Properties of the Voynich Manuscript
Part 2: Word Structure

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

107-abstract 2025-04-25 07:21:56 by stolfi

Abstract

We give here a paradigm (combinatorial description) of 'typical' words from the Voynich Manuscript (VMS), namely a fairly restrictive grammar whose language contains ■% 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.

Contents

1 Introduction **2**

2 Word paradigms **2**
2.1 Tiltman's paradigm 3
2.2 Mike Roe's paradigm 3
2.3 Robert Firth's paradigm 4

3 The new word paradigm **5**
3.1 The fine structure of Voynichese words 5

4 Justifying the fine structure model **6**
4.0.1 The crescent glyph 7
4.0.2 Multiple crescent glyphs 13

4.0.3 The circle glyphs . . . 14
4.1 The layer model . . . 15
4.1.1 Unimodality . . . 15
4.1.2 The initial element . . . 15
4.2 The final elements . . . 16
4.2.1 Abbreviation letters . . . 16
4.2.2 The leaders . . . 16
4.2.3 Leader distribution . . . 17
4.2.4 The mantle layer . . . 18
4.2.5 The core layer . . . 18
4.2.6 Distribution of the circles . . . 19

5 Abnormal words **21**

6 Sectional variariation **22**

7 Discussion and conjectures **22**

8 Conclusions **23**

A A grammar for Voynichese words **23**
A.1 Probabilistic models . . . 23
A.2 Grammar notation . . . 24
A.3 Why the frequencies? . . . 25
A.4 Predicting word frequencies . . . 25
A.5 Why a grammar? . . . 26
A.6 Implied word structure . . . 26
A.7 Coverage versus simplicity . . . 27

B Normal and abnormal words **27**

C Is ⟊ a leader? **28**

D The mantle structure **30**

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. Its history, nature, and basic statistics were summarized in a previous report [**?**]. These features strongly suggests that the book contains meaningful text in some unspecified natural language, encoded by a highly original notation or secret code.

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 review some previous word paradigms. In section section 3, we describe the new word model which is the main topic of this paper. ■[Confirm.]

2 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 [**?**], M. Roe [**?**], R. Firth [**?**], and the present author [**?**, **?**]. 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.

2.1 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 1:

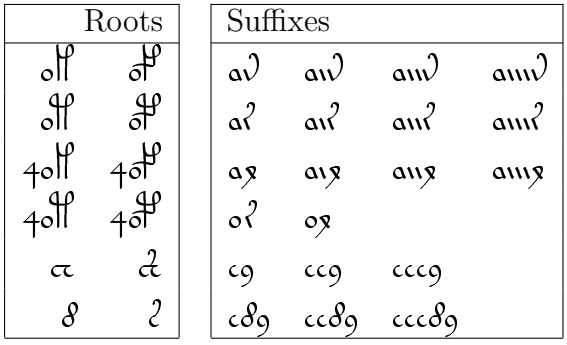


Figure 1: John Tiltman's root-suffix paradigm for VMS words.

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.

2.2 Mike Roe's paradigm

The automaton A of figure 2, devised by Mike Roe [**?**], is a typical example of those partial paradigms.

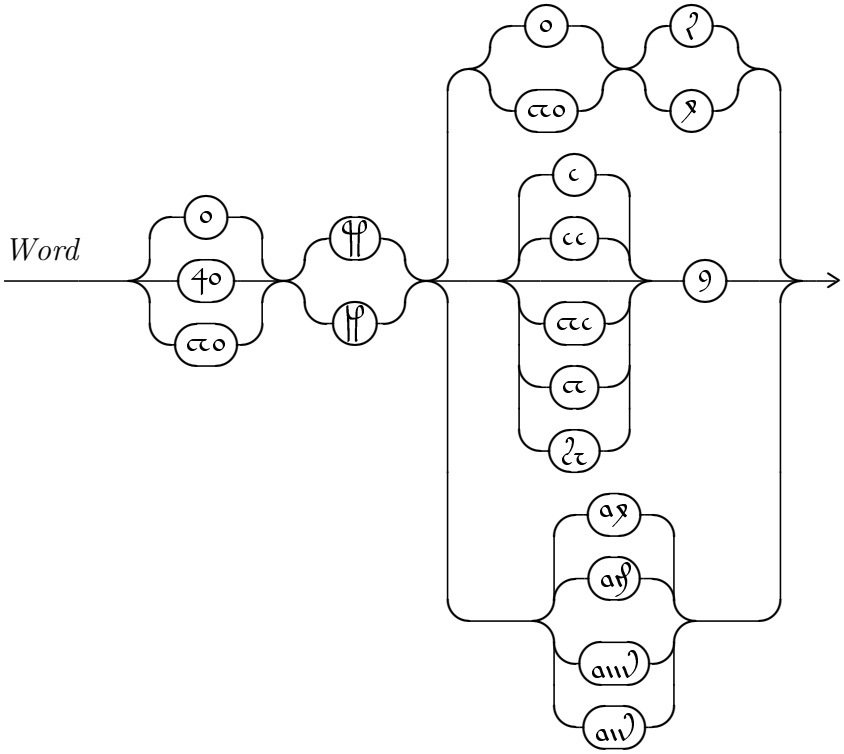


Figure 2: Mike Roe's automaton-based paradigm for VMS words.

