

Glyph Combinations across Word Breaks in the Voynich Manuscript

By Emma May Smith and Marco Ponzi¹

July 2018

Abstract

The text of the Voynich manuscript exhibits relationships between neighbouring words which have not formerly been explored. The last and first glyph of adjacent words show some dependency, and certain glyph combinations are more or less likely to occur. The patterns of preferences for glyph combinations demonstrate the existence of higher level glyph groups. The behaviour of the glyph combinations may arise due to change in a glyph caused by its neighbour.

Introduction

The Voynich Manuscript (Beinecke MS408) is an illustrated manuscript of 102 folios² purchased by Wilfrid Voynich at Villa Mondragone in 1912³. The manuscript has been carbon-dated to the early 1400s⁴. The manuscript's illustrations suggest that it deals with herbal, astrological, balneological, and pharmaceutical matters, among others. However, the contents of the manuscript are unreadable due to nearly all the text being written in an unknown script.

Multiple claims have been made over the last century for a complete or partial decipherment of the text, none of which are widely accepted today. Most have sought to explain the text either as a cipher or an obscure language. Other explanations propose a meaningless fraud or a procedurally generated output.

Script

The script comprises a number of distinct glyphs but also some which may be ligatures of others. Many glyphs are rare and some are unique. The majority of the text is written with a smaller subset of glyphs which occur more than fifty times. The number of glyphs in this subset differs according to the researcher, with counts ranging from 18 to 29 glyphs⁵.

1 The authors would like to thank René Zandbergen for his comments and advice on this paper.

2 Shailor, Barbara (1984) *Catalogue of medieval and renaissance manuscripts in the Beinecke Rare Book and Manuscript Library, Yale University*. Beinecke Rare Book and Manuscript Library. [Online: <https://pre1600ms.beinecke.library.yale.edu/docs/pre1600.ms408.HTM>. Retrieved 25 March 2018.]

3 D'Imperio, M E (1978). *The Voynich Manuscript: An Elegant Enigma*. National Security Agency.


4 Zyats, Paula, Erin Mysak, Jens Stenger, Marie-France Lemay, Anikó Bezur and David D. Driscoll (2016). *Physical Findings in: Clemens, Raymond (ed) The Voynich Manuscript*, Yale University Press, New Haven and London. pp. 23-37.

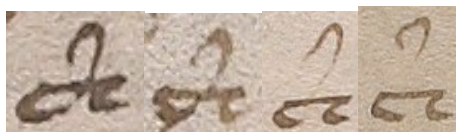
5 The First Study Group transcription represents this subset of glyphs with 18 characters; the Currier transcription uses 29. Both recognise the existence of more, rarer, glyphs. Tiltman considered two of the common glyphs to be variants of others, thus further reducing the effective glyph set to 16.

Currently the most popular form of transcription is the EVA (Extensible Voynich Alphabet) system⁶ which allows possible ligatures to be transcribed conventionally but without judgement on their nature. Our analysis of the text is based on the EVA transcription by Takeshi Takahashi⁷. The whole body of text has been considered but, since the focus of our analysis are pairs of consecutive words, single word labels did not contribute to the results.










The ongoing problem with parsing the text presents a challenge to any analysis based on textual statistics. It is necessary to define what is being counted, but different judgements will be made by individual researchers. The authors recognize that any position is contestable, but submit that resolution of the issue is outside the scope of this paper.
























For the purposes of this paper we adopt the definition of a glyph as a stroke or set of strokes which are typically contiguous while also being usually separated from other strokes or glyphs. This allows for a set of glyphs to be constructed visually without an underlying theory on the working of the script.

This definition can cause ambiguities which should be acknowledged. The glyph  has a top stroke which is connected to the other strokes in some instances and unconnected in others. However, the difference is a gradation, with intermediate instances of more or less connectedness. A series of images below shows possible variation in the same glyph.



Images 1.1-1.4: Variations on a single glyph.

Also, the glyph  is regularly connected to a following  and it could be proposed that  is a glyph in itself by our definition. However, the study in this paper is only concerned with the first and last glyphs in words. As the sequence  represents most occurrences of  it ultimately makes no difference whether the glyph is  or . Any reference to  can be construed as  if such a parsing is preferred.

For the purposes of this paper we will therefore recognise the following glyphs as analysable units:                     . Each of these glyphs occurs at least fifty times in the text; the least common being  with about 100 occurrences. We recognize that other glyphs are found within the text though in smaller numbers. The next most common is  with around 35 occurrences.

⁶ Landini, Gabriel and Zandbergen, René [Online: <http://www.voynich.nu/transcr.html#Eva>. Retrieved 26 March 2018.]

⁷ Downloaded from the Voynich Information Browser: <http://voynich.freie-literatur.de/> The text in the so-called 'Rosettes' drawing (f85v and f86r) is not included in this transcription.

