Properties of the Voynich Manuscript
Part 3: Word Structure

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

2005-todo 2025-05-04 23:39:47 by stolfi

TO DO

Character frequency Zipf-plots with and without space (Antoine Casanova)
Remove "fine structure" and "OKOKO"

2007-abstract 2025-04-28 20:42:28 by stolfi

Abstract

We give here a paradigm (combinatorial description) of "typical" words from the Voynich Manuscript (VMs), namely a fairly restrictive probabilistic grammar whose generated language contains ■% the word occurrences of the manuscript (■% of all distinct words). The grammar also predicts frequency counts for glyphs, glyph tuples, and words. We compare these predicted frequencies with the observed frequencies in the text, and discuss the implications of the discrepancies.

Contents

1 Introduction **3**
1.1 Nomenclature 3
1.2 Implications of the new paradigm 4

2 Word paradigms **4**
2.1 Tiltman's paradigm 4
2.2 Mike Roe's paradigm 5
2.3 Robert Firth's paradigm 6

3 The new word paradigm **7**
3.1 Grouping glyphs into elements 8
3.2 The OKOKO submodel 9
3.3 The layer submodel 9
3.4 Consequences of the new paradigm 10
3.5 Diagram format 11

1 Introduction

The text of the VMs is composed of discrete symbols (*glyphs*), enumerated in table 1, which are clearly partitioned into word-like strings by fairly distinct inter-word spaces and line breaks.

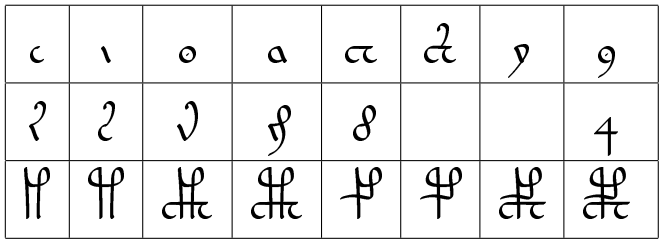


Table 1: The 22 basic glyphs of the Voynichese script that occur with significant frequencies.

Early on it was observed that the Voynichese words have a non-trivial internal structure, manifested by constraints on the number and sequence of theglyphs within each word. A number of structural models (*paradigms*) that attempt to capture these constraints have been proposed.

This note describes new paradigm that is significantly more detailed, comprehensive, and accurate than previous ones.

1.1 Nomenclature

The *lexicon* a corpus (set of texts) is the set of all distinct words that occurs in it. The members of the lexicon will be called here *lexemes* (or *word types*). A *token* is a specific occurrence of a lexeme in a corpus. The *count* of a lexeme in a corpus is the number of tokens that are occurrences of it. Its *frequency* is the count divided by the number of words in the corpus.

A paradigm is a formula or algorithm that classifies strings of letters as either *valid* or *invalid*. With respect to a given corpus, the paradigm is said to be *permissive* if it defines as valid some strings that are not lexemes, and*incomplete* if it defines as invalid some strings that are in fact are lexemes.

The lexicon of the VMS is rather irregular, so any simple paradigm will necesarily be either permissive or incomplete, or both. That kind of irregularity is the norm for natural languages; but may also be a sampling artifact, due to the relatively small size of the text (fewer than 40'000 words).

1.2 Implications of the new paradigm

The nature and complexity of the new paradigm, and its fairly uniform fit over all sections of the manuscript, can be viewed as evidence that the text has significant contents of some sort.

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 [**?**].

The new paradigm imposes severe constraints on possible decipherment theories. It apparently excludes the possibility that the text is a Vigenère-style cipher, or was generated by a random process, or is a simple transliteration of a text in a "Western" language.⋆[Confirm.]

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.

2 Word paradigms

Before describing our new word structure model [**?**, **?**], we briefly review some previous paradigms, namely those proposed by J. Tiltman [**?**], M. Roe [**?**], and R. Firth [**?**].