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.

2.3 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 1.

Roots	

Suffixes	

Table 1: 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.)

3 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 [**?**].

3.1 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 **??**. The elements are listed in table 2.

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 2: The basic elements of Voynichese, according to the fine structure model. The classes are explained in section 4.1.

The fine structure model also imposes constraints on the order in which the elements of table 2 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 $O = Y \cup A = \{$ $\}$ 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 2 and formula (1) impose some non-trivial constraints on the sequence of glyphs. Specifically, it says that the *crescent glyph* occurs either in pairs, or singly after one of the elements $\{$, , , , , , , , , , $\}$. Moreover, the letters $\{$$\}$ cannot occur between a letter and its -modifier, and cannot occur next to

each other. Finally, the letter ﻉ can occur only before {ﻍ, ﻝ, ﻑ}; and the glyphs ﻝ and ﻑ may occur only in word-final position.

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

150-newparajust 2025-04-25 07:21:56 by stolfi

4 Justifying the fine structure model

Table 2 and formula (1) can be justified by the glyph pair statistics. Generally speaking, compound elements like ﻫﻛ and ﻭﻭﻝ were identified by observing that one or more of their constituent glyphs occurs almost exclusively as part of those combinations.

4.0.1 The crescent glyph

In particular, as tables **??** and **??** suggest, the crescent glyph ﻛ either follows a gallows or bench glyph (one of {ﻕ, ﻗ, ﻫ, ﻬ, ﻩ, ﻪ, ﻫﻛ, ﻬﻛ, ﻩﻛ, ﻪﻛ}), or is adjacent to another ﻛ glyph. See also tables 3 and 4.

	□	4	๑	๐	๑	๙	๙	๘	๗	๖	๕	๔	๓	๒	๑	๐	๏	๎	ํ
□	1	·	3	2	1	1	·	·	·	4	·	·	7	13	1	4	·	·	·
4	·	·	2	4	2	·	·	1	·	·	·	·	13	8	1	2	·	·	·
๑	·	·	·	1	·	·	·	1	·	·	·	·	1	·	·	·	·	·	·
๐	2	·	8	21	4	·	·	16	1	3	1	1	41	15	7	8	1	·	·
๑	·	·	·	·	1	·	·	·	·	·	·	·	2	·	·	·	·	·	·
๙	2	·	·	1	·	·	·	·	·	2	·	·	3	1	1	·	·	·	·
๘	3	·	2	4	14	·	·	6	·	2	·	·	·	·	1	·	·	1	·
๗	1	·	·	·	1	·	·	·	·	3	·	·	·	·	·	·	·	·	·
๖	5	·	1	4	10	·	·	5	1	2	·	2	1	·	·	·	·	·	·
๕ ๑	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
๔	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
๓	4	·	153	1001	865	1	1503	5	64	·	·	9	162	76	15	34	74	40	3	3
๒	37	·	62	469	448	·	753	2	29	·	·	4	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	·	·	·	·	·	·	·
๏	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
๐	·	·	·	1	·	·	·	·	·	·	·	·	1	·	·	·	·	·	·	·
๑	·	·	3	31	142	·	49	·	·	·	·	·	·	·	1	·	1	·	·	·
๒	·	·	1	40	103	·	34	1	4	·	·	·	·	·	·	·	·	·	·	·
๓	·	·	·	6	8	·	1	·	·	·	·	·	1	·	·	·	·	·	·	·
๔	·	·	2	20	19	·	17	·	1	·	·	·	·	·	·	·	1	·	·	·

Table 3: 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 ๘๔๑ in the main text.

