

In addition to the "ordinary" glyphs of table 2, there are a hundred or so rare signs that occur only a few times in the whole text, most of them only once, such as

J. Reeds has compiled an exhaustive list of these *weirdos* [?], which by an large seem to be deformed variants or condensations of the basic glyphs above. Table 3 shows the only weirdos that occur frequently enough to qualify as possible letters.

glyph	*in tokens*	*in words*	*glyph*	*in tokens*	*in words*
	96	78		35	26
	3	3		1	1
	31	17		23	12
	7	5		13	10
	32	25		24	19
	4	4		6	6
	1	1		2	2
	2	2		2	2

Table 3: Some rare glyphs of the Voynichese script, with their occurrence counts in the text and in the lexicon.

Note the substantial gap between the frequencies of the basic glyphs of table 2 and the weirdos of table 3, which provides a convenient cutoff point. (Although the basic glyph occurs less often than the weirdo , the former is clearly part of the 'gallows with platforms' series, which has about about 2000 occurrences in total.)

It may turn out that the symbols of table 3, and perhaps a few additional ones, are indeed rare but otherwise normal symbols of the script — like æ in English. In particular, The *picnic table* glyph ϖ (35 occurrences, exclusively in the cosmological, starred-items, and herbal-B sections) behaves pretty much like the basic glyph ϱ (over 10,000 occurrences); and glyph ϛ (96 occurrences) seems to be a relative of glyph ϟ (over 1,000 occurrences). However, the other weirdos — most of which occur only once, often in special contexts like tables and diagrams — are more likley to be special symbols (like our $), abbreviations, slips of the pen, or embellished versions of the common letters above.

In any case, we have chosen to exclude most of the weirdo glyphs from the alphabet, and omit any words containing them from the text files used in our analyses. Given the extreme rarity of those symbols, this simplifying decision should not have a significant impact on the decipherment efforts.

4.5 Borrowed symbols

Although the glyph set on the whole is quite original, the general appearance of the script strongly suggest that it was inspired in European calligraphic models. Some Voynichese glyphs, such as o, a, c, are identical to Roman lowercase letters. The glyphs ϟ and τ are similar to the letters s and t in some medieval hands; and the glyph ϑ was a standard scribal abbreviation for the common Latin ending -us. These and other letter shapes also resemble some cryptographic alphabets of the time [?]. Even the characteristic gallows glyphs bear some resemblance to exaggerated and embellished ascenders used by some scribes in earlier epochs [?].

Unfortunately, these resemblances haven't provided any useful clues for decipherment, or even for locating the author at a specific time or place. The glyphs in question have fairly simple and natural shapes, so the resemblances could be simple coincidences. Most VMS scholars agree that, even if the inventor of the script did copy those symbols from existing alphabets, he probably borrowed the shapes without regard for their meaning.

4.6 Glyph structure

Except for ϑ, the basic Voynichese glyphs are combinations of a few simple pen strokes, drawn from a very limited repertoire:

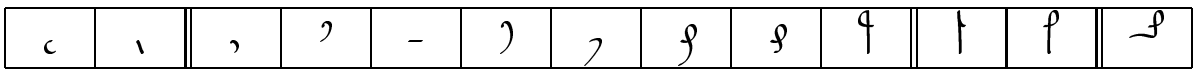
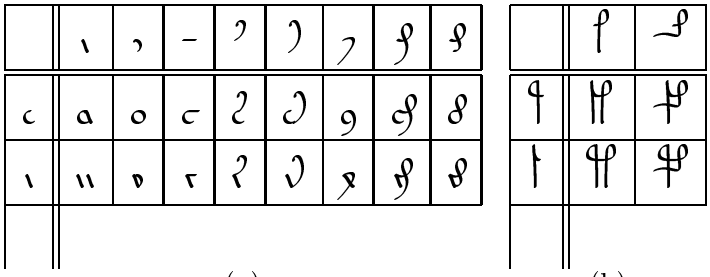


Table 4: A set of pen strokes that combine to form most of the essential Voynichese glyphs.

In particular, the strokes {ϥ, ϩ} combine with {ϱ, ϟ} in all possible ways to produce the four gallows. Also, most combinations of the strokes {c ı} with {ı , ϩ - ϩ ϩ ϱ ϱ} result in valid glyphs.



(a) (b)

Table 5: Combinations of two basic strokes that produce valid Voynichese glyphs.

Of all combinations in table 5, only ◦ does not seem to occur in the manuscript; all others occur at least a few times. The benches {ᓚ ᘐ} and the platform gallows {ᖴ ᖵ ᖶ ᖷ} are combinations of three or more of the basic strokes above. Conversely, the only glyph that does not seem to fit in the above schema is ᔭ.

This "combinatorial" structure of glyph shapes may be due solely to aesthetics and/or efficiency reasons. Namely, the author may have picked a small set of simple strokes, enumerated all combinations of two strokes, and assigned these to the alphabet, in some arbitrary order. People who devise new cipher alphabets will often follow this approach, consciously or unconsciously.

However, the shape of a glyph seems to have significant correlations with their statistical properties — an observation wich seems important, but whose implications are still obscure. This question will be discussed in more detail in section 4.10.

4.7 The question of the true alphabet

It must be stressed that the glyphs of table 2 may not be the true symbols of the Voynichese script, as understood by the VMS author. It is quite possible that, in the *true Voynichese alphabet*, some of those glyphs are only parts of letters, or composites of two or more letters. This uncertainty must be kept in mind when the text is subjected to statistical analysis.

Some hints about the true symbol boundaries could be obtained in principle by analyzing the glyph statistics around forced gaps in the text — line breaks, intruding figures, and vellum defects. However, most of those gaps seem to be ordinary word gaps, and (for reasons that will become clear later on) they give us little information about symbol boundaries within words.

Another potential source of hints are the so-called *key sequences* — about half a dozen lists of isolated glyphs, vertical or circular, found at several places in the book. Unfortunately, the interpretation of these lists is quite problematic. For one thing, no two of these lists contain the same set of symbols. Also, several glyphs that are common in the main text do not occur in any list, and vice-versa. For these and other reasons, some of these lists are suspected of being apocryphal, possibly working notes by a later owner or student of the VMS.

One must keep in mind, furthermore, that the set of letters commonly used for enumeration or labeling purposes need not match the language's alphabet. To prove this point, it

suffices to consider the classical Roman and Greek number systems (which used a subset and a superset, respectively, of the corresponding alphabets); and the fact that the German letters ü and ß are hardly ever used as enumeration tags in German texts.

In any case, we have convincing evidence that the glyphs of table 2 are *not* the true Voynichese alphabet. For instance, the EVA glyphs s and c almost never occur as independent letters, but only as parts of larger groups such as ow or ckhc. In particular, the pair cc behaves like ch and sh in many respects, and may well be a single letter of the true alphabet. Moreover, the glyphs ⫯ and ⫯ occur mostly in the first line of each paragraph; for that reason, they are suspected to be fancier variants of ⫯ and ⫯, respectively. Likewise, the glyph g often occurs in line-initial position, where it may be a calligraphic variant of o.

On the other hand, there is evidence suggesting that the glyphs ⫯ and ⫯, which so far have been considered equivalent by all VMS investigators (and were denoted by the same code in all available transcriptions), are in fact different symbols; and ditto for ⫯ and ⫯.

Anyway, in spite of all difficulties and unknowns, there is substantial agreement among VMS analysts that the 'true' Voynichese symbol set must have a couple dozen distinct symbols at most; so we are probably dealing with an alphabetic script, where each symbol corresponds roughly to one element (phoneme) of the spoken language.

4.8 Digraph statistics

The statistical properties of Voynichese, viewed as a sequence of discrete symbols, have been extensively analyzed over the last 50 years [**?**, **?**, 2, 10, **?**, **?**]. The counts of digraphs (consecutive glyph pairs) in the VMS (main text and labels) are shown in tables 6 and 7, respectively for tokens (taking word frequencies into account) and for words (ignoring the word frequencies). The symbol □ denotes a word boundary; see section 5.

	□	4	ᛁ	ᚳ	ᚪ	ᚬ	ᚹ	ᚾ	ᛇ	ᛏ	ᛞ	ᛠ	ᛢ	ᛣ	ᛥ	ᛰ	ᛱ	ᛲ	ᛳ	ᛴ	ᛵ	ᛶ	ᛷ	tot
□	·	5102	5	91	1934	8077	1777	3563	1219	5753	3142	1279	457	4	14	1156	958	119	526	193	502	33	126	36030
4	·	·	·	48	6	5031	7	·	2	3	·	2	·	·	·	15	1	1	1	9	6	·	1	5133
ᚬ	1125	20	211	342	261	62	128	2170	427	165	70	5411	2587	6	162	5802	3658	140	552	198	140	16	36	23689
ᚳ	86	·	3	4784	425	3214	3899	4873	366	150	45	5	16	10	5	412	181	42	67	135	71	3	7	18799
ᚩ	22	·	2	4779	454	2562	977	790	79	10	12	54	20	·	6	151	76	4	30	238	131	11	25	10433
ᛞ	26	·	·	2515	134	949	273	177	19	14	4	10	3	·	·	50	20	2	·	90	44	2	3	4335
ᛁ	10	·	4427	2	1	7	3	15	18	7	5	41	714	5444	62	10	4	1	4	1	3	·	·	10779
ᚪ	46	·	6107	5	3	10	13	47	40	6	3	3055	3241	111	780	30	10	1	6	7	9	3	5	13538
ᚹ	14795	8	·	7	18	33	2	173	41	259	99	31	15	·	3	683	556	20	84	5	4	·	1	16837
ᚷ	5873	2	·	51	391	517	484	432	152	665	265	29	33	·	7	994	88	32	36	·	2	3	1	10057
ᚻ	5565	·	10	16	671	333	277	34	3	118	40	12	2	·	·	19	1	1	1	1	·	1	·	7105
ᚾ	1147	1	·	42	592	360	107	24	1	73	26	4	1	1	1	16	1	2	2	1	3	·	·	2405
ᛇ	5518	·	·	1	10	13	22	8	·	1	·	1	·	1	1	1	·	·	·	·	·	·	·	5577
ᛈ	1018	·	·	·	9	7	6	6	·	3	1	·	1	·	2	·	·	·	·	·	·	·	·	1053
ᛏ	602	·	5	106	3943	450	6683	21	16	332	160	77	12	·	10	29	5	·	8	5	3	·	·	12467
ᛉ	75	·	6	3744	2833	700	730	9	5	1030	205	33	1	·	·	·	·	·	·	·	·	·	·	9371
ᛊ	57	·	2	1712	1468	693	463	18	4	959	170	12	·	·	·	1	·	·	·	·	·	1	·	5560
ᛋ	20	·	·	3	69	56	26	3	1	170	17	·	·	·	·	·	·	·	·	·	·	·	·	365
ᛒ	27	·	1	4	188	217	64	30	3	712	70	·	·	·	·	1	·	·	·	·	·	·	·	1317
ᛗ	7	·	·	253	31	99	455	33	4	1	·	·	·	·	·	·	·	·	·	·	·	·	·	883
ᛗᛗ	11	·	·	206	70	229	364	28	4	1	1	1	2	·	·	·	1	·	·	·	·	·	·	918
ᛗᛗᛗ	·	·	·	19	6	14	28	5	1	·	·	·	·	·	·	·	·	·	·	·	·	·	·	73
ᛗᛗ	·	·	·	69	21	56	49	8	·	1	·	·	·	·	·	1	·	·	·	·	·	·	·	205

Table 6: Occurence counts for consecutive pairs of basic glyphs in the VMS (main text and labels).

	□	4	۱	c	a	o	g	8	2	ɔ	ƨ	ɤ	ʔ	ɔ	ƪ	ƪ	ƭ	ƭ	ƨƨ	ƨƨ	ƨƨ	ƨƨ	tot
□	·	710	4	59	269	1726	579	549	296	963	479	331	108	1	2	300	324	82	266	44	80	19	55	7246
4	·	·	·	36	5	656	6	·	2	3	·	2	·	·	·	14	1	1	1	7	4	·	1	739
o	294	18	90	234	124	44	80	782	192	118	56	1311	651	6	70	850	733	100	235	76	68	15	29	6176
c	44	·	3	1304	208	1067	627	775	193	94	43	5	13	10	5	178	103	39	48	37	21	2	4	4823
ɔ	17	·	2	1088	172	701	292	237	56	9	11	24	11	·	4	50	42	4	25	31	25	10	9	2820
ƨ	12	·	·	515	42	259	110	85	15	13	4	7	2	·	·	25	12	2	·	16	10	2	2	1133
۱	9	·	746	2	1	7	3	11	17	7	5	34	262	825	41	10	4	1	4	1	3	·	·	1993
a	35	·	1125	5	3	10	12	42	36	6	3	952	834	56	256	27	9	1	6	5	8	2	5	3438
g	2803	8	·	7	11	31	2	123	36	100	53	29	15	·	3	259	187	17	52	4	4	·	1	3745
ɤ	898	2	·	46	199	262	220	229	103	249	119	24	20	·	7	316	62	27	26	·	2	3	1	2815
ʔ	1071	·	9	15	317	206	129	30	3	86	32	10	2	·	·	19	1	1	1	1	·	1	·	1934
ɔ	442	1	·	38	172	154	70	20	1	46	14	4	1	1	1	13	1	2	2	1	3	·	·	987
ɔ	848	·	·	1	8	13	18	7	·	1	·	1	·	1	1	1	·	·	·	·	·	·	·	900
ƨ	365	·	·	·	9	6	6	6	·	3	1	·	1	·	2	·	·	·	·	·	·	·	·	399
8	315	·	5	89	952	182	1111	19	15	142	54	54	11	·	7	27	5	·	7	4	3	·	·	3002
ƭ	27	·	6	769	452	252	159	6	5	301	98	16	1	·	·	·	·	·	·	·	·	·	·	2092
ƭƭ	25	·	2	415	270	261	102	17	4	285	92	10	·	·	·	1	·	·	·	·	·	1	·	1485
ƪ	15	·	·	3	55	49	25	3	1	110	16	·	·	·	·	·	·	·	·	·	·	·	·	277
ƪ	16	·	1	4	109	139	42	25	3	281	52	·	·	·	·	1	·	·	·	·	·	·	·	673
ƨƨ	5	·	·	76	16	47	66	12	4	1	·	·	·	·	·	·	·	·	·	·	·	·	·	227
ƨƨ	5	·	·	62	29	64	47	14	4	1	1	1	2	·	·	·	1	·	·	·	·	·	·	231
ƨƨ	·	·	·	14	5	10	20	5	1	·	·	·	·	·	·	·	·	·	·	·	·	·	·	55
ƨƨ	·	·	·	41	10	30	19	5	·	1	·	·	·	·	·	1	·	·	·	·	·	·	·	107

Table 7: Occurence counts for consecutive pairs of basic glyphs in the Voynichese lexicon (main text and labels, ignoring word frequencies).

Tables 6 and 7 reveal an interesting feature of Voynichese: repeated letters are almost non-existant.

4.9 Glyph classes

Tables 8 and 9 give the the relative frequencies of each glyph in the sample, as a function of the preceding and following glyph, respectively.