To describe the paradigms, we borrow some notation from formal language theory [**?**] to describe sets of strings. In particular, if X and Y are sets of strings, we denote by XY the set of all strings that are concatenation of a string from X and a string from Y. We also use X^* to mean the set of all concatenations of zero or more strings from X, and $X^?$ to mean at most one string from X — i.e. $\{()\} \cup X$, where $()$ denotes the empty string. Finally we use $X^{??}$ to mean the concatenation of at most two strings from X, that is, $\{()\} \cup X \cup XX$,

2.1 Tiltman's paradigm

One of the earliest paradigms is due to J. Tiltman, a British cryptographer who analyzed the Starred 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:

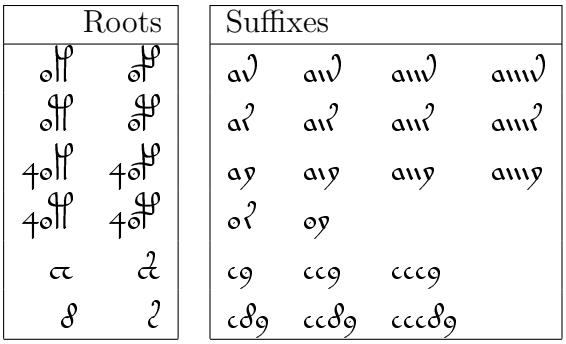


Figure 1: John Tiltman's root-suffix paradigm for VMs words.

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 3, devised by Mike Roe [**?**], is a typical example of those partial paradigms.

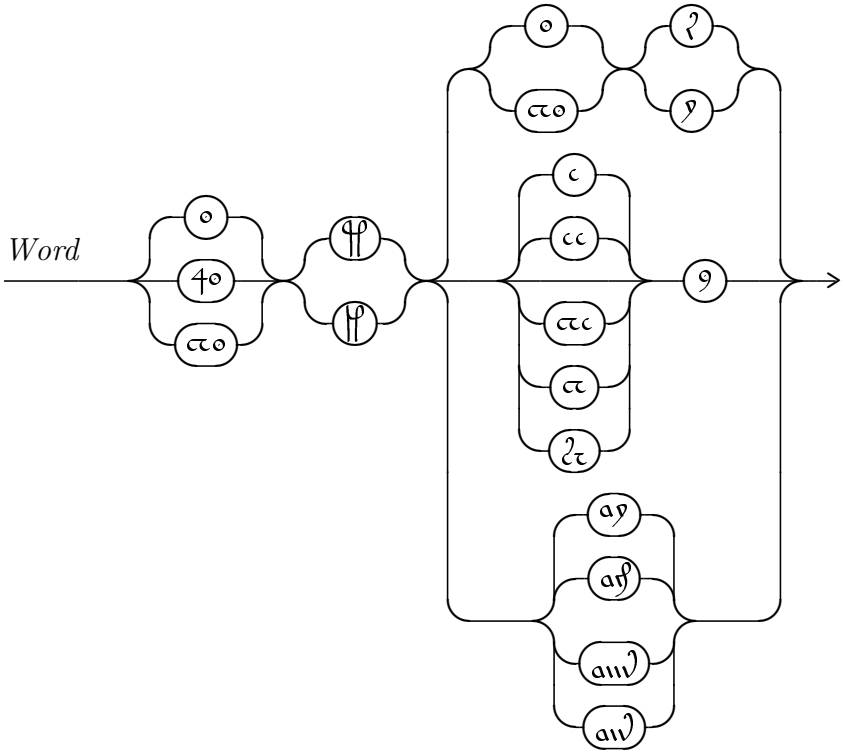


Figure 2: Mike Roe's automaton-based paradigm for VMs words.

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 2.

Roots		Suffixes	

Table 2: 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.)

2040-newparadigm 2025-04-28 20:42:28 by stolfi

3 The new word paradigm

3.1 Grouping glyphs into elements

The new paradigm, like the previous ones, states that most words are built from a small set of *elements*, each consisting of 1 to 4 of the basic glyphs of table **??**. The elements are listed in table 3.

Class	Elements										
A		13538	.0998		23689	.1747					
Y		16837	.1242								
Q		5133	.0379								
D		12417	.0916		10001	.0738		6383	.0471		2355	.0174
X		6370	.0470		2306	.0170		4100	.0302			
		4063	.0300		2029	.0150		339	.0025			
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
N		133	.0010		991	.0073						
		1324	.0098		49	.0004		581	.0043			
		4016	.0296		13	.0001		132	.0010			
		103	.0008			.			.			

Table 3: The elements of Voynichese and their assigned classes, according to the new word paradigm. The frequencies are obtained from an old transcription by T. Takahashi [**?**].

We will refer to the elements of the set $O = Y \cup A$ as the *circles*, and to elements of class H as *gallows*, either *simple* or *platf*(*orm*). The gallows with glyphs and will be called *puffs*. The elements of class D are the *dealers* (because of their EVA codes, {d, l, r, s}) and those of class X are the *benches* (because of the shape of). Finally, the elements of class N are the *tails*. Note that these terms have different meanings depending on whether they are applied to glyphs or to elements.