	□	4	๑	๐	๑	๙	๗	๘	๒	๒	๗	๔	๔	๓	๓	๓	๓	๓	๓	๓	๓
□	1	·	3	2	1	1	·	·	4	·	·	1	·	7	8	1	3	·	·	·	·
4	·	·	2	4	2	·	1	·	·	·	·	·	·	9	7	1	2	·	·	·	·
๑	·	·	·	1	·	·	1	·	·	·	·	·	·	1	·	·	·	·	·	·	·
๐	2	·	7	16	4	·	9	1	3	·	1	1	·	36	14	7	8	1	·	·	·
๙	·	·	·	·	1	·	·	·	·	·	·	·	·	2	·	·	·	·	·	·	·
๗	2	·	·	1	·	·	·	·	2	·	·	·	·	3	1	1	·	·	·	·	·
๘	3	·	2	4	13	·	4	·	2	·	·	·	·	·	·	1	·	1	·	·	·
๒	1	·	·	·	1	·	·	·	3	·	·	·	·	·	·	·	·	·	·	·	·
๒	3	·	1	4	8	·	5	1	2	·	2	·	·	1	·	·	·	·	·	·	·
๓	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
๓	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
๓	4	·	50	270	123	1	207	3	25	·	·	8	3	46	32	14	18	16	9	2	2
๓	10	·	14	128	57	·	85	1	15	·	·	3	1	28	13	4	7	11	4	·	1
๓	1	·	31	144	45	1	85	1	9	·	·	39	22	·	·	·	·	·	1	·	·
๓	1	·	19	98	27	·	52	·	5	·	·	20	3	·	·	·	·	·	·	·	·
๓	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
๓	·	·	·	1	·	·	·	·	·	·	·	·	1	·	·	·	·	·	·	·	·
๓	·	·	3	15	24	·	19	·	·	·	·	·	·	·	·	1	·	1	·	·	·
๓	·	·	1	16	20	·	12	1	2	·	·	·	·	·	·	·	·	·	·	·	·
๓	·	·	·	5	4	·	1	·	·	·	·	·	·	1	·	·	·	·	·	·	·
๓	·	·	2	12	9	·	7	·	1	·	·	·	·	·	·	·	·	1	·	·	·

Table 4: Counts of glyph pairs that occur adjacent to a single ๔ glyph. The entry in row ๘ and column ๙ is the number of occurrences of ๘๔๙ in the main text's lexicon (ignoring word frequencies).

In fact, if we look at the 9582 occurrences of single ๔ glyphs (not adjacent to another ๔

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 ⟨c⟩ glyph one can find gallows, benches, circles, leaders, or finals, all in significant numbers. See tables 5 and 6. We take this asymmetry as one piece of evidence that a single ⟨c⟩ glyph is a part the preceding gallows or bench letter.

	⟨o⟩	⟨y⟩	⟨a⟩	⟨o⟩	⟨g⟩	⟨x⟩	⟨s⟩	⟨r⟩	⟨l⟩	⟨n⟩	⟨m⟩	⟨ch⟩	⟨sh⟩	⟨cth⟩	⟨ckh⟩	⟨cph⟩	⟨cfh⟩	⟨ckhy⟩	⟨cthy⟩	⟨cfhy⟩	⟨cphy⟩
prev	.	.	.	.01	.	.	.	.	.	.	.	.43	.21	.17	.10	.	.	.02	.02	.	.01
next	.01	.	.04	.26	.22	.	.37	.	.01	.	.	.01	.	.03	.01	.	.01	.01	.01	.	.

Table 5: Distribution of basic glyphs preceding and following a single ⟨c⟩ glyph in the main text.

	⟨o⟩	⟨y⟩	⟨a⟩	⟨o⟩	⟨g⟩	⟨x⟩	⟨s⟩	⟨r⟩	⟨l⟩	⟨n⟩	⟨m⟩	⟨ch⟩	⟨sh⟩	⟨cth⟩	⟨ckh⟩	⟨cph⟩	⟨cfh⟩	⟨ckhy⟩	⟨cthy⟩	⟨cfhy⟩	⟨cphy⟩
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 6: Distribution of basic glyphs preceding and following a single ⟨c⟩ glyph in the main text's lexicon (ignoring word frequencies).

More significantly, the glyph distributions just *after* gallows-⟨c⟩ and bench-⟨c⟩ pairs, such as ⟨ckhc⟩ and ⟨cthc⟩, are similar to the distributions after the corresponding unmodified gallows and benches. See table **??**. In contrast, the glyph distributions just *before* ⟨c⟩-glyph pairs, such as ⟨cs⟩, are quite unlike those of the corresponding bare glyphs. See table **??**. In other words, the ⟨c⟩ 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 ⟨c⟩ is a gallows/bench suffix modifier — one which, in fact, does not change the glyph's character very much.

	□	𝍈	𝍉	𝍊	𝍋	𝍌	𝍍	𝍎	𝍏	𝍐	𝍑	𝍒	𝍓	𝍔	𝍕	𝍖	𝍗	𝍘	𝍙	𝍚	tot
𝍒	.	.	.09	.04	.25	.07	.	.	.	.	.	.	.	.	.	.	.	.02	.	.	.	1.0
𝍒𝍜	.	.	.22	.04	.25	.37	.	.	.02	.	.	.	.	.04	.02	.	.	.02	.	.	.	1.0
𝍓	.	.	.06	.03	.22	.04	.	.	.	.	.	.	.	.	.	.	.	.02	.	.	.	1.0
𝍓𝍜	.02	.	.22	.03	.23	.37	.	.	.	.	.	.	.	.05	.	.	.	.03	.	.	.	1.0
𝍔	.	.	.08	.30	.07	.	.	.	.	.	.	.11	.02	.	.	.	.	.	.	.	.	1.0
𝍔𝍜	.	.	.17	.04	.32	.40	.	.	.	.	.	.05	.	.	.	.	.	.	.	.	.	1.0
𝍕	.	.	.08	.26	.12	.	.	.	.	.	.	.18	.03	.	.	.	.	.	.	.	.	1.0
𝍕𝍜	.	.	.15	.03	.33	.44	.	.	.	.	.	.03	.	.	.	.	.	.	.	.	.	1.0
𝍖	.	.	.51	.04	.11	.04	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1.0
𝍖𝍜	.	.	.63	.	.14	.22	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1.0
𝍗	.	.	.40	.08	.25	.03	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1.0
𝍗𝍜	.	.	.56	.	.22	.19	.	.	.02	.	.	.	.	.	.	.	.	.	.	.	.	1.0