	□	ᔭ	ᔑ	ᓀ	ᓂ	ᓄ	ᓇ	ᓈ	ᓐ	ᓯ	ᓰ	ᓱ	ᓲ	ᓴ	ᓵ	ᓶ	ᓷ	ᓸ	ᓹ	ᓺ	ᓻ	tot
□	.	.14	.	.	.05	.22	.05	.10	.03	.16	.09	.04	.	.	.	.03	.03	.	.	.	.	.	.	1.0
ᔭ	.	.	.	.	.	.98	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1.0
ᓄ	.05	.	.	.	.	.	.	.09	.02	.	.	.23	.11	.	.	.24	.15	.	.02	.	.	.	.	1.0
ᓂ	.	.	.	.25	.02	.17	.21	.26	.02	.	.	.	.	.	.	.02	.	.	.	.	.	.	.	1.0
ᓯ	.	.	.	.46	.04	.25	.09	.08	.	.	.	.	.	.	.	.	.	.	.	.02	.	.	.	1.0
ᓰ	.	.	.	.58	.03	.22	.06	.04	.	.	.	.	.	.	.	.	.	.	.	.02	.	.	.	1.0
ᓀ	.	.	.41	.	.	.	.	.	.	.	.	.	.07	.51	.	.	.	.	.	.	.	.	.	1.0
ᓇ	.	.	.45	.	.	.	.	.	.	.	.	.23	.24	.	.06	.	.	.	.	.	.	.	.	1.0
ᓈ	.88	.	.	.	.	.	.	.	.	.02	.	.	.	.	.	.04	.03	.	.	.	.	.	.	1.0
ᓵ	.58	.	.	.	.04	.05	.05	.04	.02	.07	.03	.	.	.	.	.10	.	.	.	.	.	.	.	1.0
ᓱ	.78	.	.	.	.09	.05	.04	.	.	.02	.	.	.	.	.	.	.	.	.	.	.	.	.	1.0
ᓲ	.48	.	.	.02	.25	.15	.04	.	.	.03	.	.	.	.	.	.	.	.	.	.	.	.	.	1.0
ᓴ	.99	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1.0
ᓶ	.97	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1.0
ᓐ	.05	.	.	.	.32	.04	.54	.	.	.03	.	.	.	.	.	.	.	.	.	.	.	.	.	1.0
ᓷ	.	.	.	.40	.30	.07	.08	.	.	.11	.02	.	.	.	.	.	.	.	.	.	.	.	.	1.0
ᓸ	.	.	.	.31	.26	.12	.08	.	.	.17	.03	.	.	.	.	.	.	.	.	.	.	.	.	1.0
ᓹ	.05	.	.	.	.19	.15	.07	.	.	.47	.05	.	.	.	.	.	.	.	.	.	.	.	.	1.0
ᓺ	.02	.	.	.	.14	.16	.05	.02	.	.54	.05	.	.	.	.	.	.	.	.	.	.	.	.	1.0
ᓻ	.	.	.	.29	.04	.11	.52	.04	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1.0
ᓼ	.	.	.	.22	.08	.25	.40	.03	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1.0
ᓽ	.	.	.	.26	.08	.19	.38	.07	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1.0
ᓾ	.	.	.	.34	.10	.27	.24	.04	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1.0

Table 8: Next-symbol probabilities for basic glyphs in the VMS text. The table should be read by rows; i.e., the value '.21' in row ᓄ and column ᓈ means that 21% of the occurrences of ᓄ in the text are followed by ᓈ.

	□	4	ı	c	a	o	9	8	2	cc	2c	9	2	3	9																
□	.	.99	.	.	.14	.34	.11	.29	.51	.55	.72	.13	.06	.	.	.12	.17	.33	.40	.22	.55	.45	.61						
4	.	.	.	.	.	.21	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.						
o	.03	.	.02	.02	.02	.	.	.17	.18	.02	.02	.54	.36	.	.15	.62	.66	.38	.42	.22	.15	.22	.18						
c	.	.	.	.25	.03	.14	.23	.39	.15	.	.	.	.	.	.	.04	.03	.12	.05	.15	.08	.04	.03						
cc	.	.	.	.25	.03	.11	.06	.06	.03	.	.	.	.	.	.	.02	.	.	.02	.27	.14	.15	.12						
2c	.	.	.	.13	.	.04	.02	.	.	.	.	.	.	.	.	.	.	.	.	.10	.05	.03	.						
ı	.	.	.41	.	.	.	.	.	.	.	.	.	.10	.98	.06	.	.	.	.	.	.	.	.						
a	.	.	.57	.	.	.	.	.	.02	.	.	.30	.46	.02	.74	.	.	.	.	.	.	.04	.02						
9	.41	.	.	.	.	.	.	.	.02	.02	.02	.	.	.	.	.07	.10	.05	.06	.	.	.	.						
9	.16	.	.	.	.03	.02	.03	.03	.06	.06	.06	.	.	.	.	.11	.02	.09	.03	.	.	.04	.						
2	.15	.	.	.	.05	.	.02	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.						
2	.03	.	.	.	.04	.02	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.						
3	.15	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.						
9	.03	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.						
8	.02	.	.	.	.29	.02	.40	.	.	.03	.04	.	.	.	.	.	.	.	.	.	.	.	.						
卅	.	.	.	.20	.21	.03	.04	.	.	.10	.05	.	.	.	.	.	.	.	.	.	.	.	.						
卄	.	.	.	.09	.11	.03	.03	.	.	.09	.04	.	.	.	.	.	.	.	.	.	.	.	.						
艹	.	.	.	.	.	.	.	.	.	.02	.	.	.	.	.	.	.	.	.	.	.	.	.						
艹	.	.	.	.	.	.	.	.	.	.07	.02	.	.	.	.	.	.	.	.	.	.	.	.						
𝌆	.	.	.	.	.	.	.03	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.						
𝌆	.	.	.	.	.	.	.02	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.						
𝌇	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.						
𝌇	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.						
tot	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0						

Table 9: Previous-symbol probabilities for the basic glyphs in the VMS text. The table should be read by columns; i.e., the entry '.23' in column 9, row c means that 23% of the occurrences of 9 in the text are preceded by c.

Tables 10 and 11 give the same statistics for the Voynichese lexicon (ignoring repeated words).

	□		ᛁ																				tot
□	.	.10	.	.	.04	.24	.08	.08	.04	.13	.07	.05	.	.	.	.04	.04	.	.04	.	.	.	.	1.0
	.	.	.	.05	.	.89	.	.	.	.	.	.	.	.	.	.02	.	.	.	.	.	.	.	1.0
	.05	.	.	.04	.02	.	.	.13	.03	.02	.	.21	.11	.	.	.14	.12	.02	.04	.	.	.	.	1.0
	.	.	.	.27	.04	.22	.13	.16	.04	.02	.	.	.	.	.	.04	.02	.	.	.	.	.	.	1.0
	.	.	.	.39	.06	.25	.10	.08	.02	.	.	.	.	.	.	.02	.	.	.	.	.	.	.	1.0
	.	.	.	.45	.04	.23	.10	.08	.	.	.	.	.	.	.	.02	.	.	.	.	.	.	.	1.0
ᛁ	.	.	.37	.	.	.	.	.	.	.	.	.02	.13	.41	.02	.	.	.	.	.	.	.	.	1.0
	.	.	.33	.	.	.	.	.	.	.	.	.28	.24	.02	.07	.	.	.	.	.	.	.	.	1.0
	.75	.	.	.	.	.	.	.03	.	.03	.	.	.	.	.	.07	.05	.	.	.	.	.	.	1.0
	.32	.	.	.02	.07	.09	.08	.08	.04	.09	.04	.	.	.	.	.11	.02	.	.	.	.	.	.	1.0
	.55	.	.	.	.16	.11	.07	.02	.	.04	.02	.	.	.	.	.	.	.	.	.	.	.	.	1.0
	.45	.	.	.04	.17	.16	.07	.02	.	.05	.	.	.	.	.	.	.	.	.	.	.	.	.	1.0
	.94	.	.	.	.	.	.02	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1.0
	.91	.	.	.	.02	.02	.02	.02	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1.0
	.10	.	.	.03	.32	.06	.37	.	.	.05	.02	.02	.	.	.	.	.	.	.	.	.	.	.	1.0
	.	.	.	.37	.22	.12	.08	.	.	.14	.05	.	.	.	.	.	.	.	.	.	.	.	.	1.0
	.02	.	.	.28	.18	.18	.07	.	.	.19	.06	.	.	.	.	.	.	.	.	.	.	.	.	1.0
	.05	.	.	.	.20	.18	.09	.	.	.40	.06	.	.	.	.	.	.	.	.	.	.	.	.	1.0
	.02	.	.	.	.16	.21	.06	.04	.	.42	.08	.	.	.	.	.	.	.	.	.	.	.	.	1.0
	.02	.	.	.33	.07	.21	.29	.05	.02	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1.0
	.02	.	.	.27	.13	.28	.20	.06	.02	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1.0
	.	.	.	.25	.09	.18	.36	.09	.02	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1.0
	.	.	.	.38	.09	.28	.18	.05	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1.0

Table 10: Next-symbol probabilities for basic glyphs in the Voynichese lexicon. The table should be read by rows; i.e., the value '.13' in row ᥴ and column ᦐ means that 13% of the occurrences of ᥴ in the lexicon are followed by ᦐ.

	□	4	ı	c	a	o	9	8	2	cc	2c	9	?	J	8								
□	.	.96	.	.	.08	.28	.15	.18	.30	.34	.42	.12	.06	.	.	.14	.22	.30	.40	.19	.35	.35	.51
4	.	.	.	.	.	.11	.	.	.	.	.	.	.	.	.	.	.	.	.	.03	.02	.	.
o	.04	.02	.05	.05	.04	.	.02	.26	.19	.04	.05	.47	.34	.	.18	.41	.49	.36	.35	.33	.29	.27	.27
c	.	.	.	.27	.06	.17	.17	.26	.20	.03	.04	.	.	.	.	.09	.07	.14	.07	.16	.09	.04	.04
cc	.	.	.	.23	.05	.11	.08	.08	.06	.	.	.	.	.	.	.02	.03	.	.04	.14	.11	.18	.08
2c	.	.	.	.11	.	.04	.03	.03	.02	.	.	.	.	.	.	.	.	.	.	.07	.04	.04	.02
ı	.	.	.37	.	.	.	.	.	.02	.	.	.	.14	.92	.10	.	.	.	.	.	.	.	.
a	.	.	.56	.	.	.	.	.	.04	.	.	.34	.43	.06	.64	.	.	.	.	.02	.03	.04	.05
9	.39	.	.	.	.	.	.	.04	.04	.04	.05	.	.	.	.	.12	.13	.06	.08	.02	.02	.	.
8	.12	.	.	.	.06	.04	.06	.08	.10	.09	.11	.	.	.	.02	.15	.04	.10	.04	.	.	.05	.
?	.15	.	.	.	.09	.03	.03	.	.	.03	.03	.	.	.	.	.	.	.	.	.	.	.02	.
2	.06	.	.	.	.05	.02	.02	.	.	.02	.	.	.	.	.	.	.	.	.	.	.	.	.
J	.12	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
8	.05	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
8	.04	.	.	.02	.28	.03	.30	.	.02	.05	.05	.02	.	.	.02	.	.	.	.	.02	.	.	.
	.	.	.	.16	.13	.04	.04	.	.	.11	.09	.	.	.	.	.	.	.	.	.	.	.	.
	.	.	.	.09	.08	.04	.03	.	.	.10	.08	.	.	.	.	.	.	.	.	.	.	.02	.
	.	.	.	.	.02	.	.	.	.	.04	.	.	.	.	.	.	.	.	.	.	.	.	.
	.	.	.	.	.03	.02	.	.	.	.10	.05	.	.	.	.	.	.	.	.	.	.	.	.
	.	.	.	.02	.	.	.02	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
tot	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Table 11: Previous-symbol probabilities for the basic glyphs in the Voynichese lexicon. The table should be read by columns; i.e., the entry '.17' in column 9, row c means that 17% of the occurrences of 9 in the lexicon are preceded by c.

As tables 8 and 9 show, the next- and previous-glyph distributions are highly non-uniform, with many "forbidden" glyph pairs. Moreover, the glyphs can be grouped into several distinct classes with similar and characteristic distributions (indicated by vertical and horizontal lines in the tables). These strong features bring to mind the phonological/ortographical constraints typical of natural languages. Unfortunately, all atempts to match the Voynichese glyph classes with the symbol classes of known languages have been in vain. In particular, Sukhotin's vowel/consonant identification algorithm [**?**] does not produce a convincing bipartition of the basic glyph set [2].

On the other hand, those failures could mean only that the alphabet assumed in those studies — typically, some variant of table 2 — was so far from the true Voynichese alphabet

that the key features of the digraph distribution were distorted beyond recognition.

4.10 Glyph shape and statistics

Even on casual inspection it is obvious that glyph classes implied by the context statistics are strongly correlated with the glyph shapes. It has long been known, for example, that the four gallows occur in similar contexts, which are different from the contexts of other letters.

In order to explain this phenomenon, it has been conjectured that the shape of the glyph could be related to its pronunciation; or, even, that the strokes could represent specific phonetic traits, such as voiced/unvoiced, long/short, front/back, high/low, etc.. (There are plenty of examples of alphabets displaying such "phonetic correlation." Traces of it can be seen even in the Roman alphabet itself: compare for example the shapes and sounds of `C` and `G`, `P` and `B`, `M` and `N`, `S` and `Z`.) Under this hypothesishe, the apparent connection between glyph shape and statistics in the VMS could be a consequence of phonetic rules, such as exist in all natural languages, that force similar-sounding phonemes to occur in similar contexts.

However, a closer look at the adjacent-glyph statistics shows some unexpected features that do not seem to fit the above theory. If we break down each glyph into its component strokes, according to tables 5(a) and 5(b), we find that all glyphs on the same row of either table (i.e., with the same stroke on the left side) seem to have similar previous-glyph distributions; and any two glyphs in the same column (i.e., with the same right stroke) will have similar next-glyph distributions.

This asymmetric correlation seems hard to explain in terms of phonetic mapping. Traits like duration, stress, and place of articulation are usually manifested simultaneously on each phoneme, not serially. Therefore, it seems unlikely that one trait of a phoneme would be strongly correlated with the previous phoneme, while another would be strongly correlated with the next one. Even if the strokes represented atomic articulatory motions, or phoneme pairs, we would expect to see more strokes and more single-stroke characters (corresponding to vowels).

4.11 Glyph entropy

The entropy h_1 of a random glyph from the text is about 3.83 bits, fairly similar to the entropy of a random letter in English (3.97) and Latin (3.91). However, the next-character entropy h_2 is 2.21 bits, against 3.06 for English and 3.21 for Latin. This apparent anomaly has been discussed at length [**?**, **?**] and has led some investigators to doubt the existence of meaningful contents in the VMS. However, this anomaly too can be explained as a consequence of using the wrong alphabet. In fact, it turns out that the higer-order entropies h_k for $k > 2$ are actually a bit higher for Voynichese than for Latin or English text. See figure 2.

Missing figure auto/hk-plots-basic.eps.eps

Figure 2: Entropy (expected information contents) of a random glyph from the text, given the preceding $k-1$ glyphs, as a function of k. Word spaces were treated as letters.

The relative flatness of the plot between $k = 2$ and $k = 5$ in figure **??** shows that although there is a strong correlation between a Voynichese glyph and the preceding one (see section 4.10), there is almost no correlation between symbols spaced two or three positions apart — which is unlike the situation in English and Latin, where the correlation decreases gradually as the separation increases.