⋆[Is it worth including as dealer?]

Any glyph sequence that cannot be parsed into the concatenation of elements from table 3, like , , or , is by definition invalid under the new word paradigm.

In particular, the crescent glyph may occur either in pairs, as occurrences of the element , or just once, as a modifier, after one of the elements {, , , , , , , , , , }. And the stick glyph can occur only before {, , }, and at most three of them may occur in a row.

Note that some glyphs may be parts of elements of different classes. For example, the glyph is part of elements of class N when preceded by one to three glyphs; otherwise, by itself, it is an element of class D.

Note also that while the ch modifier may follow the elements and , as well as all platform gallows, it cannot follow the simle puffs and . This well-known anomaly is still lacking an explanation. Therefore the combinations and could be excluded from the set H; but we chose to retain them in order to simplify the model.

By table 3 alone, the parsing of a word into elements is ambiguous. For instance, a hypthetical word could be parsed as [][ch][chch], [][chch][ch] or as [ch][ch][ch]. We make the parsing non-ambiguous by the additional rule that any maximal substring of n consecutive ch glyphs should be parsed as a modifier of the previous element if $n = 1$, otherwise as zero or more ch elements followed by ch or chch.

Another anomaly of the table above is the exclusion of (without preceding sticks) from the set N of tail elements. The frequencies of , , and suggest that should be a tail too. We chose to exclude it to reduce the ambiguity of the parsing – namely, to ensure that each element of a word that fits the new paradigm can be assigned to exactly one class.

3.2 The OKOKO submodel

A first constraint imposed by the new model on the structure of a Voynichese word is that there is at most one member of the set *circle* elements $O = Y \cup A = \{a, o, g\}$ before, between, or after elements of the set $K = Q \cup H \cup X \cup D \cup N$. Namely, the model implies that every Voynichese word as the form

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

We refer to this formula as the *OKOKO submodel*.

3.3 The layer submodel

We define the *skeleton* of a word as being what remains of it after erasing the circle elements $\{a, o, g\}$ The nexw paradigm also says that the skeleton of most VMs words can be parsed into according to the formula

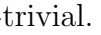
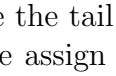
$$Q^?D^?X^?H^?X^?D^{??}N^? \tag{2}$$

We refer to this formula as the *layer submodel*.

3.4 Consequences of the new paradigm

The OKOKO sumbodel (1) says that circle elements cannot occur inside any of the K elements, including before an ch-modifier. Also, two circle elements cannot occur

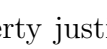
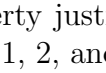
occur next to each other. Thus the words , , are rejected by that submodel.

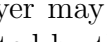
Although each factor in the layer submodel (2) may be empty, the formula is definitely non-trivial. It excludes words with two gallows like  , with two or two before a gallows, with before anything except a circle, as in  , and many more. It implies that, apart from circles, can appear only in word-initial position, while the tail elements like and may occur only in word-final position.

Suppose we assign "densities" 0 to element classes Q and N, 1 to class D, 2 to class X, and 3 to class H. Formula (2) 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 an element of maximum density in the word, in either direction, the density can only decrease (or remain constant). In particular, the sequence of densities cannot have a substring $\alpha\beta\gamma$ where $\beta < \alpha$ and $\beta < \gamma$, like 102 of 312.

This property justifies the name "layer submodel", as we can view the elements of densities 0, 1, 2, and 3 as the *atmosphere*, *crust*, *mantle* and *core* layers of a word; where each layer may be missing, on either side, but when present the layers must nest as suggested by these names.

The Q prefix, when present, consists of a single glyph. can occur only at the beginning of a normal word, although in principle it may be preceded by a circle element.

According to the layer submodel, the maximum number of elements in the skeleton is 8, achieved by any string of the form $QDXHXDDN$, such as  or  .

The OKOKO and layer submodels allow any circle (at most one) to be inserted before the skeleton, at the end of the skeleton, or at any point between two K elements. Thus in principle an allowed word can have 17 elements, for example  or .

The OKOKO and layer submodels allow the empty string. This case is declared invalid by definition. The shortest valid word is then any of the elements of table 3 by itself, such as or .