Table 7: Distributions of basic glyphs in the main text just after some digraphs ending with single 𝍜, compared to the distributions after the the corresponding 𝍜-less glyph.

	□	c□																						
□	.	.02	.10	.	.14	.	.33	.	.29	.	.56	.	.73	.	.13	.02	.18	.09	.42	.07	.35	.03	.22	.	.55	.
4	.	.	.	.	.	.	.22	.	.	.	.	.	.	.	.	.04	.	.06	.	.04	.	.03	.	.	.	.
9	.41	.	.	.	.	.	.	.	.	.	.03	.	.02	.	.07	.	.10	.	.06	.	.05	.	.	.	.	.
a	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
o	.03	.04	.	.	.02	.02	.	.	.17	.	.02	.	.02	.03	.62	.13	.65	.11	.41	.14	.36	.22	.22	.	.15	.
8	.02	.05	.40	.	.29	.	.02	.	.	.	.03	.	.04	.	.	.	.	.	.	.	.	.03	.	.	.	.02
x	.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	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
ɑ	.	.07	.06	.42	.03	.47	.11	.41	.06	.44	.	.07	.	.09	.02	.50	.	.55	.02	.60	.	.47	.27	.58	.14	.62
ʕ	.	.65	.02	.22	.	.19	.04	.19	.	.22	.	.03	.	.03	.	.28	.	.18	.	.16	.	.16	.10	.40	.05	.35
4	.	.02	.04	.14	.21	.18	.03	.22	.	.19	.10	.63	.05	.73	.	.	.	.	.	.	.	.	.	.	.	.02
4	.	.02	.03	.07	.11	.10	.03	.13	.	.12	.09	.23	.04	.09	.	.	.	.	.	.	.	.	.	.	.	.
4	.	.	.	.	.	.	.	.	.	.	.02	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	.	.	.	.	.	.	.	.	.	.	.07	.	.02	.03	.	.	.	.	.	.	.	.	.	.	.	.
4	.	.	.03	.07	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.03	.	.	.	.
4	.	.	.02	.05	.	.	.	.02	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
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 8: 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	x	?	2	v	9	ɑ	ɔ	4	4	4	4	4	4	4	4	tot
ɑ	.	.	.10	.06	.25	.08	.	.	.02	.	.	.	.	.02	.	.	.	.	.	.	.	1.0
ɑc	.	.	.15	.06	.32	.25	.	.	.03	.	.	.	.	.06	.04	.02	.02	.02	.	.	.	1.0
ɔ	.	.	.09	.04	.23	.07	.	.	.	.	.	.	.	.02	.	.	.	.	.	.	.	1.0
ɔc	.03	.	.15	.04	.34	.22	.	.	.04	.	.	.	.	.07	.03	.	.02	.03	.	.	.	1.0
4	.	.	.07	.21	.11	.	.	.	.	.	.	.15	.05	.	.	.	.	.	.	.	.	1.0
4c	.	.	.12	.08	.38	.22	.	.	.02	.	.	.10	.06	.	.	.	.	.	.	.	.	1.0
4	.02	.	.07	.17	.16	.	.	.	.	.	.	.20	.07	.	.	.	.	.	.	.	.	1.0
4c	.	.	.12	.08	.44	.23	.	.	.02	.	.	.09	.	.	.	.	.	.	.	.	.	1.0
4	.02	.	.27	.07	.21	.06	.	.	.02	.	.	.	.	.	.	.	.	.	.	.	.	1.0
4c	.	.	.38	.05	.24	.30	.	.	.	.	.	.	.	.	.	.02	.	.02	.	.	.	1.0
4	.02	.	.19	.13	.28	.06	.	.	.02	.	.	.	.	.	.	.	.	.	.	.	.	1.0
4c	.	.	.38	.02	.31	.23	.	.02	.04	.	.	.	.	.	.	.	.	.	.	.	.	1.0

Table 9: 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□																						
□	.	.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
x	.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	.	.	.	.	.	.	.	.	.	.	.	.	.	.
ɔ	.12	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
ɣ	.05	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
ɑ	.	.14	.08	.36	.05	.37	.12	.37	.08	.42	.	.11	.	.10	.03	.34	.03	.43	.04	.47	.	.47	.13	.53	.11	.60
ɔ	.	.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	.	.	.	.	.	.	.	.	.	.	.	.
ᚠ	.	.	.	.	.	.	.	.	.	.	.04	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
ᚠ	.	.	.	.	.03	.	.02	.	.	.	.10	.	.05	.03	.	.	.	.	.	.	.	.	.	.	.	.
ᚠc	.	.	.02	.07	.	.02	.	.02	.	.04	.	.	.	.	.	.	.	.	.	.	.	.03	.	.03	.	.
ᚠc	.	.	.	.06	.	.	.	.02	.	.02	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
ᚠc	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
ᚠc	.	.	.	.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 10: 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.