This confusing situation highlights a basic limitation of character-based analysis: the results may change quite radically if the input text is modified by fairly simple variable-length or multi-valued encodings. Thus, we should not expect useful clues from character entropy studies, until we somehow identify the correct symbol boundaries and identities.

In particular, we should not expect character-based statistics to prove or disprove that the VMS text is some secret cipher, or a plaintext in some "exotic" language (possibly with an original spelling system). The statistics do tell us, however, that the text is not a simple Caesar encryption of any major European language. (If it were, the code would have been broken decades ago.) They also seem to rule out simple Vigenère or polyalphabetic substitution ciphers, since such codes tend to flatten out the character and digraph distributions. In fact, if the VMS is encrypted, the code is probably an original system devised by the author.

In any case, extensive analyses by R. Zandebergen, G. Landini, M. Perakh, and others have shown that the letter and n-gram distributions are fairly consistent through the whole book, with modest but significant deviations at all scales [**?**, **?**]. These properties are at least consistent with the theory that the VMS contains a meaningful text in natural language.

4.12 Are the word spaces reliable?

Considering that certain glyphs, like $\mathfrak{d}$ or $_9$, occur mostly at the end of words, it has been conjectured that the Voynichese word spaces are either part of the alphabet [**?**] or "nulls" inserted according to specific rules in order to confuse the lay reader [**?**].

However, if we compute the entropy of the glyphs that may follow a specific glyph or glyph sequence, counting word space as a distinct symbol, we find that the highest values generally occur after a word break. Coincidentally, the same phenomenon is observed in our English and Latin samples. We read this fact as evidence that the Voynichese words and word spaces are indeed what they seem to be.

5 The Voynichese words

The VMS as we can read it today contains about ■ tokens, of which ■ are in the running text and ■ in the illustration labels and other isolated tokens. Ignoring repetitions,

the *Voynichese lexicon* contains ■ distinct words — ■ in the maintext and ■ in labels. (It should be stressed that these counts exclude the lost folios, and tokens which contain unreadable glyphs or weirdos.)

5.1 Word frequency distribution

Word-based statistical analysis of the Voynichese text has generally been more rewarding than character-based analysis [**?**, 11, 12]. For one thing, the word frequencies satisfy Zipf's frequency-versus-rank law, roughly to the same extent as other natural-language texts [4]. See figure 3.

Missing figure langs-text-zipf.eps Missing figure langs-labs-zipf.eps

Figure 3: Plot of word frequency versus word frequency rank (Zipf's plot), for Voynichese plain text (left) and labels (right), compared to samples of English and Latin text. The sloping line is the ideal inverse law freq $= C/$ rank. The English and Latin texts were truncated so as to match the token count of the Voynichese samples.

As shown by figure 3 (left), the Voynichese word frequencies are not far from Zipf's ideal distribution. In fact, for ranks 3 and higher, the VMs distribution is closer to the ideal than that of the Latin sample. The Voynichese label words, on the other hand, have a fairly flat frequency-rank plot, that does not follow Zipf's law at all, and is quite unlike the plots for the two other languages. Indeed, there are very few repeated words among the labels; the most common ones — ♃, ♂ — occur only 10 times each in the whole book. Within some sections, especially the cosmological and zodiacal ones, label words typically occur only once — as one would expect from labels in an atlas.

Looking more closely at the main text plot, we see that the frequency of the most common word in the VMS main text (♁♃, 2.5%) is considerably lower than the frequency of the most common word in English (`the`, 8.2%) or Latin (`et`, 6.6%). In fact, the Voynichese plot is consistently lower and flatter than the English one up to rank 20 or so. This feature may be an indication of polymorphism, i.e. the most common words have two or three different variants or spellings, about equally common.

Incidentally, the ten most common words in the Latin sample are

```
et in est ad non ut qui de quod cum autem quae eius si sunt
```

The low Latin word frequencies for ranks 3 onwards could be attributed to the inflection of certain words (`est` and `sunt`; `qui`, `quod`, and `quae`; etc.). If inflections were supressed, the Latin rank-frequency plot would probably get closer to the ideal. Indeed, it seems possible to rectify the Voynichese plot by identifying some common words in pairs by a suitable similarity criterion, like ♁♃ = ♃, ♐♌♁ = ♋♌♁, etc.

5.2 Lexicon size

The long tail of Zipf's distribution makes it difficult to estimate or even define, the *lexical complexity* (number of distinct words) in a natural language. However, if we can say that the lexical complexity of the VMS main text (6525 words in about 35,000 tokens) lies between that of our English and Latin samples (4801 and 8263 words, respectively). It should be noted that the Latin sample is actually the join of two very different texts. ⋆[Fix this!] R. Zandbergen has produced plots of vocabulary size as a function of text size, which show small discontinuities at section boundaries.

5.3 Word entropy

As one may expect from the similarity of the Zipf plots, the entropy of a single random token from the Voynichese text (10.1219 bits) [**?**] is quite similar to the values observed in Latin and English (10.6160 and 9.1758 bits, respectively). However, as R. Zandbergen observed, the average entropy g_k of the k-th glyph in a random Voynichese token, given the preceding $k-1$ symbols, is *lower* than the corresponding value for English or Latin when $k = 2$, but is *higher* (and more uniform) for $k \geq 3$. See figure 4.

Missing figure auto/entropy-profile-voyn-basic.eps.eps

Figure 4: Entropy (expected information contents) of the kth glyph in a random token from the text, given the preceding $k-1$ glyphs, as a function of k. Word end was treated as a glyph.

5.4 The most popular words

Tables 12 and 15 show some of the most common and least common words in the main text of the manuscript.

886 .0253	212 .0061	140 .0040	96 .0027
548 .0156	211 .0060	137 .0039	95 .0027
515 .0147	196 .0056	133 .0038	95 .0027
462 .0132	190 .0054	130 .0037	91 .0026
437 .0125	182 .0052	126 .0036	91 .0026
403 .0115	173 .0049	118 .0034	88 .0025
365 .0104	169 .0048	116 .0033	87 .0025
360 .0103	166 .0047	115 .0033	85 .0024
348 .0099	158 .0045	115 .0033	85 .0024
338 .0096	153 .0044	113 .0032	81 .0023
305 .0087	152 .0043	112 .0032	79 .0023
305 .0087	151 .0043	110 .0031	79 .0023
285 .0081	151 .0043	109 .0031	78 .0022
278 .0079	147 .0042	107 .0031	77 .0022
270 .0077	147 .0042	107 .0031	75 .0021
269 .0077	146 .0042	105 .0030	75 .0021
266 .0076	146 .0042	102 .0029	74 .0021
264 .0075	142 .0041	97 .0028	73 .0021
240 .0069	141 .0040	97 .0028	72 .0021
219 .0063	141 .0040	96 .0027	72 .0021

Table 12: The 80 most common words in the manuscript, with their total token counts and relative frequencies.

Table 13: A random sample (40 words) of the least common words in the main text of the manuscript (one occurrence each).

Note that the frequency of a word bears no obvious relation to its structure, except that the most common words tend to be shorter than average. Once again, these features are universal characteristics of natural languages, and exclude certan encryption methods which, like Vigenère's, map the same plaintext word to many different code words.

Tables 14 and **??** show some of the most common and least common words in figure labels. Observe that the most common label word has only ■ as much relative frequency as the most common word of the main text. In other terms, Voynichese label words are by

and large unique. Note also that the most common words in the plain text are rarely used in labels.

10 .0100	7 .0070	4 .0040	4 .0040
10 .0100	6 .0060	4 .0040	4 .0040
9 .0090	6 .0060	4 .0040	4 .0040
9 .0090	6 .0060	4 .0040	4 .0040
8 .0080	6 .0060	4 .0040	4 .0040
8 .0080	5 .0050	4 .0040	4 .0040
7 .0070	5 .0050	4 .0040	3 .0030
7 .0070	5 .0050	4 .0040	3 .0030
7 .0070	5 .0050	4 .0040	3 .0030
7 .0070	5 .0050	4 .0040	3 .0030

Table 14: The 40 most common words in the figure labels, with their total token counts and relative frequencies.

Table 15: A random sample (40 words) of the least common words in the figure labels (one occurrence each).

The labels on the illustrations are too long and complex to be letters, too irregular to be numbers, and too diverse to be random garbage; hence it is almost certain that they are lexical items of the language. But, as we shall see, their internal structure is quite similar to that of text words. This is a strong arument for the hypothesis that the Voynichese words are indeed words in the usual sense.

5.5 Word frequencies per section

Table 16 lists the 25 most common words in each section of the manuscript.

pha	hea	heb	cos	str	zod	bio
.044	.052	.024	.022	.018	.030	.038
.020	.029	.020	.017	.018	.029	.036
.015	.019	.020	.016	.016	.028	.033
.015	.016	.020	.016	.015	.016	.024
.014	.014	.016	.015	.014	.012	.023
.014	.013	.015	.014	.013	.012	.022
.014	.013	.014	.014	.013	.011	.018
.013	.013	.012	.011	.012	.010	.017
.011	.012	.012	.011	.011	.010	.015
.010	.010	.011	.011	.011	.009	.014
.010	.009	.011	.011	.011	.009	.013
.010	.009	.010	.010	.009	.009	.013
.010	.008	.010	.008	.009	.009	.012
.010	.007	.010	.008	.008	.009	.011
.010	.007	.008	.007	.007	.008	.011
.009	.007	.008	.007	.007	.008	.010
.009	.006	.008	.006	.007	.008	.009
.008	.006	.007	.006	.006	.008	.009
.008	.006	.007	.006	.006	.008	.008
.008	.006	.007	.006	.006	.007	.008
.007	.006	.007	.005	.006	.007	.007
.007	.005	.006	.005	.006	.007	.007
.007	.005	.006	.005	.005	.006	.007
.007	.005	.006	.005	.005	.006	.007

Table 16: The 25 most common words in each section and their relative frequencies in the section.

As it can be seen from the table, some words are fairly common in all sections, while some words are largely confined to one section. Detailed analysis reveals even more significant variations in word frequencies from page to page. Once again, this combination of regularity and variation is consistent with the thesis that Voynichese is a meaningful text, and would hardly be seen in randomly generated gibberish.

5.6 Token length distribution

The average token length (number of basic glyphs) is 4.5 for running Voynichese text, and 5.1 for the VMS labels. These numbers are similar to the average token length in typical

English and Latin texts, respectively 4.4 and 5.4. However, the distribution of token lengths is distinctively anomalous; see figure 5.

Missing figure langs-t-lengths.eps

Figure 5: Relative token frequencies, as a function of token length (number of basic glyphs), in Voynichese plain text and figure labels, compared to English and Latin text.

Note that Voynichese has comparatively few words of length 2 and 3, or greater than 7. Although our measure of word length can be questioned, a mere change of alphabet would not solve the problem — it would change the horizontal scale of the plot, but would have little effect on the shape of the distribution. Therefore, the abrupt fall-off at both ends of the graph is likely to be a real feature of the language, and not an artifact of the choice of alphabet.

Several theories have been advanced to explain the anomalous lack of long tokens [**?**, **?**]. Some of these theories can be dismissed because they would imply in significant deviations from Zipf's law. In any case, the phenomenon seems to be intimately connected to the structure of the words — which we address in section 6.

5.7 Word length distribution

When we plot the relative count of distinct *words* of each given length, irrespective of how many times each word occurs in the text, we obtain a rather striking result. See figure 7.

Missing figure langs-w-lengths.eps

Figure 6: Relative count of distinct words, as a function of word length, in Voynichese plain text and figure labels, compared to English and Latin text.

The almost exact match between the plain text and label distributions, and their symmetry around the mean length (5.5), are quite remarkable coincidences that cry out for an explanation.

In fact, the relative count w_k of words of length k fits almost perfectly the binomial distribution of degree 9, shifted by 1; i.e.

$$w_k \approx \frac{1}{2^9}\binom{9}{k-1}$$

See figure **??**.

Missing figure binom-w-lengths.eps

Figure 7: Relative count of distinct words, as a function of word length, in Voynichese plain text and figure labels, compared to the binomial distribution of 9 fair coins, shifted by 1.

This result means that the length of a random word from the lexicon has the same distribution as the sum of nine 0-1 random binary variables, plus one. An encoding that could generate this kind of distribution is described in section E.

6 Word paradigms

It has long been known that the Voynichese words have a non-trivial internal structure [**?**], manifested by restrictions on the order and position of the glyphs. Several structural models or *paradigms* for the Voynichese lexicon (or subsets thereof) have been proposed over the last 80 years, e.g by J. Tiltman [13], M. Roe [5], R. Firth [1], and the present author [8, 7]. We will review some of those paradigms below, and then present a new one, which is the main topic of this paper.

To describe sets of words, we borrow some standard notation from formal language theory [**?**]. In particular, we'll use X^* to mean the concatenation of zero or more strings from set X, and $X^?$ to mean at most one string from X — i.e. $\{()\} \cup X$, where () denotes the empty string.

6.1 Tiltman's paradigm

One of the earliest paradigms is due to J. Tiltman, a British cryptographer who analyzed the starred-item section in the ■s. Titlman observed that many words of that sample could be formed by combining a certain set of roots with a certain set of suffixes, listed in table 8:

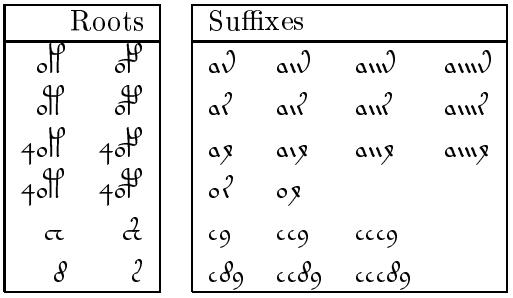


Figure 8: John Tiltman's root-suffix paradigm for VMS words.

Tiltman's paradigm generates 240 distinct words, of which 149 occur in the VMS text, with 10863 occurrences in total. That means 2.16% of all words, and 30.15% of all tokens.

6.2 Mike Roe's paradigm

The automaton A of figure 9, devised by Mike Roe [**?**], is a typical example of those partial paradigms.

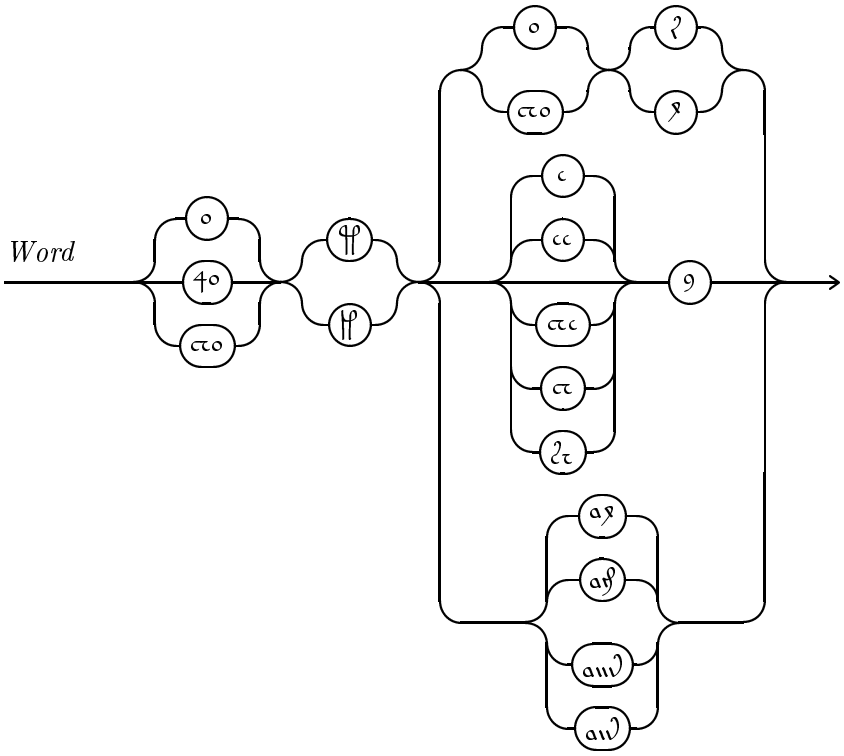


Figure 9: Mike Roe's automaton-based paradigm for VMS words.