However, the number, types, and positions of the circles are qute limited in practice. For instance, while by itself is fairly common, the single-element skeleton *Ch* seems to require at least one circle, turning it into or . The rules needed to capture these constraints seem to be fairly complex, and are still being sorted out. Chiefly for that reaon, the new paradigm 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 (with very few exceptions) only occurs at in word-initial or word-final position. Yet the grammar indifferently allows either ,

or ᩋ at any slot within the crust layer, and either ᩉ or ᩈ 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.

3.5 Diagram

Formula (2) can be represented by the following syntactic diagram:

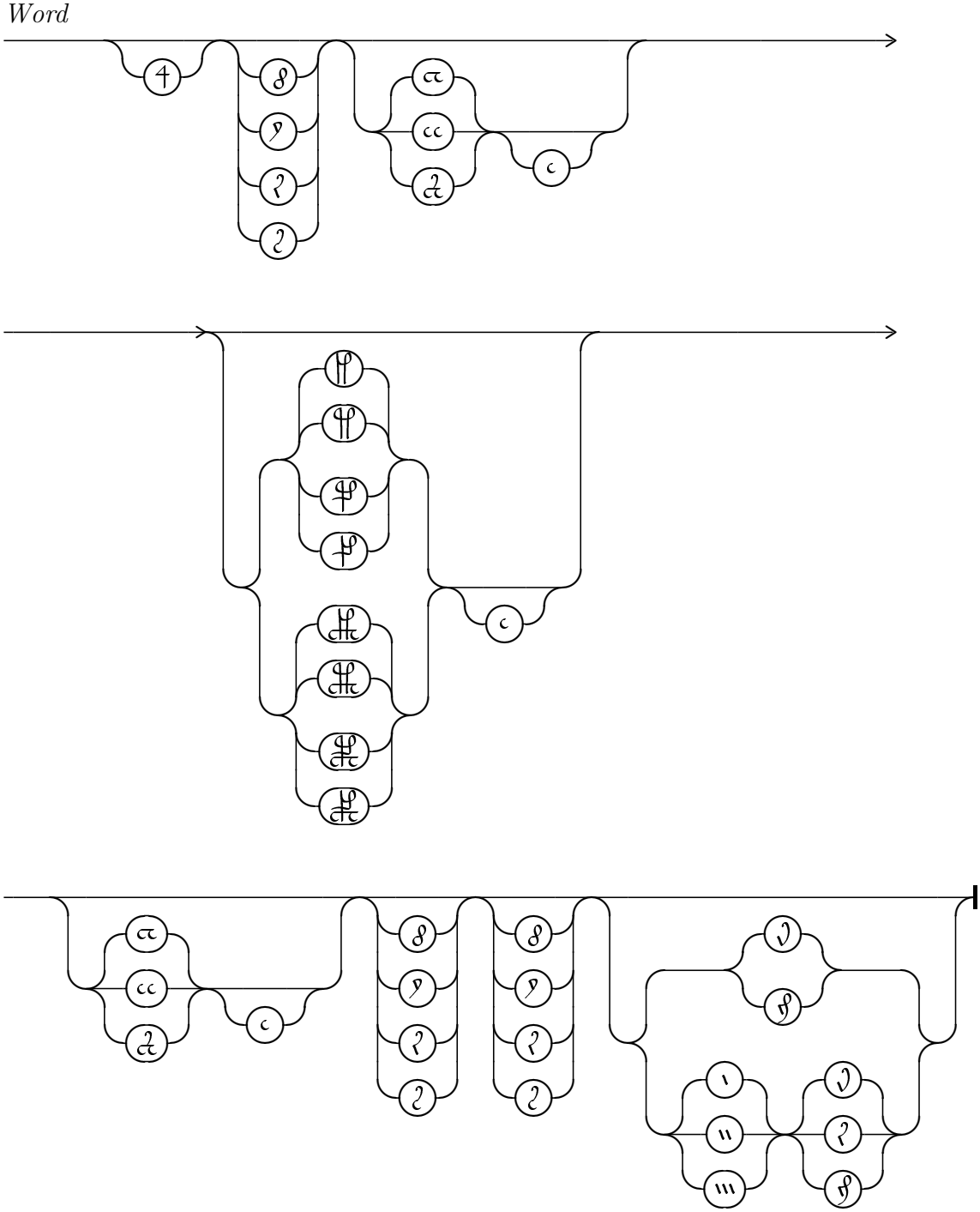


Figure 3: A representation of the layer submodel 2 as a non-deterministic automaton.