4.0.2 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 ct or ch.

The reader may have noticed that the inclusion of both cc and ccc creates ambiguities in the parsing of some words; for instance, ctcccg could be parsed as either ct·ccc·g or ctc·cc·g. Observing that groups like pctc are far more common than pcct, we arbitrarily chose to resolve the ambiguity by parsing pccc as p·ccc rather than pc·cc.

4.0.3 The circle glyphs

The "circle" glyphs {a, o, g} 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 c 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 {y, ol, aiin} and the letters q and x, are almost always preceded by a or o. In particular, the words {aq, oq, ax, ox} are quite common, while {qaqo, xa, xo} are essentially non-existent. On the other hand, the glyphs 4 and y are usually followed by a circle letter, but rarely preceded by one.

	□	4	c	x	y	q	l	ol	y	ct	ch	p	p	f	f	ckh	ckh	ckh	ckh
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 11: Distribution in the main text of basic glyphs adjacent to a single o glyph.

	□	4	c	x	y	q	l	ol	y	ct	ch	p	p	f	f	ckh	ckh	ckh	ckh
prev	.15	.	.03	.03	.30	.05	.04	.	.	.04	.01	.22	.11	.	.01	.	.01	.	.
next	.	.	.	.22	.	.23	.	.01	.06	.	.	.	.	.	.	.	.	.	.

Table 12: Distribution in the main text of basic glyphs adjacent to a single a glyph.

	□	4	c	x	8	2	2	9	9	ch	ch	sh	sh	ckh	ckh	cph	cph
prev	.10	.	.24	.03	.40	.01	.01	.	.	.06	.02	.04	.03	.	.	.03	.02	.	.
next	.89	.	.	.	.01	.	.	.	.	.02	.01	.04	.03	.	.	.	.	.	.

Table 13: Distribution in the main text of basic glyphs adjacent to a single 9 glyph.

	□	4	c	x	8	2	2	9	9	ch	ch	sh	sh	ckh	ckh	cph	cph
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 14: Distribution of basic glyphs adjacent to a single circle glyph ({a o: *evaby*}).

160-layermodel 2025-04-25 07:21:56 by stolfi

4.1 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 2. Throughout this section, we will ignore any occurrences of the glyphs $A \cup Y = \{\mathfrak{a}, \mathfrak{o}, \mathfrak{9}\}$; their distribution will be discussed separately in section 4.2.6. 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.

4.1.1 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'.]

4.1.2 The initial element

The Q prefix, when present, consists of a single ⟨4⟩ glyph. can occur only at the beginning of a normal word, although in a few instances (less than 0.4% of all ⟨4⟩s,) it is preceded by ⟨o⟩ or ⟨9⟩.

The letter ⟨4⟩ rarely occurs at beginning of paragraphs or in labels, which may mean that it is a grammatical particle (article, preposition, etc.).

4.2 The final elements

Elements of class N can occur only at the end of the word. They comprise the glyphs ⟨ɔ⟩ and ⟨ƒ⟩, and clusters consisting of one to four ⟨ı⟩ glyphs, followed by one of the letters {⟨ɔ⟩, ⟨ƒ⟩, ⟨8⟩, ⟨x⟩, ⟨ʔ⟩, ⟨ʕ⟩}.

Actually, as shown in table 15, only a few of those 24 potential ⟨ı⟩-containing clusters occur in significant numbers.

Class	Elements															
F	⟨ıı8⟩	5	.0008	⟨ıı9⟩	28	.0044	⟨ıiʔ⟩	581	.0923	⟨ıiʕ⟩	11	.0018	⟨iiɔ⟩	1324	.2104	⟨iiƒ⟩	49	.0078
	⟨iii8⟩	9	.0014	⟨iii9⟩	12	.0019	⟨iiiʔ⟩	132	.0210	⟨iiiʕ⟩	7	.0011	⟨iiiɔ⟩	4016	.6381	⟨iiiƒ⟩	13	.0021
	⟨iiii8⟩	1	.0002	⟨iiii9⟩	1	.0002	⟨iiiiʔ⟩	1	.0002	⟨iiiiʕ⟩	·	.	⟨iiiiɔ⟩	103	.0164	⟨iiiiƒ⟩	·	.
	⟨iiiii8⟩	·	.	⟨iiiii9⟩	·	.	⟨iiiiiʔ⟩	·	.	⟨iiiiiʕ⟩	·	.	⟨iiiiiɔ⟩	1	.0002	⟨iiiiiƒ⟩	·	.