Roe's paradigm, more conservative than Tiltman's, generates 78 words, and all but one of them are found in the reference text, with 3804 occurrences in total. That means 1.1% of all words, and 10.6% of all tokens. The one exception is ; since the similar-looking occurs only once, we can ascribe the absence of to sampling error.

6.3 Robert Firth's paradigm

Robert Firth's paradigm is similar to Tiltman's, but uses different (and larger) set of roots and a suffixes, listed in table 17.

Table 17: Robert Firth's root-suffix paradigm.

Firth's paradigm generates 496 distinct words, of which 366 appear in the reference text and account for 16074 tokens. That corresponds to 5% of all words and 44.6% of all tokens. (Actually, in Firth's paradigm the word spaces were not considered significant; with that assumption, the model may turn out to cover an even larger fraction of the text.)

7 The new word paradigm

We now describe a new paradigm that is more general and accurate than the previous models. The paradigm consists of two parts: the *fine structure model*, detailed in the rest of this section, defines local constraints on the order of glyphs within a word; and the *layer model*, the topic of section **??**, defines a decomposition of the typical word into seven quite distinct parts. A more detailed and quantitative version of the paradigm will be presented and discussed in section **??**.

The new paradigm fits equally well the words from ordinary text and to figure labels, and therefore strengthens the claim that the text words are indeed semantic units. The paradigm also provides strong support for John Grove's theory that many ordinary-looking words occur prefixed with a spurious letter [**?**].

7.1 The fine structure of Voynichese words

The fine structure model says that most words are built from a small set of *elements*, each consisting of 1 to 3 of the basic glyphs of table 2. The elements are listed in table 18.

Class	Elements											
Q	5133	.0379										
Y	16837	.1242										
A	13538	.0998		23689	.1747							
H	7680	.0566		4569	.0337		365	.0027		1313	.0097	
	1691	.0125		991	.0073		·	.		4	.	
	653	.0048		733	.0054		55	.0004		144	.0011	
	230	.0017		185	.0014		18	.0001		61	.0004	
X	6370	.0470		2306	.0170		4100	.0302				
	4063	.0300		2029	.0150		339	.0025				
D	12417	.0916		10001	.0738		6383	.0471		2355	.0174	
N	133	.0010		991	.0073							
	1324	.0098		49	.0004		581	.0043				
	4016	.0296		13	.0001		132	.0010				
	103	.0008										

Table 18: The basic elements of Voynichese, according to the fine structure model. The classes are explained in section 7.3.

The fine structure model also imposes constraints on the order in which the elements of table 18 may follow each other. Specifically, it says that the prototypical Voynichese word has the form formula

$$O^?(KO^?)^* \;=\; O^?KO^?KO^?\cdots KO^? \tag{1}$$

where $O = Y \cup A = \{$ a, o, y $\}$ is the set of *circle elements*, and $K = Q \cup H \cup X \cup D \cup N$ is the set of all other elements.

Table 18 and formula (1) impose some non-trivial constraints on the sequence of glyphs. Specifically, it says that the *crescent glyph* c occurs either in pairs, or singly after one of the elements {ch, sh, ckh, cph, cfh, ckhh, cphh, cfhh, ct, cth, cc}. Moreover, the letters {a, o, y} cannot occur between a letter and its c-modifier, and cannot occur next to each other. Finally, the letter s can occur only before {d, l, r}; and the glyphs l and r may occur only in word-final position.

Formula (1) fits more than ■% of the VMS tokens, and ■% of its words.

7.2 Justifying the fine structure model

Table 18 and formula (1) can be justified by the glyph pair statistics. Generally speaking, compound elements like ch and ol were identified by observing that one or more of their constituent glyphs occurs almost exclusively as part of those combinations.

7.2.1 The crescent glyph

In particular, as tables 8 and 9 suggest, the crescent glyph ౮ either follows a gallows or bench glyph (one of {ౘ, ౙ, ౹, ౺, ౻, ಹ, ಺, ಻, ಼}), or is adjacent to another ౮ glyph. See also tables 19 and 20.

	□	𑇄	๐	๐	๑	๑	๘	๒	๒	๏	๏	๓	๓	𑇞	𑇞	𑇞	𑇞	𑇞	𑇞	𑇞	𑇞
□	1	·	3	2	1	1	·	·	4	·	·	1	·	7	13	1	4	·	·	·	·
𑇄	·	·	2	4	2	·	1	·	·	·	·	·	·	13	8	1	2	·	·	·	·
๐	·	·	·	1	·	·	1	·	·	·	·	·	·	1	·	·	·	·	·	·	·
๐	2	·	8	21	4	·	16	1	3	·	1	1	1	41	15	7	8	1	·	·	·
๑	·	·	·	·	1	·	·	·	·	·	·	·	·	2	·	·	·	·	·	·	·
๑	2	·	·	1	·	·	·	·	2	·	·	·	·	3	1	1	·	·	·	·	·
๘	3	·	2	4	14	·	6	·	2	·	·	·	·	·	·	1	·	·	1	·	·
๒	1	·	·	·	1	·	·	·	3	·	·	·	·	·	·	·	·	·	·	·	·
๒	5	·	1	4	10	·	5	1	2	·	·	2	·	1	·	·	·	·	·	·	·
๏	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
๏	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
๓	4	·	153	1001	865	1	1503	5	64	·	·	9	3	162	76	15	34	74	40	3	3
๓	37	·	62	469	448	·	753	2	29	·	·	4	1	91	25	5	9	51	23	·	3
𑇞	1	·	60	527	285	1	652	1	10	·	·	76	24	·	·	·	·	·	1	·	·
𑇞	1	·	32	314	146	·	412	·	8	·	·	27	3	·	·	·	·	·	·	·	·
𑇞	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
𑇞	·	·	·	1	·	·	·	·	·	·	·	·	1	·	·	·	·	·	·	·	·
𑇞	·	·	3	31	142	·	49	·	·	·	·	·	·	·	·	1	·	1	·	·	·
𑇞	·	·	1	40	103	·	34	1	4	·	·	·	·	·	·	·	·	·	·	·	·
𑇞	·	·	·	6	8	·	1	·	·	·	·	·	·	1	·	·	·	·	·	·	·
𑇞	·	·	2	20	19	·	17	·	1	·	·	·	·	·	·	·	·	1	·	·	·

Table 19: Counts of glyph pairs that occur adjacent to a single ๔ glyph. The entry in row ๘ and column ๑ is the number of occurrences of ๘๑ in the main text.

	□	q	a	o	y	l	d	r	s	n	m	ch	sh	k	t	f	p	ckh	cth	cfh	cph
□	1	·	3	2	1	1	·	·	4	·	·	1	·	7	8	1	3	·	·	·	·
q	·	·	2	4	2	·	1	·	·	·	·	·	·	9	7	1	2	·	·	·	·
a	·	·	·	1	·	·	1	·	·	·	·	·	·	1	·	·	·	·	·	·	·
o	2	·	7	16	4	·	9	1	3	·	1	1	1	36	14	7	8	1	·	·	·
y	·	·	·	·	1	·	·	·	·	·	·	·	·	2	·	·	·	·	·	·	·
l	2	·	·	1	·	·	·	·	2	·	·	·	·	3	1	1	·	·	·	·	·
d	3	·	2	4	13	·	4	·	2	·	·	·	·	·	·	1	·	·	1	·	·
r	1	·	·	·	1	·	·	·	3	·	·	·	·	·	·	·	·	·	·	·	·
s	3	·	1	4	8	·	5	1	2	·	·	2	·	1	·	·	·	·	·	·	·
n	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
m	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
ch	4	·	50	270	123	1	207	3	25	·	·	8	3	46	32	14	18	16	9	2	2
sh	10	·	14	128	57	·	85	1	15	·	·	3	1	28	13	4	7	11	4	·	1
k	1	·	31	144	45	1	85	1	9	·	·	39	22	·	·	·	·	·	1	·	·
t	1	·	19	98	27	·	52	·	5	·	·	20	3	·	·	·	·	·	·	·	·
f	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
p	·	·	·	1	·	·	·	·	·	·	·	·	1	·	·	·	·	·	·	·	·
ckh	·	·	3	15	24	·	19	·	·	·	·	·	·	·	·	1	·	1	·	·	·
cth	·	·	1	16	20	·	12	1	2	·	·	·	·	·	·	·	·	·	·	·	·
cfh	·	·	·	5	4	·	1	·	·	·	·	·	·	1	·	·	·	·	·	·	·
cph	·	·	2	12	9	·	7	·	1	·	·	·	·	·	·	·	·	1	·	·	·

Table 20: Counts of glyph pairs that occur adjacent to a single e glyph. The entry in row d and column y is the number of occurrences of dey in the main text's lexicon (ignoring word frequencies).

In fact, if we look at the 9582 occurrences of single e glyphs (not adjacent to another e glyph), we find that only 310 of them (3.2%) are not preceded by a gallows or bench. On the other hand, after a single e glyph one can find gallows, benches, circles, leaders, or finals, all in significant numbers. See tables 21 and 22. We take this asymmetry as one piece of

evidence that a single ᵹ glyph is a part the preceding gallows or bench letter.

	□	4	a	o	9	ɣ	8	ʔ	ɔ	ɔ	ɣ	ɑ	ɑ	ɑ	ɑ	ɑ	ɑ	ɑ	ɑ	ɑ
prev	.	.	.	.01	.	.	.	.	.	.	.	.43	.21	.17	.10	.	.	.02	.02	.	.01
next	.01	.	.04	.26	.22	.	.37	.	.01	.	.	.01	.	.03	.01	.	.01	.01	.01	.	.

Table 21: Distribution of basic glyphs preceding and following a single ᵹ glyph in the main text.

	□	4	a	o	9	ɣ	8	ʔ	ɔ	ɔ	ɣ	ɑ	ɑ	ɑ	ɑ	ɑ	ɑ	ɑ	ɑ	ɑ	ɑ
prev	.01	.01	.	.05	.	.	.01	.	.01	.	.	.37	.17	.17	.10	.	.	.03	.02	.	.01
next	.01	.	.06	.32	.15	.	.22	.	.03	.	.	.03	.01	.06	.03	.01	.02	.01	.01	.	.

Table 22: Distribution of basic glyphs preceding and following a single ᵹ glyph in the main text's lexicon (ignoring word frequencies).

More significantly, the glyph distributions just *after* gallows-ᵹ and bench-ᵹ pairs, such as ɑᵹ and ɑᵹ, are similar to the distributions after the corresponding unmodified gallows and benches. See table **??**. In contrast, the glyph distributions just *before* ᵹ-glyph pairs, such as ᵹ8, are quite unlike those of the corresponding bare glyphs. See table **??**. In other words, the ᵹ glyph transmits to the right the presence of the preceding gallows or bench, but does not transmit to the left any information on the following glyph. Once again, we interpret these observations as hints that single ᵹ is a gallows/bench suffix modifier — one which, in fact, does not change the glyph's character very much.

	□	4	9	a	o	8	ɣ	ʔ	ɔ	ɔ	ɣ	ɔ	ɑ	ɑ	ɑ	ɑ	ɑ	ɑ	ɑ	ɑ	tot
ɔ	.	.	.09	.04	.25	.07	.	.	.	.	.	.	.	.	.	.	.	.02	.	.	.	1.0
ɔᵹ	.	.	.22	.04	.25	.37	.	.	.02	.	.	.	.	.04	.02	.	.	.02	.	.	.	1.0
ɔ	.	.	.06	.03	.22	.04	.	.	.	.	.	.	.	.	.	.	.	.02	.	.	.	1.0
ɔᵹ	.02	.	.22	.03	.23	.37	.	.	.	.	.	.	.	.05	.	.	.	.03	.	.	.	1.0
ɑ	.	.	.08	.30	.07	.	.	.	.	.	.	.11	.02	.	.	.	.	.	.	.	.	1.0
ɑᵹ	.	.	.17	.04	.32	.40	.	.	.	.	.	.05	.	.	.	.	.	.	.	.	.	1.0
ɑ	.	.	.08	.26	.12	.	.	.	.	.	.	.18	.03	.	.	.	.	.	.	.	.	1.0
ɑᵹ	.	.	.15	.03	.33	.44	.	.	.	.	.	.03	.	.	.	.	.	.	.	.	.	1.0
ɑ	.	.	.51	.04	.11	.04	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1.0
ɑᵹ	.	.	.63	.	.14	.22	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1.0
ɑ	.	.	.40	.08	.25	.03	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1.0
ɑᵹ	.	.	.56	.	.22	.19	.	.	.02	.	.	.	.	.	.	.	.	.	.	.	.	1.0

Table 23: Distributions of basic glyphs in the main text just after some digraphs ending with single ᵹ, compared to the distributions after the the corresponding ᵹ-less glyph.

	□	c□	ɡ	cɡ	ɑ	cɑ	o	co	ȣ	cȣ	ɾ	cɾ	ȡ	cȡ	ᚠ	cᚠ	ᚠ	cᚠ	ᚡ	cᚡ	ᚡ	cᚡ	ᚢ	cᚢ	ᚢ	cᚢ
□	.	.02	.10	.	.14	.	.33	.	.29	.	.56	.	.73	.	.13	.02	.18	.09	.42	.07	.35	.03	.22	.	.55	.
ɬ	.	.	.	.	.	.	.22	.	.	.	.	.	.	.	.	.04	.	.06	.	.04	.	.03	.	.	.	.
ɡ	.41	.	.	.	.	.	.	.	.	.	.03	.	.02	.	.07	.	.10	.	.06	.	.05	.	.	.	.	.
ɑ	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
o	.03	.04	.	.	.02	.02	.	.	.17	.	.02	.	.02	.03	.62	.13	.65	.11	.41	.14	.36	.22	.22	.	.15	.
ȣ	.02	.05	.40	.	.29	.	.02	.	.	.	.03	.	.04	.	.	.	.	.	.	.	.	.03	.	.	.	.02
ɤ	.16	.04	.03	.	.03	.	.02	.	.03	.	.06	.	.06	.	.11	.	.02	.	.03	.	.09	.03	.	.	.	.
ʔ	.15	.02	.	.	.05	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
ʕ	.03	.09	.	.	.04	.	.02	.	.	.	.	.02	.	.	.	.	.	.	.	.	.	.	.	.	.	.
ɔ	.16	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
ɣ	.03	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
ɾ	.	.07	.06	.42	.03	.47	.11	.41	.06	.44	.	.07	.	.09	.02	.50	.	.55	.02	.60	.	.47	.27	.58	.14	.62
ȡ	.	.65	.02	.22	.	.19	.04	.19	.	.22	.	.03	.	.03	.	.28	.	.18	.	.16	.	.16	.10	.40	.05	.35
ᚠ	.	.02	.04	.14	.21	.18	.03	.22	.	.19	.10	.63	.05	.73	.	.	.	.	.	.	.	.	.	.	.	.02
ᚠ	.	.02	.03	.07	.11	.10	.03	.13	.	.12	.09	.23	.04	.09	.	.	.	.	.	.	.	.	.	.	.	.
ᚡ	.	.	.	.	.	.	.	.	.	.	.02	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
ᚡ	.	.	.	.	.	.	.	.	.	.	.07	.	.02	.03	.	.	.	.	.	.	.	.	.	.	.	.
ᚢ	.	.	.03	.07	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.03	.	.	.	.
ᚢ	.	.	.02	.05	.	.	.	.02	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
ᚢ	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
ᚢ	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
tot	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Table 24: Distributions of basic glyphs in the main text just before some digraphs beginning with single c, compared to the distributions before the corresponding c-less glyph.