Word Structure

The text is composed of strings of glyphs separated by gaps, and was written from top left to bottom right. Conventionally, the strings of glyphs are referred to as “words” and the gaps as “spaces”. There has been dispute over the assumptions inherent in such names, that words and spaces are only apparent and not real. We do not wish to enter into such a dispute in this paper, merely taking the existence of strings and gaps as observable phenomena.

The structure of words in the Voynich Manuscript has attracted comment since at least the 1950s⁸. The regularity of the structure has been noted on multiple occasions, with some researchers suggesting model “grammars”⁹. For our purposes, it is enough to note that some glyphs occur more commonly in the first or last positions of words, and thus after or before an adjacent space.

	Initial	Final		Initial	Final		Initial	Final
◦	3786	928	⌘	996	3045	𐌒	123	0
Ɱ	3023	29	𐌒	990	75	𐌒	118	23
4	2732	4	𐌒	862	53	𐌒	33	0
8	2307	577	𐌒	473	26	˘	18	7
2	2065	31	2	456	2987	8	16	84
9	1458	4377	𐌒	434	14	8	14	932
Ɱ	1260	94	𐌒	177	5	2	5	3114
2	1132	1030	˘	146	98			

Table 1: The number of occurrences in first and last position of a word for each glyph.

Table 1 above shows how many times each glyph occurs in the first or last position of a word. Some glyphs appear much more often in one position than the other, such as **Ɱ** and **2**. A few do not appear commonly in either position either due to being uncommon overall, such as **𐌒** and **8**, or almost always internal to a word, such as **˘**.

Table 2 below shows glyphs arranged by which position they are commonly found in. ‘Common’ here is defined as occurring at least 150 times in that position. Most glyphs are common in first position in a

⁸ Tiltman, John H (1967). *The Voynich Manuscript: The Most Mysterious Manuscript in the World*. NSA Technical Journal. National Security Agency. XII (3). Tiltman quotes his original report to William Friedman made in 1951.

⁹ Many such models, partial and complete, have been suggested. The most thoroughly evidenced is provided by Stolfi, Jorge (2000). *A Grammar for Voynichese Words*.

[Online: <http://www.ic.unicamp.br/~stolfi/voynich/00-06-07-word-grammar>. Retrieved 26 March 2018.]

word, but many are uncommon in the last position. A few glyphs are uncommon in either position. Because of this ꞥ, Ꞧꞧ, ꞦꞨ, ꞩ, Ɦ, and Ɜ will not be further considered in this paper.

Common in first position	Common in last position	Not common in either position
o, Ɡ, Ɬ, Ɪ, ꞯ, Ʞ, Ʇ, Ʝ, Ꭓ, Ꞵ, ꞵ, Ꞷ, ꞷ, Ꞹ, ꞹ, Ꞻ, ꞻ, Ꞽ, ꞽ, Ꞿ, ꞿ	Ʇ, Ʝ, Ꭓ, Ꞵ, ꞵ, Ꞷ, ꞷ, Ꞹ, ꞹ, Ꞻ, ꞻ, Ꞽ, ꞽ, Ꞿ, ꞿ	ꞥ, Ꞧꞧ, ꞦꞨ, ꞩ, Ɦ, Ɜ

Table 2: Common glyph positions within words.

It should also be noted that different word structures are observed when a word is adjacent to a line break¹⁰. Some glyphs are more or less common in the first or last position of a word when also in the first or last position of a line. Because of this, the following statistics will be based on only those glyphs adjacent to spaces inside lines (word breaks) and not to spaces at the start or end of lines (line breaks).

Word Break Combinations

Each space is adjacent to two words and lies immediately between the two glyphs at the end and beginning of those words. Because the structure of words restricts glyph position, only certain glyphs will occur before or after word breaks. This glyph ‘phrase’, a combination of the last glyph of the previous word and the first glyph of the following word, is the object of our study. We will call these two glyphs separated by a space, “word break combinations”.

The distribution of glyphs in the first and last positions of words, given in Table 1, suggests that the number of commonly occurring combinations is less than the total possible combinations between all glyphs in the script.

Using the number of occurrences of each glyph in first and last positions of words we can calculate the number of times that each combination is expected to occur, assuming that word order is random. This is then compared with the actual number of occurrences of each combination. The full tables of potential combinations, their expected occurrences, their actual occurrences, and the percentage difference are given in Tables 3.1-3.8 (page 6).

In order to evaluate the impact of different transcriptions, the statistics were also computed on the basis of the Zandbergen-Landini EVA transcription¹¹. The file uses two different symbols (dot and comma) to identify clear word separations and possible but ambiguous spaces. Only considering unambiguous spaces, the Zandbergen-Landini transcription contains fewer adjacent word pairs than the Takahashi transcription (30,269 vs 31,673). Considering uncertain spaces in addition to unambiguous spaces results in more word pairs than in Takahashi transcription (32,930 vs 31,673). All counts were therefore normalized to match the total number of adjacent word pairs in Takahashi transcription.