Table 15: All the potential final elements of Voynichese.

The asymmetry between ⟨ıiʔ⟩ and ⟨ıiʕ⟩ is puzzling, considering that ⟨ʔ⟩ and ⟨ʕ⟩ are similar in other respects. Also disconcerting is the fact that the glyphs ⟨ɔ⟩ and ⟨ƒ⟩ are almost exclusively word-final, whereas ⟨ʔ⟩ occurs both internally and as part of ⟨ıiʔ⟩ and ⟨iiiʔ⟩ elements. However such asymmetries are common in natural languages.

4.2.1 Abbreviation letters

It seems that the letter ﬗ is inordinately common at the end of lines, and before interruptions in the text due to intruding figures. The letter ﬗ, like the N groups, is almost always preceded by ᥀ or ᥆ (862 tokens in 950, 91%). We note also that ᦟ᥀ﬗ and ᥀ﬗ are the most common -᥀ﬗ words, just as ᦟ᥀ᥐᥣ and ᥀ᥐᥣ are the most common -᥀ᥐᥣ words. Perhaps ﬗ is an abbreviation for ᥐᥣ (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 ﬗ, 56 (5.8%) are preceded by ᥀ᥑ or ᥀ᥑ rather than ᥀ alone.

The rare letter ᦟ, like ﬗ, occurs almost exclusively at the end of words (24 tokens out of 27); however, unlike ﬗ, it is *not* preceded by ᥀. We note that ᦟ looks like an ﬗ, except that the leftmost stroke is rounded like that of an ᥀. Perhaps ᦟ is an abbreviation of ᥀ﬗ?

There are 32 tokens that end in ﬗ, but not as ᥀ﬗ, ᥆ﬗ, or ᥑﬗ. It is possible that these tokens are actually instances of ᦟ that were incorrectly transcribed as ﬗ — a fairly common mistake.

4.2.2 The leaders

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

After the intial and final groups, the next inner layer consists of leaders — the letters $D = \{ᦠ, ᦡ, ᦢ, ᦣ, ᦤ\}$ — with their $\{᥀, ᥆, ᥉\}$, 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 5).

4.2.3 Leader distribution

Table 16 shows the distribution of number of leaders in words without mantle or core, tabulated separately for words with and without the initial ᦵ letter:

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

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

Table 16: Distribution of number of leaders in words without core and mantle, with and without 𐤀. Each # represents a leader

In words that have a non-empty mantle or core, the crust is divided in two blocks. Table 17 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 17: 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 17, 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.

4.2.4 The mantle layer

The mantle layer consists primarily of the "bench" letters: 𐤒 and 𐤓, and the 𐤊𐤊 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 𐤊 letters, except those that follow a core letter; and any 𐤍 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 c letters and circles. If we ignore the latter, and replace ch by q, the most common combinations in normal words are:

68	0.00799	c	3292	0.38661	q			
185	0.02173	cc	3851	0.45226	qc			
90	0.01057	ccc	917	0.10769	qcc			
2	0.00023	cccc	24	0.00282	qccc			
3	0.00035	cq	42	0.00493	qcq	17	0.00200	qccq
2	0.00023	cqcc	7	0.00082	qcqc	2	0.00023	qccqc
5	0.00059	ccq						
2	0.00023	ccqcc						

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 c 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 ccc groups ought to be parsed as c.cc, cc.c, or neither; and so on.

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

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

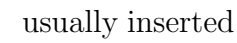
Almost exactly half of the normal words have an empty core, while the other half has a core that consists of a *single* gallows letter, possibly with platform. There are

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 .

4.2.6 Distribution of the circles

Up to now we have ignored the presence of the "circle" letters {a, o, y} 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 y (with very few exceptions)] only occurs at in word-initial or word-final position. Yet the grammar indifferently allows either y, o or a at any slot within the crust layer, and either y 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

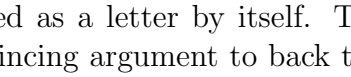
different "leaders", which can be inferred from the digraph and trigraph statistics:

21		6		18	
394		1		44	
27		2		63	
21		1		23	
75		730		199	
30		72		152	
12		126		103	
4		95		133	
11		803		127	
35		69		156	
1		107		61	
2		121		68	
179		7		4	
396		2		17	
45		2		7	
28		1		16	

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

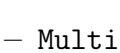
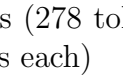
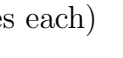
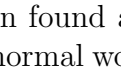
Circles are less common within the mantle layer, but fairly common at the boundaries of those two layers. Again, the present version of the grammar doesn't try to capture these nuances: it allows an optional circle before every core or mantle letter.