	□	4	9	a	o	8	ɣ	ʔ	ɔ	ʋ	ɣ	ɑ	ɑ									tot
ɑ	.	.	.10	.06	.25	.08	.	.	.02	.	.	.	.	.02	.	.	.	.	.	.	.	1.0
ɑc	.	.	.15	.06	.32	.25	.	.	.03	.	.	.	.	.06	.04	.02	.02	.02	.	.	.	1.0
ɑ	.	.	.09	.04	.23	.07	.	.	.	.	.	.	.	.02	.	.	.	.	.	.	.	1.0
ɑc	.03	.	.15	.04	.34	.22	.	.	.04	.	.	.	.	.07	.03	.	.02	.03	.	.	.	1.0
	.	.	.07	.21	.11	.	.	.	.	.	.	.15	.05	.	.	.	.	.	.	.	.	1.0
c	.	.	.12	.08	.38	.22	.	.	.02	.	.	.10	.06	.	.	.	.	.	.	.	.	1.0
	.02	.	.07	.17	.16	.	.	.	.	.	.	.20	.07	.	.	.	.	.	.	.	.	1.0
c	.	.	.12	.08	.44	.23	.	.	.02	.	.	.09	.	.	.	.	.	.	.	.	.	1.0
	.02	.	.27	.07	.21	.06	.	.	.02	.	.	.	.	.	.	.	.	.	.	.	.	1.0
c	.	.	.38	.05	.24	.30	.	.	.	.	.	.	.	.	.	.02	.	.02	.	.	.	1.0
	.02	.	.19	.13	.28	.06	.	.	.02	.	.	.	.	.	.	.	.	.	.	.	.	1.0
c	.	.	.38	.02	.31	.23	.	.02	.04	.	.	.	.	.	.	.	.	.	.	.	.	1.0

Table 25: Distributions of basic glyphs in the main text's lexicon (ignoring word frequencies), just after some digraphs ending with single c, compared to the distributions after the the corresponding c-less glyph.

	□	c□	9	c9	a	ca	o	co	8	c8	ɑ	cɑ	ƨ	cƨ	ᚠ	cᚠ	ᚠ	cᚠ	ᚡ	cᚡ	ᚢ	cᚢ	ᚢ	cᚢ	ᚣ	cᚣ	ᚣ	cᚣ
□	.	.04	.15	.	.08	.02	.25	.	.18	.	.34	.	.42	.	.15	.05	.24	.11	.43	.08	.32	.03	.20	.	.35	.		
4	.	.	.	.	.	.	.12	.	.	.	.	.	.	.	.	.07	.	.09	.	.05	.	.03	.03	.	.02	.		
9	.38	.	.	.	.	.	.	.	.04	.	.04	.	.05	.	.12	.	.12	.	.07	.	.06	.	.02	.	.02	.		
a	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.02	.	.03	.		
o	.04	.07	.02	.	.04	.05	.	.02	.26	.02	.04	.	.05	.03	.39	.27	.46	.19	.32	.21	.33	.23	.31	.03	.28	.		
8	.04	.11	.30	.04	.28	.	.03	.	.	.	.05	.	.05	.	.	.	.	.	.	.	.	.03	.02	.	.	.07		
ƨ	.12	.07	.05	.	.05	.	.04	.	.07	.	.09	.	.10	.	.16	.02	.05	.	.04	.	.10	.03	.	.	.	.		
ʔ	.15	.04	.03	.	.08	.	.03	.	.	.	.03	.	.03	.	.	.	.	.	.	.	.	.	.	.	.	.		
ʔ	.06	.11	.02	.02	.04	.	.02	.	.	.	.02	.03	.	.	.	.	.	.	.	.	.	.	.	.	.	.		
ᘐ	.12	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.		
ᘓ	.05	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.		
ɑ	.	.14	.08	.36	.05	.37	.12	.37	.08	.42	.	.11	.	.10	.03	.34	.03	.43	.04	.47	.	.47	.13	.53	.11	.60		
ƨ	.	.36	.03	.17	.	.10	.05	.18	.03	.17	.	.04	.	.03	.	.21	.	.17	.	.18	.	.13	.07	.37	.05	.27		
ᚠ	.	.04	.04	.13	.14	.23	.04	.20	.	.17	.11	.53	.09	.71	.	.	.	.	.	.	.	.	.	.	.	.07		
ᚠ	.	.04	.03	.08	.08	.14	.04	.14	.	.11	.10	.27	.08	.10	.	.	.	.	.	.	.	.	.	.	.	.		
ᚡ	.	.	.	.	.	.	.	.	.	.	.04	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.		
ᚡ	.	.	.	.	.03	.	.02	.	.	.	.10	.	.05	.03	.	.	.	.	.	.	.	.	.	.	.	.		
ᚢ	.	.	.02	.07	.	.02	.	.02	.	.04	.	.	.	.	.	.	.	.	.	.	.	.03	.	.03	.	.		
ᚢ	.	.	.	.06	.	.	.	.02	.	.02	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.		
ᚢ	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.		
ᚢ	.	.	.	.03	.	.	.	.02	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.03	.	.		
tot	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0		

Table 26: Distributions of basic glyphs in the main text's lexicon (ignoring word frequencies), just before some digraphs beginning with single c, compared to the distributions before the corresponding c-less glyph.

7.2.2 Multiple crescent glyphs

If we look at all strings of consecutive c glyphs, we find ■ instances that are preceded by a gallows or bench glyph, and ■ instances that are not. Of the former, ■ (■%) consist of a single c; of the latter, ■ (■%) consist of either two or three c. Thus we feel justified in parsing single c as modifiers of the preceding glyph, and treating cc and ccc as elements on their own.

Moreover, as we observed before, the glyphs that may follow a single c glyph that follows a gallows or bench glyph *g* are those that may follow the glyph *g* by itself; whereas the glyphs that may follow a double or triple c glyph are those that may follow an ɑ or ƨ.

The reader may have noticed that the inclusion of both cc and ccc creates ambiguities in the parsing of some words; for instance, ɑcccc9 could be parsed as either ɑ·ccc·9 or ɑc·cc·9.

Observing that groups like ᚻᚳᚳ are far more common than ᚻᚳᚳ, we arbitrarily chose to resolve the ambiguity by parsing ᚻᚳᚳᚳ as ᚻ·ᚳᚳᚳ rather than ᚻᚳ·ᚳᚳ.

7.2.3 The circle glyphs

The "circle" glyphs {ᚪ, ᚩ, ᚹ} are found interspersed among other elements. In fact the number of circle glyphs is almost exaclty half the number of non-circle elements. If circles and non-circles were intermixed at random, we would expect about ■ double-circle and ■ triple-circle sequences. Instead we see only ■ doublets and ■ triplets. Obviously repetition of circle glyphs is strongly avoided.

Unlike the ᚳ glyph, which can be confidently viewed as a modifier for the preceding letter, it is still an open question whether the circle glyphs are independent letters, or modifiers for adjacent letters, or both. The final groups {ᚸ, ᚿ, ᛁᚿ} and the letters ᚲ and ᚷ, are almost always preceded by ᚪ or ᚩ. In particular, the words {ᚪᚲ, ᚩᚲ, ᚪᚷ, ᚩᚷ} are quite common, while {ᚲᚪᚲᚩ, ᚷᚪ, ᚷᚩ} are essentially non-existent. On the other hand, the glyphs ᚾ and ᚸ are usually followed by a circle letter, but rarely preceded by one.

	□	ᚾ	ᚳ	ᚷ	ᚸ	ᚲ	ᚳ	ᚿ	ᚹ	ᚱ	ᚣ	ᚻ	ᚼ	ᚽ	ᚾ	ᛁᚾ	ᛁᛁᚾ	ᛁᛁᛁᚾ	ᛁᛁᛁᛁᚾ
prev	.33	.22	.14	.02	.02	.01	.01	.	.	.11	.04	.03	.03	.	.01	.	.01	.	.
next	.05	.	.01	.23	.09	.11	.02	.	.01	.01	.	.25	.16	.01	.02	.01	.01	.	.

Table 27: Distribution in the main text of basic glyphs adjacent to a single ᚩ glyph.

	□	ᚾ	ᚳ	ᚷ	ᚸ	ᚲ	ᚳ	ᚿ	ᚹ	ᚱ	ᚣ	ᚻ	ᚼ	ᚽ	ᚾ	ᛁᚾ	ᛁᛁᚾ	ᛁᛁᛁᚾ	ᛁᛁᛁᛁᚾ
prev	.15	.	.03	.03	.30	.05	.04	.	.	.04	.01	.22	.11	.	.01	.	.01	.	.
next	.	.	.	.22	.	.23	.	.01	.06	.	.	.	.	.	.	.	.	.	.

Table 28: Distribution in the main text of basic glyphs adjacent to a single ᚪ glyph.

	□	ᚾ	ᚳ	ᚷ	ᚸ	ᚲ	ᚳ	ᚿ	ᚹ	ᚱ	ᚣ	ᚻ	ᚼ	ᚽ	ᚾ	ᛁᚾ	ᛁᛁᚾ	ᛁᛁᛁᚾ	ᛁᛁᛁᛁᚾ
prev	.10	.	.24	.03	.40	.01	.01	.	.	.06	.02	.04	.03	.	.	.03	.02	.	.
next	.89	.	.	.	.01	.	.	.	.	.02	.01	.04	.03	.	.	.	.	.	.

Table 29: Distribution in the main text of basic glyphs adjacent to a single ᚹ glyph.

	□	ᚾ	ᚳ	ᚷ	ᚸ	ᚲ	ᚳ	ᚿ	ᚹ	ᚱ	ᚣ	ᚻ	ᚼ	ᚽ	ᚾ	ᛁᚾ	ᛁᛁᚾ	ᛁᛁᛁᚾ	ᛁᛁᛁᛁᚾ
prev	.22	.10	.14	.02	.21	.02	.02	.	.	.08	.03	.08	.05	.	.01	.01	.01	.	.
next	.30	.	.01	.16	.04	.11	.01	.	.02	.01	.	.12	.08	.	.01	.	.	.	.

Table 30: Distribution of basic glyphs adjacent to a single circle glyph ({ᚪ ᚩ: *evaby*}).

7.3 The layer model

The elements of the fine-structure model can be partitioned into seven distinct classes Q, A, Y, H, X, D, N, listed in table 18. Throughout this section, we will ignore any occurrences of the glyphs $A \cup Y = \{$ol, or, ar$\}$; their distribution will be discussed separately in section 7.4.6. After erasing those glyphs, it turns out that almost every VMS word can be parsed into five nested layers, each consisting of elements from the same class. More precisely, almost every word is generated by the formula

$$Q^{?} D^{?} X^{\alpha} H^{?} X^{\beta} D^{\gamma} N^{?} \tag{2}$$

where $\alpha + \beta$ and γ are 0, 1, or 2.

7.3.1 Unimodality

Although each factor in formula (2) may be empty, the formula is definitely non-trivial: it rules out, for example, words with two core letters bracketing a mantle or crust letter. More generally, suppose we assign "densities" 1, 2, and 3 to the three main letters classes above, and ignore the remaining letters. The paradigm then says that the density profile of a normal word is a single unimodal hill, without any internal minimum. In other words,as we move away from any maximum-density letter in the word, in either direction, the density can only decrease (or remain constant). The possible density profiles (ignoring repeated digits) are

1	2	3				
12	21	13	31	23	32	
121	123	131	132	231	232	321
1231	1232	1321	2321			
12321						

Note that these are a proper subset of the possible three-level profiles. In particular, the profiles 212, 213, 312, 313, and 323 are excluded by our paradigm.

Formula (2) fits more than ■% of the tokens, and ■% of the words.

★ [Here we should mention the remarkable evenness and independence of the two traits, 'has gallows' and 'has benches'.]

7.3.2 The initial element

The Q prefix, when present, consists of a single qo glyph. can occur only at the beginning of a normal word, although in a few instances (less than 0.4% of all qos,) it is preceded by ol or ar.

The letter qo rarely occurs at beginning of paragraphs or in labels, which may mean that it is a grammatical particle (article, preposition, etc.).

7.4 The final elements

Elements of class N can occur only at the end of the word. They comprise the glyphs $\mathcal{O}$ and $\mathcal{y}$, and clusters consisting of one to four $\mathcal{i}$ glyphs, followed by one of the letters $\{\mathcal{O}, \mathcal{y}, \mathcal{s}, \mathcal{x}, \mathcal{l}, \mathcal{d}\}$.

Actually, as shown in table 31, only a few of those 24 potential $\mathcal{i}$-containing clusters occur in significant numbers.

Class	Elements														
F	$\mathcal{is}$	5	.0008	$\mathcal{ix}$	28	.0044	$\mathcal{il}$	581	.0923	$\mathcal{id}$	11	.0018	$\mathcal{iO}$	1324	.2104	$\mathcal{iy}$	49	.0078
	$\mathcal{iis}$	9	.0014	$\mathcal{iix}$	12	.0019	$\mathcal{iil}$	132	.0210	$\mathcal{iid}$	7	.0011	$\mathcal{iiO}$	4016	.6381	$\mathcal{iiy}$	13	.0021
	$\mathcal{iiis}$	1	.0002	$\mathcal{iiix}$	1	.0002	$\mathcal{iiil}$	1	.0002	$\mathcal{iiid}$	·	·	$\mathcal{iiiO}$	103	.0164	$\mathcal{iiiy}$	·	·
	$\mathcal{iiis}$	·	·	$\mathcal{iiix}$	·	·	$\mathcal{iiil}$	·	·	$\mathcal{iiid}$	·	·	$\mathcal{iiiO}$	1	.0002	$\mathcal{iiiy}$	·	·

Table 31: All the potential final elements of Voynichese.

The asymmetry between $\mathcal{il}$ and $\mathcal{id}$ is puzzling, considering that $\mathcal{l}$ and $\mathcal{d}$ are similar in other respects. Also disconcerting is the fact that the glyphs $\mathcal{O}$ and $\mathcal{y}$ are almost exclusively word-final, whereas $\mathcal{l}$ occurs both internally and as part of $\mathcal{il}$ and $\mathcal{iil}$ elements. However such asymmetries are common in natural languages.