On the other hand, the grammar does impose some restrictions about the circle slots just before an IN group (where only ◌ and ◌ are allowed), before ◌ and ◌ (where only ◌ is allowed), before other core or mantle letters (where only ◌ or ◌ 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 ◌ glyph) is parsed as a letter by itself. Thus  is parsed as . 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.

5 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 ꞇ after an ꞔ (28 tokens). The most common is ꞔꞇꞇꞔ (3 tokens).
- `WeirdOther`: abnormal words that did not seem to fit in any of the above categories (213 tokens). Apart from isolated letters like ꞌ (7 tokens) and ꞅ (4 tokens) — mainly in the circular text on page f57v — the most common are ꞩꞏ (6 tokens), ꞏꞇꞏ, ꞔꞏ, and ꞔꞇꞏ (3 tokens each). Note that the latter are probably the result of misreading ꞑ as ꞏ 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. ꟼ + ꞏꞩꞏꞔꞔ = ꟼꞏꞩꞏꞔꞔ), or prepending a spurious non-gallows letter to a suitable normal word (e.g. ꞩ+ ꞇꞇꞑ = ꞩꞇꞇꞑ). Indeed, it is quite possible that most of the normal-looking line-initial words are in fact such "crypto-Grove" words.

180-secvar 2025-04-25 07:21:56 by stolfi

6 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`.

190-discussion 2025-04-25 07:21:56 by stolfi

7 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. [**?**]

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 con-

secutive Voynichese words, reminiscent of the Turkish/Hungarian "vowel harmony" rule. [**?**])

195-conclusions 2025-04-25 07:21:56 by stolfi

8 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).

210-wdgrammar 2025-04-25 07:21:56 by stolfi

A A grammar for Voynichese words

A.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.

A.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 ꞩaiaiv uses the rule R → ꞩ once, and R → i twice; therefore, 100 occurrences of ꞩaiaiv in the text would count as 100 uses of R → ꞩ and 200 of R → i.

A.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.

A.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, ſolccſ9 and ſolſaiin are all very popular (866, 305, 266 occurrences, respectivey), while ſccſ9 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 -ccſ9.

A.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.)

A.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.

A.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  ʃ
       9  0.00153  1.00000  ʔ
```

These rules do not accomodate words containing ııı, ıπ, or ıʃ — like ɔıııɔ ʔɔ‖ɑıπ, or ʃɑıʃ (1 occurrence each). Yet ııı with count of 1 would be a logical extrapolation of the ı series; and, in other contexts, ʃ and π clearly belong to the same class as ʔ, ɣ, ʔ.

220-wdnormab 2025-04-25 07:21:56 by stolfi

B 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 5.

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 $\{$ɕ, ɑ, ɔ, ɡ$\}$ are frequent enough to merit special attention.

C Is ⟨4⟩ a leader?

A natural question is whether the ⟨4⟩ 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 ⟨4⟩. ■[Recompute table 16, for words (not tokens), and looking at total word length (not just leader count).]

As we can see, crust-only words without the ⟨4⟩ prefix have between 0 and 3 leaders (most often 1 or 2, 1.57 on the average). Those with ⟨4⟩ have between 0 and 2 leaders (most often 1 or 2, 1.17 on the average), not counting the ⟨4⟩ glyph. We could say that the ⟨4⟩ 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 ⟨4⟩-letters. (The "#" denotes the core and/or mantle component, and ⟨?⟩ denotes a generic leader.)

without⟨4⟩			*with*⟨4⟩ *as affix*			*with*⟨4⟩ *as leader*		
5130	0.25594	#	1277	0.27713	⟨4⟩#			
10572	0.52744	#⟨?⟩	3100	0.67274	⟨4⟩#⟨?⟩			
820	0.04091	⟨?⟩#	45	0.00977	⟨4⟩⟨?⟩#	1277	0.27713	⟨4⟩#
1565	0.07808	#⟨?⟩⟨?⟩	144	0.03125	⟨4⟩#⟨?⟩⟨?⟩			
1579	0.07878	⟨?⟩#⟨?⟩	38	0.00825	⟨4⟩⟨?⟩#⟨?⟩	3100	0.67274	⟨4⟩#⟨?⟩
59	0.00294	⟨?⟩⟨?⟩#	0	0.00000	⟨4⟩⟨?⟩⟨?⟩#	45	0.00977	⟨4⟩⟨?⟩#
112	0.00559	#⟨?⟩⟨?⟩⟨?⟩	2	0.00043	⟨4⟩#⟨?⟩⟨?⟩⟨?⟩			
103	0.00514	⟨?⟩#⟨?⟩⟨?⟩	1	0.00022	⟨4⟩⟨?⟩#⟨?⟩⟨?⟩	144	0.03125	⟨4⟩#⟨?⟩⟨?⟩
94	0.00469	⟨?⟩⟨?⟩#⟨?⟩	1	0.00022	⟨4⟩⟨?⟩⟨?⟩#⟨?⟩	38	0.00825	⟨4⟩⟨?⟩#⟨?⟩
1	0.00005	⟨?⟩⟨?⟩⟨?⟩#	0	0.00000	⟨4⟩⟨?⟩⟨?⟩⟨?⟩#	0	0.00000	⟨4⟩⟨?⟩⟨?⟩#
2	0.00010	#⟨?⟩⟨?⟩⟨?⟩⟨?⟩						
0	0.00000	⟨?⟩#⟨?⟩⟨?⟩⟨?⟩				2	0.00043	⟨4⟩#⟨?⟩⟨?⟩⟨?⟩
3	0.00015	⟨?⟩⟨?⟩#⟨?⟩⟨?⟩				1	0.00022	⟨4⟩⟨?⟩#⟨?⟩⟨?⟩
3	0.00015	⟨?⟩⟨?⟩⟨?⟩#⟨?⟩				1	0.00022	⟨4⟩⟨?⟩⟨?⟩#⟨?⟩
0	0.00000	⟨?⟩⟨?⟩⟨?⟩⟨?⟩#				0	0.00000	⟨4⟩⟨?⟩⟨?⟩⟨?⟩#
1	0.00005	⟨?⟩⟨?⟩#⟨?⟩⟨?⟩⟨?⟩						

If we view the ⟨4⟩ letter as an independent affix (second column), the distribution of leader patterns in ⟨4⟩-words seems similar to that of words without ⟨4⟩ (first column),

except for a noticeable bias in the former towards shorter words. Note in particular that #ʔ and ɬ#ʔ are the most popular patterns in the two classes. On the other hand, if we try to view ɬ 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.

D The mantle structure

Again, after ignoring circles, mapping $\partial\tau$ to α, and mapping all gallows to #, the most common core/mantle combinations in this class are

withoutplatform				*withplatform*			
5820	0.38477		#	737	0.37335	σ	#τ
2160	0.14280		#ι	295	0.14944	σ	#$\tau\iota$
2339	0.15463		#$\iota\iota$	44	0.02229	σ	#$\tau\iota\iota$
189	0.01250		#$\iota\iota\iota$	2	0.00101	σ	#$\tau\iota\iota\iota$
4	0.00026		#$\iota\iota\iota\iota$				
1611	0.10651		#α	8	0.00405	σ	#$\tau\alpha$
1102	0.07285		#$\alpha\iota$				
101	0.00668		#$\alpha\iota\iota$				
2	0.00013		#$\alpha\iota\iota\iota$				
88	0.00582		#$\iota\alpha$				
40	0.00264		#$\iota\alpha\iota$				
2	0.00013		#$\iota\alpha\iota\iota$				
27	0.00179		#$\iota\iota\alpha$				
6	0.00040		#$\iota\iota\alpha\iota$				
11	0.00073		#$\alpha\alpha$				
1	0.00007		#$\alpha\alpha\iota$				
6	0.00040		#$\alpha\iota\alpha$				
502	0.03319	α	#	514	0.26039	$\alpha\sigma$	#τ
94	0.00621	α	#ι	126	0.06383	$\alpha\sigma$	#$\tau\iota$
64	0.00423	α	#$\iota\iota$	2	0.00101	$\alpha\sigma$	#$\tau\iota\iota$
6	0.00040	α	#$\iota\iota\iota$				
144	0.00952	α	#α	1	0.00051	$\alpha\sigma$	#$\tau\alpha$
36	0.00238	α	#$\alpha\iota$				
5	0.00033	α	#$\alpha\iota\iota$				
3	0.00020	α	#$\iota\alpha$				
2	0.00013	α	#$\alpha\alpha$				
355	0.02347	$\alpha\iota$	#	183	0.09271	$\alpha\iota\sigma$	#τ
69	0.00456	$\alpha\iota$	#ι	45	0.02280	$\alpha\iota\sigma$	#$\tau\iota$
35	0.00231	$\alpha\iota$	#$\iota\iota$	1	0.00051	$\alpha\iota\sigma$	#$\tau\iota\iota$
2	0.00013	$\alpha\iota$	#$\iota\iota\iota$				
51	0.00337	$\alpha\iota$	#α				
18	0.00119	$\alpha\iota$	#$\alpha\iota$				
2	0.00013	$\alpha\iota$	#$\alpha\iota\iota$				
88	0.00582	$\alpha\iota\iota$	#	4	0.00203	$\alpha\iota\iota\sigma$	#τ
12	0.00079	$\alpha\iota\iota$	#ι	3	0.00152	$\alpha\iota\iota\sigma$	#$\tau\iota$
11	0.00073	$\alpha\iota\iota$	#$\iota\iota$	1	0.00051	$\alpha\iota\iota\sigma$	#$\tau\iota\iota$
5	0.00033	$\alpha\iota\iota$	#α				
2	0.00013	$\alpha\iota\iota$	#$\alpha\iota$				

Note that we have sorted this table as if the single ɕ 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 ɕ as part of the core.