7.4.1 Abbreviation letters

It seems that the letter $\mathcal{y}$ is inordinately common at the end of lines, and before interruptions in the text due to intruding figures. The letter $\mathcal{y}$, like the N groups, is almost always preceded by $\mathcal{a}$ or $\mathcal{o}$ (862 tokens in 950, 91%). We note also that $\mathcal{say}$ and $\mathcal{ay}$ are the most common -$\mathcal{ay}$ words, just as $\mathcal{saiO}$ and $\mathcal{aiO}$ are the most common -$\mathcal{aiO}$ words. Perhaps $\mathcal{y}$ is an abbreviation for $\mathcal{iO}$ (and/or other N groups), used where space is tight.

On the other hand, the truth may not be that simple. of the 950 tokens that contain $\mathcal{y}$, 56 (5.8%) are preceded by $\mathcal{ai}$ or $\mathcal{ai}$ rather than $\mathcal{a}$ alone.

The rare letter $\mathcal{9}$, like $\mathcal{y}$, occurs almost exclusively at the end of words (24 tokens out of 27); however, unlike $\mathcal{y}$, it is *not* preceded by $\mathcal{a}$. We note that $\mathcal{9}$ looks like an $\mathcal{y}$, except that the leftmost stroke is rounded like that of an $\mathcal{a}$. Perhaps $\mathcal{9}$ is an abbreviation of $\mathcal{ay}$?

There are 32 tokens that end in $\mathcal{y}$, but not as $\mathcal{ay}$, $\mathcal{oy}$, or $\mathcal{iy}$. It is possible that these tokens are actually instances of $\mathcal{9}$ that were incorrectly transcribed as $\mathcal{y}$ — a fairly common mistake.

7.4.2 The leaders

⋆ [Rewrite, removing references to the crust layer.]

After the intial and final groups, the next inner layer consists of leaders — the letters $D = \{\mathcal{x}, \mathcal{s}, \mathcal{l}, \mathcal{d}, \mathcal{r}\}$ — with their $\{\mathcal{a}, \mathcal{o}, \mathcal{9}\}$, if any. In normal words, this layer comprises either the whole word (almost exactly 25% of the normal tokens), or a prefix and a suffix thereof (75%). ⋆[Note that these percentages are a consequence of the gallows/bench trait statistics.]

There are 459 tokens (1.3%) where leader letters occur bracketed by non-crust letters on both sides. Most of these exceptions are actually instances of what we call "Grove words" (see section 8).

7.4.3 Leader distribution

Table 32 shows the distribution of number of leaders in words without mantle or core, tabulated separately for words with and without the initial ɬ letter:

Without ɬ		
221	0.02662	.
3565	0.42941	#
4066	0.48976	##
413	0.04975	###
36	0.00434	####
1	0.00012	#####

With ɬ		
38	0.08482	ɬ
299	0.66741	ɬ#
109	0.24330	ɬ##
2	0.00446	ɬ###

Table 32: Distribution of number of leaders in words without core and mantle, with and without ɬ. Each # represents a leader

In words that have a non-empty mantle or core, the crust is divided in two blocks. Table 33 shows the joint distribution of prefix and suffix lengths.

prefix length	suffix length					
	0	1	2	3	4	avg
0	5130	10572	1565	112	2	0.81
1	820	1579	103			0.71
2	59	94	3	2		0.67
3	1	3				0.75
avg	0.16	0.15	0.07	0.00	0.00	

Table 33: Distribution of number of leaders in the crust prefix and suffix of words with core or mantle.

From the row and column averages in table 33, it is clear that prefix length and suffix length (number of leaders) are nearly independent variables. There slight negative dependence that can be noticed between the two may well be the result of transcribers inserting bogus word breaks in longer words.

In any case, the average lengths are 0.14 leaders in the prefix, 0.80 in the suffix, and 0.94 in the whole word. Note that this number is substantially less than the average length of crust-only words; in other words, the presence of core or mantle letters seems to reduce the 'need' for leaders.

7.4.4 The mantle layer

The mantle layer consists primarily of the "bench" letters:  and  , and the  group, which, in its n-gram statistics, seems to be a variant of those two. As explained above, we include in the mantle also single letters, except those that follow a core letter; and any  letters prefixed to the above.

Almost exactly 1/4 or the normal tokens have a non-empty mantle, but no core. In those words, the mantle typically consists of one or two benches, combined of course with single letters and circles. If we ignore the latter, and replace by , the most common combinations in normal words are:

68	0.00799		3292	0.38661				
185	0.02173		3851	0.45226				
90	0.01057		917	0.10769				
2	0.00023		24	0.00282				
3	0.00035		42	0.00493		17	0.00200	
2	0.00023		7	0.00082		2	0.00023	
5	0.00059							
2	0.00023							

In words that have gallows letters, the mantle is normally split into two contiguous segments, a prefix and a suffix, and either or both of them may be empty.

★ [Here we need some tabulations?]

The implied structure of the mantle is probably the weakest part of our paradigm. Actually, we still do not know whether the single after the core is indeed a modifier for the gallows letter (as the grammar implies); or whether the pedestal of a platform gallows is to be counted as part of the mantle; or whether the groups ought to be parsed as , , or neither; and so on.

Allowing for both and in the mantle could make the grammar ambiguous. Fortunately, it turns out that the only ambiguous string that is common enough to matter is . (The string occurs only 4 times in the whole manuscript.) Our grammar parses as followed by .

7.4.5 The core layer

The core layer of a normal word, by definition, consists of the "gallows" letters {, , , } or their "pedestal" variants {, , , }; each possibly prefixed by one or more round letters, and followed by a single or . Alternative platforms such as and , and incomplete platforms such as are extremely rare (abot 30 occurrences), and are classified as `AbnormalWord` by the grammar.

A string of two or more letters following a gallows letter is parsed *from right to left*, into zero or more pairs, which are assigned to the mantle, and possibly a single , which is interpreted as part of the core. Thus is parsed as   and as . We have no strong arguments for this rule, except that it avoids ambiguity.

Almost exactly half of the normal words have an empty core, while the other half has a core that consists of a *single* gallows letter, possibly with platform. There are 326 words with two or more gallows. Here is a breakdown of the normal gallows by type:

7084	0.39876		633	0.03563	
4162	0.23428		701	0.03946	
299	0.01683		42	0.00236	
1159	0.06524		129	0.00726	
1749	0.09845		223	0.01255	
966	0.05438		180	0.01013	
3	0.00017		15	0.00084	
3	0.00017		58	0.00326	

Note the almost absolute lack of after and . The anomaly of these counts can be appreciated by comparing the ratios with , , and .

7.4.6 Distribution of the circles

Up to now we have ignored the presence of the "circle" letters $\{a, o, 9\}$ These are usually inserted between the other letters, as in or . The insertion is strongly context-dependent, of course. As several people have observed, two circles in consecutive positions occur with abnormaly low frequency — much less than implied by the frequencies of individual letters. Our decision to attach the circles in the crust to adjacent letters (see the `OR` symbol) was dictated by this observation.

Actually, the rules about which circles may appear in each position seem to be fairly complex, and are still being sorted out. Chiefly for that reaon, the grammar is quite permissive on this point, and may in fact predict significant frequency for many words that have in fact a forbidden circle pattern.

For instance, it is well-known that 9 (with very few exceptions)] only occurs at in word-initial or word-final position. Yet the grammar indifferently allows either 9, o or a at any slot within the crust layer, and either 9 or o within the core and mantle layers. We considered distinguishing initial from medial circle slots in the grammar, but that would have required the duplication several rules.

Our grammar also fails to record the unequal distribution of the circles next to different

"leaders", which can be inferred from the digraph and trigraph statistics:

21	𐬀𐬀	6	𐬀𐬁𐬀	18	𐬀𐬂𐬀
394	𐬁𐬀	1	𐬁𐬁𐬀	44	𐬁𐬂𐬀
27	𐬂𐬀	2	𐬂𐬁𐬀	63	𐬂𐬂𐬀
21	𐬃𐬀	1	𐬃𐬁𐬀	23	𐬃𐬂𐬀
75	𐬀𐬁	730	𐬀𐬁𐬁	199	𐬀𐬂𐬁
30	𐬁𐬁	72	𐬁𐬁𐬁	152	𐬁𐬂𐬁
12	𐬂𐬁	126	𐬂𐬁𐬁	103	𐬂𐬂𐬁
4	𐬃𐬁	95	𐬃𐬁𐬁	133	𐬃𐬂𐬁
11	𐬀𐬂	803	𐬀𐬁𐬂	127	𐬀𐬂𐬂
35	𐬁𐬂	69	𐬁𐬁𐬂	156	𐬁𐬂𐬂
1	𐬂𐬂	107	𐬂𐬁𐬂	61	𐬂𐬂𐬂
2	𐬃𐬂	121	𐬃𐬁𐬂	68	𐬃𐬂𐬂
179	𐬀𐬃	7	𐬀𐬁𐬃	4	𐬀𐬂𐬃
396	𐬁𐬃	2	𐬁𐬁𐬃	17	𐬁𐬂𐬃
45	𐬂𐬃	2	𐬂𐬁𐬃	7	𐬂𐬂𐬃
28	𐬃𐬃	1	𐬃𐬁𐬃	16	𐬃𐬂𐬃

Generally speaking, the letters 𐬀 and 𐬁 seem to be attracted to the slots before 𐬂 and 𐬁, and seem to avoid slots before 𐬀 and 𐬃. To record these preferences in the grammar, it would be necessary to split the `R` symbol into separate symbols `R` → 𐬂 | 𐬁 and `D` → 𐬀 | 𐬃, and similarly for `OR`.

Circles are less common within the mantle layer, but fairly common at the boundaries of those two layers. Again, the present version of the grammar doesn't try to capture these nuances: it allows an optional circle before every core or mantle letter.

On the other hand, the grammar does impose some restrictions about the circle slots just before an `IN` group (where only 𐬁 and 𐬀 are allowed), before 𐬄 and 𐬅 (where only 𐬀 is allowed), before other core or mantle letters (where only 𐬉 or 𐬀 are allowed) and the slot at the very end of the word (ditto).

We have arbitrarily chosen to parse each circle as if it were a modifier of the next non-circle letter; except that a circle at the end of the word (usually a 𐬉 glyph) is parsed as a letter by itself. Thus 𐬀𐬁𐬆𐬰𐬀𐬉 is parsed as 𐬀𐬁·𐬆·𐬰·𐬀𐬃·𐬉. We have no convincing argument to back this choice, except that circles behave quite differently from the more numerous non-circles, so placing both at the same level in the grammar would obscure the structure of the non-circles.

8 Abnormal words

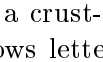
The words that do not fit into our paradigm are collected in the gramamr under the symbol `AbnormalWord`. These words comprise 1295 tokens (3.7%) in the main text, and 127 tokens

(12.4%) in the labels. The vast majority are rare words that occur only once in the whole manuscript. They were manually sorted into a few major classes, according to their main "defect" as we perceived it:

- `Multiple`: words that do not have a properly nested layer structure, and seem to be two more normal words joined together (716 tokens, 55% of the abnormal words). These can be subdivided into:
  - `MultiCore`: words with two or more gallows (208 tokens). The most common is (3 occurrences).
  - `MultiCoreMantle`: words with crust letters surrounded by core or mantle letters (278 tokens). The most common are and (4 occurrences each)
  - `EmbeddedAIN`: words which contain the `A.IN` groups in non-final position (206 tokens). The most common are and (5 occurrences each).
  - `EmbeddedYQ`: abnormal words which contain the letter in non-final, non-initial position; or the letter in non-initial position (24 tokens). The most common is (2 occurrences).
- `GroveWord`: this class was defined by John Grove, who noticed that the rare words often found at the beginning of lines, such as , could be interpreted as normal words prefixed with a spurious gallows letter. Of the abnormal tokens in the text, 213 (16%) fit this description.
- `Weird`: the remaining 366 abnormal tokens (28%) are not easily interpreted as joined words or Grove's gallows-prefixed words. We have sorted them into:
  - `WeirdM`: words that have one of the letters or not preceded by a circle (57 tokens). Apart from the letter by itself (13 occurrences), the most common is (4 occurrences).
  - `WeirdI`: words that contain letter in any context other than an `IN` group (68 tokens). The most common is (2 occurrences).
  - `WeirdSE`: abnormal words that contain single after an (28 tokens). The most common is (3 tokens).
  - `WeirdOther`: abnormal words that did not seem to fit in any of the above categories (213 tokens). Apart from isolated letters like (7 tokens) and (4 tokens) — mainly in the circular text on page f57v — the most common are (6 tokens), , , and (3 tokens each). Note that the latter are probably the result of misreading as in otherwise normal (and common) words.

It is quite possible that, when the VMS is deciphered, we will discover that some of these abnormal words are in fact quite "normal". Indeed, although most "abnormal" words occur only once, some classes of abnormal words may be sufficiently frequent and well defined to

deserve recognition in the grammar. One such candidate, for example, is `EmbeddedAIN`, the set of words that have `A.IN` groups in non-final position.

Conversely, the grammar is probably too permissive in many points, so that many words that it classifies as normal are in fact errors or non-word constructs. See the section about circle letters, for example. For instance, there must be many apparently "normal" tokens which are in fact "Grove words". These could result from prepending a spurious gallows letter to a crust-only normal word (e.g.  +  =), or prepending a spurious non-gallows letter to a suitable normal word (e.g.  +  =). Indeed, it is quite possible that most of the normal-looking line-initial words are in fact such "crypto-Grove" words.

9 Sectional variariation

The rule frequencies vary somewhat from section to section, as shown in the appendices **??** and **??**.

The pages included in each section are listed in section **??**. The special section `txt.n` is the whole text of the manuscript, as used in the main grammar page. For each of those sections, we considerd only paragraph, circular, radial, and "signature" text; excluding labels and key-like sequences. The special section `lab.n` consist of all labels.

It is not surprising to find variations from section to section. What is surprising is that the variations are modest; the basic paradigm seems to hold for the whole text, and the alternatives of each rule generally have similar relative frequencies.

In fact, even those modest differences may not be significant. It has been established that the Voynichese word distribution, like that of natural languages, is highly non-uniform (Zipf-like), largely unconnected to word structure, and highly variable from section to section. Therfore, the rule frequencies in any given section are likely to be dominated by the few most common words in that section — just as the frequency of the digraph `th` in English is largely determined by the frequency of words `the` and `that`.

10 Discussion and conjectures

Perhaps the most important feature of the paradigm is its existence. The non-trivial word structure, especially the three-layer division, pose severe constraints on cryptological explanations. In particular, simple Vigenere-style ciphers, such as the codes considerd by Strong and Brumbaugh, seem to be out of the question, as they would hardly generate the observed word structure.

In fact, the existance of a non-trivial word structure strongly suggests that the Voynichese "code" operates on isolated words, rather than on the text as a whole. (This conclusion is supported also by statistical studies of Voynichese word frequencies, and by the existence of labels and other non-linear text.)

The complexity of the paradigm also discredits the claims that the VMS is nonsense gibberish. It seems unlikely that a 15th century author would invent a random pseudo-

language with such a complex, unnatural structure — and stick to it for 240+ pages, some of them quite boring — only to impress clients, defraud a gullible collector, embarass a rival scholar, or just for the fun of it.

The paradigm has implications also for theories that assume a straightforward (non-encrypted) encoding of some obscure language. The layered word structure does not obviously match the word structure of Indo-European languages. Semitic languages such as Arabic, Hebrew, or Ethiopian could berhaps be transliterated into Voynichese, but not by any traightforward mapping.

In fact, if the VMS is not encrypted, the layered structure suggests that the "words" are single syllables (a conclusion that is also supported by the comparatively narrow range of "word" lengths). However, the number of different "words" is far too large compared to the number of syllables in Indo-European languages. So either the script allows multiple spellings for the same syllable, or we must look for languages with large syllable inventory — e.g. East Asian languages such as Cantonese, Vietnamese, or Tibetan. [6]

Another possibility is that the VMS "words" are isolated stems and affixes of an agglutinative language, such as Turkish, Hungarian, or several Amerind languages. (Indeed, there is evidence of a strong correlation between certain features of consecutive Voynichese words, reminiscent of the Turkish/Hungarian "vowel harmony" rule. [9])

A Digital transcription of the VMS

Preparation of the VMS text for computer analysis requires an encoding of the glyphs into bytes. Several encoding schemes of *transcription alphabets*, loosely based on the glyphs of table 2, have been devised for this purpose. The encodings which are still in common use are listed in table 34.

Glyph	FSG ∼1950	Currier ∼1960	Frogguy ∼1992	EVA ∼1996
	C	C	c	e
	I	I	i	i
	G	9	9	y
	4	4	4	q
	A	A	a	a
	O	O	o	o
	8	8	8	d
	E	E	x	l
	R	R	2	r
	2	2	s	s
	L	D	v	n
	K	J	ig	m
	T	S	ct	Ch
	S	Z	c't	Sh
	D	F	lp	k
	H	P	qp	t
	DZ	X	clpt	CKh
	HZ	Q	cqpt	CTh
	F	V	lj	f
	P	B	qj	p
	FZ	Y	cljt	CFh
	PZ	X	cqjt	CPh

Table 34: Encoding of the essential Voynichese glyphs in some transcription systems.

The *FSG* (*First Study Group*) encoding was used by the very first computerized VMS analysis effort, undertaken between 1944 and 1946 by an informal VMS research team set up at NSA by the noted cryptographer W. Friedman. [**?**, **?**]. Their partial transcription of the VMS into punched cards was recovered in 1995 by J. Reeds and J. Guy [**?**], and was until quite recently the only publicly available digital edition of the text. The *Currier* alphabet was defined by P. Currier for his independent transcription effort; it was proposed as a "standard" by the 1976 workshop organized by M. D'Imperio [**?**, **?**]. The *Frogguy* encoding is an 'anlytical' alphabet developed by J. Guy in 1991, where each character represents a pen stroke rather than a whole glyph [**?**]. The EVA alphabet was defined by R. Zandbergen and G. Landini in 1996 [**?**], and seems to be the most popular one at the moment.

Actually all these systems use additional symbols for some rare glyphs (like π = EVA x = FSG Y) or common glyph combinations (like ω = FSG M). Fortunately, due to the discrete nature of the script, any of these alphabets can be trivially mapped to any other, with negligible loss of information.

B The reference sample

All statistics presented in the previous sections were derived from an almost complete *reference sample* of the VMS transcription, containing 35027 running text tokens and 1003 label tokens. The reason for not using the whole transcription is that all versions that are presently available are contain a significant fraction of reading errors, as well as explicit marks of 'unreadable' characters. If such problematic tokens were included in the samples, they would be improperly counted as failures of the paradigm and introduce a negative bias in the computed failure rate.

To reduce the impact of transcription errors, we took advantage of the fact that almost every part of the VMS text has been transcribed by at least two people, often by three or more. Note that if two people disagree about the reading of some token, at least one of them must be in error. Therefore, whenever we had several readers for a token, for every character position (in the EVA encoding) we used the reading that was reported by the majority of the readers. If there was no definite majority for any character (in particular, if we had only two readers for a token, and they disagreed), we excluded the token from the reference sample.

We also excluded from the sample any tokens which contained very rare characters ("weirdos") like σ or α. Word breaks were not defined by majority vote, but by taking the union of all breaks reported by the various transcribers.

Table 35 gives the number of text words in each section, and the percentage of rejected words.

	Tokens					Words				
Sec	Total	Accepted		Discarded		Total	Accepted		Discarded	
hea.1	6866	6703	97.6	163	2.4	2131	1980	92.9	151	7.1
hea.2	868	823	94.8	45	5.2	554	509	91.9	45	8.1
heb.1	2901	2820	97.2	81	2.8	1189	1111	93.4	78	6.6
heb.2	557	510	91.6	47	8.4	331	288	87.0	43	13.0
cos.1	185	146	78.9	39	21.1	73	63	86.3	10	13.7
cos.2	1491	1353	90.7	138	9.3	868	733	84.4	135	15.6
cos.3	884	713	80.7	171	19.3	533	380	71.3	153	28.7
bio.1	6828	6555	96.0	273	4.0	1536	1325	86.3	211	13.7
zod.1	1010	701	69.4	309	30.6	641	379	59.1	262	40.9
pha.1	926	858	92.7	68	7.3	485	418	86.2	67	13.8
pha.2	1426	1309	91.8	117	8.2	684	587	85.8	97	14.2
str.1	755	670	88.7	85	11.3	483	402	83.2	81	16.8
str.2	10768	10097	93.8	671	6.2	3225	2779	86.2	446	13.8
unk.1	213	202	94.8	11	5.2	162	153	94.4	9	5.6
unk.2	140	134	95.7	6	4.3	103	97	94.2	6	5.8
unk.3	47	44	93.6	3	6.4	46	43	93.5	3	6.5
unk.4	302	292	96.7	10	3.3	226	216	95.6	10	4.4
unk.5	342	309	90.4	33	9.6	246	214	87.0	32	13.0
unk.6	489	431	88.1	58	11.9	297	247	83.2	50	16.8
unk.7	387	357	92.2	30	7.8	235	208	88.5	27	11.5
tot.n	37385	35027	93.7	2358	6.3	8105	6525	80.5	1580	19.5
mid.n	27380	25685	93.8	1695	6.2	5630	4485	79.7	1145	20.3

Table 35: Counts of plain text tokens and words for each section: in the complete transcription, in the reference sample, and in the rejected subset.

Table 36 gives the analgous data for labels.

	Tokens					Words				
Sec	Total	Accepted		Discarded		Total	Accepted		Discarded	
hea.1	1	1	100.0	0	0.0	1	1	100.0	0	0.0
cos.1	10	9	90.0	1	10.0	10	9	90.0	1	10.0
cos.2	255	237	92.9	18	7.1	225	208	92.4	17	7.6
cos.3	122	82	67.2	40	32.8	112	72	64.3	40	35.7
bio.1	147	142	96.6	5	3.4	127	122	96.1	5	3.9
zod.1	360	287	79.7	73	20.3	303	233	76.9	70	23.1
pha.1	97	86	88.7	11	11.3	92	81	88.0	11	12.0
pha.2	162	143	88.3	19	11.7	155	136	87.7	19	12.3
unk.4	15	14	93.3	1	6.7	15	14	93.3	1	6.7
unk.8	2	2	100.0	0	0.0	2	2	100.0	0	0.0
tot.n	1171	1003	85.7	168	14.3	882	721	81.7	161	18.3

Table 36: Counts of label tokens and words for each section: in the complete transcription, in the reference sample, and in the rejected subset.

Although the percentage of rejected text is fairly high (6.3% of the tokens, 20.3% of the words), and even higher for labels (14.3% of the tokens, 18.3% of the words), we believe that the sample is not significantly biased for its intended purpose, namely to estimate the fraction of Voynichese language tokens that fit our paradigm.

For one thing, the vast majority of of the 'bad' tokens were rejected because the transcribers did not agree on the reading of some character, or because they agreed that some glyph was unreadable. Such conditions are mostly due to writing or reading accidents — cramped or careless writing, vellum defects, manuscript damage, poor reproduction quality, etc. — which affect all tokens equally, independently of their structure.

At most, we could expect a slight bias towards loss of longer words, since the probability of misreading or obliterating some glyph in a token may depend on its length. However, as figures 10 and 11 shows, that bias is not visible — the token and word length distributions are practically unchanged by the sampling.

Missing figure cleanup-text-t-len-cmp.eps Missing figure cleanup-labs-t-len-cmp.eps

Figure 10: Effect of sampling on the token length distribution for normal text (left) and labels (right).

Missing figure cleanup-text-w-len-cmp.eps Missing figure cleanup-labs-w-len-cmp.eps

Figure 11: Effect of sampling on the word length distribution for normal text (left) and labels (right).

As for the rare glyphs, some of them are likely to be ordinary glyphs that were mangled by slips of the pen or embellished for aesthetic reasons. Tokens that contain such accidents can be eliminated from the sample without biasing the results, for the same reasons that

apply to contentious or unreadable tokens. Other weirdos may be abbreviations or logographic symbols, like our $^{\text{th}}$, & and \$; given that our aim is to identify the nature of the underlying language, there is no point in including such non-linguistic tokens from analysis. Finally, some of the weirdos — for instance, π and $\mathcal{S}$ — may indeed be rare but legitimate letters of the alphabet, like ü or æ in English; but these are so few that their exclusion from the sample will have negligible effect on the conclusions.

Our word-breaking rule, based on the union of all transcribers, may have introduced a bias in the sample, by preferably deleting longer words, randomly cutting them into pieces, and adding the latter to the sample set. However, the omission of an inter-word space by the scribe seems more likely than the insertion of a bogus one; so the bias in the space-insertion rule probably brings the sample closer to the true text, as intended by the author. In any case, the bias is limited by the rather low rate of disagreement (■%) between the transcribers.

The English text used for inter-language comparisons was H. G. Wells's *War of the Worlds*, extracted from a Gutenberg Project electronic edition. The Latin text was the concatenation of the Rule of the Benedictine monks and the the Vulgate Bible (Old Testament). Both texts were cleansed by removing all numerals and punctuation, converted to lower case, and truncated to so as to have the same total token count as the corresponding Voynichese samples (35027 for text, 1003 for labels).

C A grammar for Voynichese words

C.1 Probabilistic models

Qualitative word paradigms, such as those described in section **??** have some inherent limitations when we rty to apply them to real texts. The Zipf law studies mentioned above support the view that the set V of words used in the Voynich manuscript is only a finite sample of a much larger probabilistic language $\hat{V}$. Therefore, any regularity in the distribution $\hat{V}$ will be obscured by sampling error, which leads to the random exclusion of words whose probability is $\approx 0.5/|V|$. One can gauge the magnitude of this problem by observing that about ■% of the words in V occur only once in the text, and they account for ■% of the tokens. To overcome this limitation, we need to use a *probabilistic* word model, that allows us to take sampling errors into account when evaluating its fit to the data.

One could attempt to build such a model by purely automatic methods, e.g. by interpreting the k-gram frequencies as probabilities in a kth order Markov ■ process. However, a k-th order model with an alphabet of size m has m^k potential states. For $m = 20$ and $k = 6$ (the typical length of a Voynichese word), the number of states would far exceed the number of letters in the VMS text (about ■). The estimated transition probabilities for such model would then be grossly inaccurate; the resulting automaton would be merely a frequency table for the $(k+1)$-letter substrings of the VMS tokens, giving little insight into the mechanisms underlying those frequencies.

Fortunately, inspection of the word frequencies reveals some simple but surprisingly strong constraints in the arrangement of the letters within a word. Therefore, we have chosen to build our models by a semi-automatic method: we specify the qualitative structure of the model, and use the observed word frequencies to adjust its quantitative parameters.

C.2 Grammar notation

We choose to describe the model as a probabilistic grammar, rather than a probabilistic automaton. Although the grammar turns out to be regular, and therefore equivalent to some finite automaton, we find that the former is more readable, and gives more insight into the underlying "linguistic" mechanisms responsible for the structure.

The terminal strings generated by the grammar are word-like strings in the basic EVA alphabet. The notation should be fairly straightforward. The alternatives for each non-terminal symbol are listed together, one per line, in the format

$$
\begin{array}{l}
NTSYMB \rightarrow \\
\quad \begin{array}{cccc}
COUNT_1 & FREQ_1 & CUMFREQ_1 & DEF_1 \\
COUNT_2 & FREQ_2 & CUMFREQ_2 & DEF_2 \\
\ldots & \ldots & \ldots & \ldots \\
COUNT_m & FREQ_m & CUMFREQ_m & DEF_m
\end{array}
\end{array}
$$

where *NTSYMB* is the non-terminal symbol being defined, and each DEF_i is an alternative replacement for it. In conventional notation, without frequency data, the rule above would be written

$$NTSYMB \ \rightarrow \ DEF_1 \mid DEF_2 \mid ... \mid DEF_m$$

In the rewrite strings DEF_i, the terminal strings are in Voynichese script; while non-terminal symbols are in Roman letters. The period "." here denotes the empty string, and is also used as a symbol separator or concatenation operator. The comments in italics are not part of the model.

The fields to the left of each alternative define its frequency of use. Specifically, $COUNT_i$ is the number of times the alternative gets used when parsing the VMS text; $FREQ_i$ is its relative frequency (that is, the ratio of $COUNT_i$ relative to the total *COUNT* of all alternatives of *NTSYMB*); and $CUMFREQ_i$ is the sum of all previous $FREQ_j$ in the section, up to and including $FREQ_i$.

The fields $COUNT_i$, $FREQ_i$, and $CUMFREQ_i$ take into account the word frequencies in the text, as well as the number of times each rule is used in each word. Thus, for example, the derivation of ſaiin uses the rule R → ſ once, and R → i twice; therefore, 100 occurrences of ſaiin in the text would count as 100 uses of R → ſ and 200 of R → i.

C.3 Why the frequencies?

The primary purpose of the *COUNT* and *FREQ* fields is to express the relative "normalness" of each word pattern. We think that, at the present state of knowledge, this kind of statistical information is essential in any useful word paradigm.

The text is contaminated by sampling, transcription, and possibly scribal errors, amounting to a few percent of the text tokens — which is probably the rate of many rare but valid word patterns. Thus, a purely qualitative model would have to either exclude too many valid patterns, or allow too many bogus ones. By listing the rule frequencies, we can be more liberal in the grammar, and list many patterns that are only marginally attested in the data, while clearly marking them as such.

C.4 Predicting word frequencies

Apart from their primary purpose, the FREQ fields also allow us to assign a *predicted frequency* to each word, which is obtained by mutiplying the *FREQ* fields in all rules used in the word's derivation, and adding these numbers for all possible derivations. (Actually there is at most one, since the grammar happens to be unambiguous.)

It would be nice if the predicted word frequencies matched the frequencies observed in the Voynich manuscript. Unfortunately this is not quite the case, at least for the highly condensed grammar given here.

The mismatch between observed and predicted frequecies is largely due to dependencies between the various choices that are made during the derivation. For instance, suppose the grammar contained the following rules:

```
Word:
    100   1.00   1.00   Y.Y

Y:
    100   0.50   0.50   9
    100   0.50   1.00   o
```

This grammar generates the words oo, o9, 9o and 99, and assigns to them the same predicted frequency (0.25). However, the rule counts and frequencies are equally consistent with a text where oo and 99 occur 50 times each, while o9 and 9o do not occur at all — or vice-versa. In other words, the grammar does not say wether the choice of the first `Y` affects the choice of the second `Y`.

These dependencies are actually quite common in Voynichese (and in all natural languages). In English text one will find plenty of `can`, `cannot`, and `man`, but hardly any `mannot`. In Voynichese ſaiin, ſolkeedy and ſolkaiin are all very popular (866, 305, 266 occurrences, respectivey), while ſeedy is essentially nonexistent (3 occurrences). Our paradigm fails to notice this assymetry, since it allows independent choices between ſ- and ſol-, and between -aiin and -eedy.

C.5 Why a grammar?

Although our paradigm is formulated as a context-free grammar, it actually defines a *regular* (or *rational*) stochastic language. Therefore, the grammar could be replaced, in priciple, by an equivalent probabilistic finite-state automaton (i.e., a Markov-style model).

However, we believe that the grammar notation is more convenient and readable than the equivalent automaton, for several reasons. For one thing, it is more succint: a single grammar rule with N symbols on the right-hand side would normally translate into N or more states in the automaton. Moreover, although our grammar is unambiguous, it is not left-to-right deterministic; therefore the equivalent automaton would be either non-deterministic, or would have a very large number of "still undecided" states.

(In fact, our grammar is not recursive, and thus generates a large but *finite* set of words. we could have simplified some rules by making them recursive (e.g. `CrS`), but then the rule probabilities would be much harder to interpret.)

C.6 Implied word structure

The grammar not only specifies the valid words, but also defines a parse tree for each word, which in turn implies a nested division of the same into smaller parts.

Some of this "model-imposed" structural information may be significant; for example, we belive that our parsing of each word into three nested layers must correspond to a major feature of the VMS encoding or of its underlying plaintext.

However, the reader should be warned that the overriding design goals for the grammar were to reproduce the set of observed set of words as accurately as possible, while ensuring unambiguous parsing. Therefore, one should not give too much weight to the finer divisions and associations implied by our parse trees. For example, our grammar arbitrarily associates each ₒ letter to the letter at its right, although the evidence for such association is ambiguous at best.

Said another way, there are many grammars that would generate the same set of words, even the same word distributions, but with radically different parsings. Further study is needed to decide which details of the word decomposition are "real" (necessary to match the data), and which are arbitrary.

C.7 Coverage versus simplicity

When designing the grammar, we tried to strike a useful balance between a simple and informative model and one that would cover as much of the corpus as possible. In particular, we generally omitted rules that were used by only one or two tokens from the corpus, since those could be abbreviations, split words, or transcription errors. However, some of those rules seemed quite natural in light of the overall structure of the paradigm. It may be worth restoring some of those low frequency rules, for the sake of making the grammar more logical.

For example, the present grammar defines

```
IN:
    1770  0.30066  0.30066  ı.N
    4019  0.68269  0.98335  ıı.N
      98  0.01665  1.00000  ııı.N

N:
    5246  0.89112  0.89112  ɔ
     554  0.09411  0.98522  ʔ
      24  0.00408  0.98930  ɣ
      54  0.00917  0.99847  ʕ
       9  0.00153  1.00000  ʔ
```

These rules do not accomodate words containing ııı, ıɾ, or ıʕ — like oıııɔ ʔoſſaıɾ, or ʕaıʕ (1 occurrence each). Yet ııı with count of 1 would be a logical extrapolation of the ı series; and, in other contexts, ʕ and ɾ clearly belong to the same class as ʔ, ɣ, ʔ.

D Normal and abnormal words

The grammar's starting non-terminal symbol (the *axiom* or *root*) is `Word`. For convenience, the grammar actually generates *all* the words that occur in the VMS transcription. Our paradigm proper consists of the sub-grammar rooted at the symbol `NormalWord`. The exceptions — VMS words that do not follow our paradigm — are listed as derivations of the symbol `AbnormalWord`.

It should be noted that that normal words account for over 88% of all label tokens, and over 96.5% of all the tokens (word instances) in the text. The exceptions (less than 4 every 100 text words) can be ascribed to several causes, including physical "noise" and transcription errors. (Different people transcribing the same page often disagree on their reading, with roughly that same frequency.). Indeed, most "abnormal" words are still quite similar to normal words, as discussed in section 8.

Among the EVA letters not listed above, most are so rare that it seems pointless to include them in the "normal word" paradigm. Only the letters {ɛ, ɑ, o, ɡ} are frequent enough to merit special attention.

E A code with binomial length distribution

Here is a code that would produce a lexicon with a binomial distribution of word lengths, similar to that observed in the VMS (figure **??**).

In the first step, we assign to each word of the lexicon a distinct binary number. Then we write down the positions of the '1' bits in each number, in a fixed order, denoting each position by a distinct symbol. For simplicity, let's assume that the lexicon contains at most

$2^1 0$ words; then each bit position can be represented by a decimal digit, counting from 0 the unit end. Finally, we add a marker '#' after the last digit. Let's call the resulting string the *decimal code* of the word. For example:

Binary number	0	1	10	11	100	101	110	111	1000	1001	...
Decimal code	#	0#	1#	10#	2#	20#	21#	210#	3#	30#	...

(Note that the binary numbering step is merely a pedagogical device; once the concept is understood, the decimal codes can be enumerated directly with little effort.)

If the lexicon size is 2^m for some integer m, each of the m bit positions will be 1 in exactly half of the words. In that case, a word drawn randomly from the lexicon will have k ones with probability

$$\text{binom}(n, k, \frac{1}{2}) = \frac{1}{2^k}\binom{m}{k}$$

It follows that the relative count of words whose decimal codes have length k is $\text{binom}(n, k-1, 1/2)$. In particular, if the lexicon has about $2^9 = 512$ words, the code length distribution will have minimum 1, mean 5.5, and maximum 10.

E.0.1 Word scrambling

The distribution of word lengths will remain unchanged if the symbols of each codeword are permuted according to some deterministic rule (one which will return the same result for the same input word). For instance, we could list the even digits in increasing order, then the marker #, then the odd digits in decreasing order:

Binary number	10100	10101	10110	10111	11000	11001	11010	11011	11100	11101	...
Decimal code	24#	024#	24#1	024#1	4#5	04#5	4#51	04#51	4#53	04#53	...

Note that the structure of these scrambled codes is strangely similar to the crust-core-mantle paradigm: in both cases the symbols are, in some sense, unimodally sorted — first ascending, then descending.

In fact, we can apply to the decomal codewords any deterministic, one-to-one, and length-preserving transformation, without disturbing the word-length distribution. For example, since the digits after the # marker are all od, we can subtract 1 from them:

Binary number	10100	10101	10110	10111	11000	11001	11010	11011	11100	11101	...
Decimal code	24#	024#	24#0	024#0	4#4	04#4	4#40	04#40	4#42	04#42	...

★ [Mention Rene's suggestion that the letters in each word were sorted.]

★ [Mention that the uniformity and independence of the gallows and bench traits also has parallels in the decimal code above]

★ [Recall that O-slots can be filled/unfilled with 50% probability. Does this help us understand the model?]

Note that if the decimal codes were assigned to the words at random, or in alphabetical order, the *token* length distribution would be fairly symmetrical, and similar to the word

length distribution. On the other hand, if a new code is assigned in sequence to each new word that appears in some plaintext, then the most common words will tend to have shorter codes, and the token length distribution will be biased towards the left — as in figure **??**

F Is ϟ a leader?

A natural question is whether the ϟ letter should be counted as a leader (or a mutated form of some other leader), or as an idependent trait. We may get some clues by looking at the number of words as a function of word length, for words with and without ϟ. ⋆[Recompute table 32, for words (not tokens), and looking at total word length (not just leader count).]

As we can see, crust-only words without the ϟ prefix have between 0 and 3 leaders (most often 1 or 2, 1.57 on the average). Those with ϟ have between 0 and 2 leaders (most often 1 or 2, 1.17 on the average), not counting the ϟ glyph. We could say that the ϟ prefix counts as 0.4 of a leader.

In words that have a split crust (non-empty core and/or mantle), the leaders are mostly located in the crust suffix. Here are the counts for various patterns of leaders, in words with and without ϟ-letters. (The "#" denotes the core and/or mantle component, and ʔ denotes a generic leader.)

*without*ϟ			*with*ϟ *asaffix*			*with*ϟ *asleader*		
5130	0.25594	#	1277	0.27713	ϟ#			
10572	0.52744	#ʔ	3100	0.67274	ϟ#ʔ			
820	0.04091	ʔ#	45	0.00977	ϟʔ#	1277	0.27713	ϟ#
1565	0.07808	#ʔʔ	144	0.03125	ϟ#ʔʔ			
1579	0.07878	ʔ#ʔ	38	0.00825	ϟʔ#ʔ	3100	0.67274	ϟ#ʔ
59	0.00294	ʔʔ#	0	0.00000	ϟʔʔ#	45	0.00977	ϟʔ#
112	0.00559	#ʔʔʔ	2	0.00043	ϟ#ʔʔʔ			
103	0.00514	ʔ#ʔʔ	1	0.00022	ϟʔ#ʔʔ	144	0.03125	ϟ#ʔʔ
94	0.00469	ʔʔ#ʔ	1	0.00022	ϟʔʔ#ʔ	38	0.00825	ϟʔ#ʔ
1	0.00005	ʔʔʔ#	0	0.00000	ϟʔʔʔ#	0	0.00000	ϟʔʔ#
2	0.00010	#ʔʔʔʔ						
0	0.00000	ʔ#ʔʔʔ				2	0.00043	ϟ#ʔʔʔ
3	0.00015	ʔʔ#ʔʔ				1	0.00022	ϟʔ#ʔʔ
3	0.00015	ʔʔʔ#ʔ				1	0.00022	ϟʔʔ#ʔ
0	0.00000	ʔʔʔʔ#				0	0.00000	ϟʔʔʔ#
1	0.00005	ʔʔ#ʔʔʔ						

If we view the ϟ letter as an independent affix (second column), the distribution of leader patterns in ϟ-words seems similar to that of words without ϟ (first column), except for a noticeable bias in the former towards shorter words. Note in particular that #ʔ and ϟ#ʔ

are the most popular patterns in the two classes. On the other hand, if we try to view 4 as a leader (third column), the distributions don't match at all. Thus the first interpretation seems to be the most correct of the two.

G The mantle structure

Again, after ignoring circles, mapping ʔτ to ɑ, and mapping all gallows to #, the most common core/mantle combinations in this class are

withoutplatform				*withplatform*			
5820	0.38477		#	737	0.37335	ɔ	#τ
2160	0.14280		#ɔ	295	0.14944	ɔ	#τɔ
2339	0.15463		#ɔɔ	44	0.02229	ɔ	#τɔɔ
189	0.01250		#ɔɔɔ	2	0.00101	ɔ	#τɔɔɔ
4	0.00026		#ɔɔɔɔ				
1611	0.10651		#ɑ	8	0.00405	ɔ	#τɑ
1102	0.07285		#ɑɔ				
101	0.00668		#ɑɔɔ				
2	0.00013		#ɑɔɔɔ				
88	0.00582		#ɔɑ				
40	0.00264		#ɔɑɔ				
2	0.00013		#ɔɑɔɔ				
27	0.00179		#ɔɔɑ				
6	0.00040		#ɔɔɑɔ				
11	0.00073		#ɑɑ				
1	0.00007		#ɑɑɔ				
6	0.00040		#ɑɔɑ				
502	0.03319	ɑ	#	514	0.26039	ɑɔ	#τ
94	0.00621	ɑ	#ɔ	126	0.06383	ɑɔ	#τɔ
64	0.00423	ɑ	#ɔɔ	2	0.00101	ɑɔ	#τɔɔ
6	0.00040	ɑ	#ɔɔɔ				
144	0.00952	ɑ	#ɑ	1	0.00051	ɑɔ	#τɑ
36	0.00238	ɑ	#ɑɔ				
5	0.00033	ɑ	#ɑɔɔ				
3	0.00020	ɑ	#ɔɑ				
2	0.00013	ɑ	#ɑɑ				
355	0.02347	ɑɔ	#	183	0.09271	ɑɔɔ	#τ
69	0.00456	ɑɔ	#ɔ	45	0.02280	ɑɔɔ	#τɔ
35	0.00231	ɑɔ	#ɔɔ	1	0.00051	ɑɔɔ	#τɔɔ
2	0.00013	ɑɔ	#ɔɔɔ				
51	0.00337	ɑɔ	#ɑ				
18	0.00119	ɑɔ	#ɑɔ				
2	0.00013	ɑɔ	#ɑɔɔ				
88	0.00582	ɑɔɔ	#	4	0.00203	ɑɔɔɔ	#τ
12	0.00079	ɑɔɔ	#ɔ	3	0.00152	ɑɔɔɔ	#τɔ
11	0.00073	ɑɔɔ	#ɔɔ	1	0.00051	ɑɔɔɔ	#τɔɔ
5	0.00033	ɑɔɔ	#ɑ				
2	0.00013	ɑɔɔ	#ɑɔ				
49	0.00324	ɔ	#	3	0.00152	ɔɔ	#τ
15	0.00099	ɔ	#ɔ				
14	0.00093	ɔ	#ɔɔ				

60

Note that we have sorted this table as if the single ⟨c⟩ following the core was part of the mantle suffix. As the table shows, prefixes are generally shorter than suffixes, and, for a given prefix or suffix, the frequency generally decreases as the other affix gets more complicated.

The dilemma of the mantle structure is illustrated in the following pages, which show the same distribution of split core-mantles above in different formats:

- `mantle1.html`: Sorted by total length, ignoring platform.
- `mantle2.html`: Sorted by total length, including platform.
- `mantle3.html`: Parsing the ⟨c⟩ as part of the core.

H Conclusions

It is hard to resist the impression that the Voynichese tokens are indeed words of the language (or at least 'units of meaning' of some sort).

References

[1] Robert Firth. ??? http://www.research.att.com/ reeds/voynich/firth/24.txt, 1995.

[2] Jacques B. M.Guy. The distribution of letters ⟨*c*⟩ and ⟨*o*⟩ in the Voynich Manuscript: Evidence for a real language? *Cryptologia*, XXI(1):51–54, January 1997.

[3] David Kahn. *The Codebreakers.* Macmillan, 1967.

[4] Gabriel Landini. Zipf's laws in the Voynich Manuscript. WWW document at //web .bham.ac.uk/G.Landini/, file evmt/zipf.htm, November 1997.

[5] Mike Roe. ???, ¡1997? message to the Voynich mailing list.

[6] J. Stolfi. The generalized chinese theory. http://www.dcc.unicamp.br/ stolfi/voynich /97-11-23-tonal/, 1997.

[7] J. Stolfi. The voynich manuscript. http://www.dcc.unicamp.br/ stolfi/voynich/99-07-31-cbm99-slides/, July 1997. transparencies from a talk presented at the Brazilian Mathematics Coloquium.

[8] J. Stolfi. OKOKOKO: The fine structure of voynichese words. http://www.dcc.unicamp .br/ stolfi/voynich/Notes/017/Note-017.html, 1998.

[9] J. Stolfi. ???? Messages to the Voynich mailing list, 13.jun.2000, June 2000.

[10] Jorge Stolfi. A prefix-midfix-suffix decomposition of Voynichese words. WWW document at //www.dcc.unicamp.br/~stolfi/, file voynich/97-11-12-pms/, December 1997.

[11] Jorge Stolfi. Scatterplots of VMs pages. WWW document at //www.dcc.unicamp.br/, file ~stolfi/voynich/98-06-19-page-plots/, July 1998.

[12] Jorge Stolfi. Where are the bits? Local entropy distribution of various languages . WWW document at //www.dcc.unicamp.br/, file ~stolfi/voynich/98-07-09-local-entropy/, July 1998.

[13] Brig. J.Tiltman. Untitled remarks, 1951. Reproduced in D'Imperio, Fig.27.