10 Currier, Prescott H (1992 [1976]). *Papers on the Voynich Manuscript*.

11 Version 1, 24 September 2017 [Online: http://www.voynich.nu/data/ZL_ivtff_1b.txt. Retrieved 9 April 2018.]

The normalized counts of the 100 word break combinations based on the ten most frequent word initial glyphs and the ten most frequent word-final glyphs were compared. In Takahashi transcription, 80 of the 100 combinations result in an actual-expected difference of 4 occurrences or more. For all these combinations, the positive or negative sign of the difference agrees with that computed on the Zandbergen-Landini transcription, both excluding and including unclear spaces. We can therefore conclude that the observed deviations from the expected counts do not depend on the specific transcription used.

The Voynich text is broadly divided into two parts which are statistically distinct. Their identification by Prescott Currier¹² has led to these two parts being referred to as Currier A and B. Although they likely represent the two poles of a textual spectrum, with the text changing continuously throughout the manuscript, it is necessary to be aware of their difference.

The chart below shows the ratio between actual and expected counts for the most common word break combinations in both Currier forms. Currier A is in light grey, Currier B in black; the 100% ratio correspond to an observed value identical to the expected one. The results show that combinations in Currier B tend to deviate wider from the expected than in Currier A. This happens both because those combinations which deviate in Currier A typically deviate even more in Currier B (such as 9.4) and because some combinations that do not show any substantial deviation in Currier A appear to deviate from the expected in Currier B (such as 2.2).

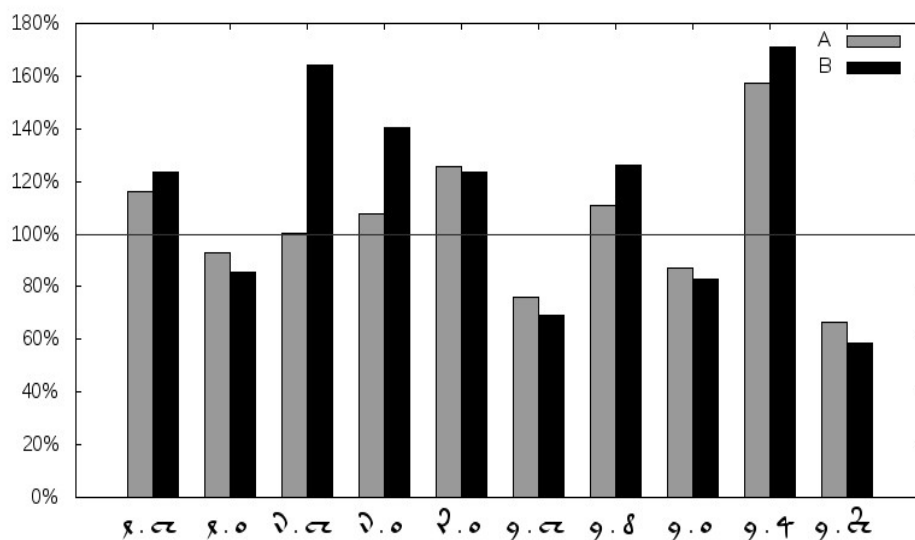


Chart 1: Word break combinations in Currier A and B.

¹² Currier, Prescott H (1992 [1976]). *Papers on the Voynich Manuscript*.

[Online: http://www.voynich.nu/extra/img/curr_main.pdf. Retrieved 26 March 2018.] These papers were presented at a seminar on the manuscript in 1976. They were collected and published in 1992.

Tables: 3.1-3.8: The expected and actual occurrences of word break combinations arranged by last glyph of preceding word.

After 。	ㄩ	ㄩ	ㄩ	ㄩ	ㄩ	ㄩ	ㄩ	ㄩ	。	ㄩ	ㄩ	ㄩ	ㄩ	ㄩ	ㄩ
Expected	35	19	5	6	16	188	99	160	238	36	64	93	25	15	43
Actual	86	30	13	22	32	112	47	89	108	32	48	153	39	77	144
Percentage	246	158	260	366	200	60	47	56	45	89	75	165	156	513	335

After ㄩ	ㄩ	ㄩ	ㄩ	ㄩ	ㄩ	ㄩ	ㄩ	ㄩ	。	ㄩ	ㄩ	ㄩ	ㄩ	ㄩ	ㄩ
Expected	424	232	65	79	193	2308	1210	1964	2916	441	780	1139	312	186	524
Actual	516	370	102	55	147	1667	773	3479	2549	324	112	1255	383	267	796
Percentage	122	159	157	70	76	72	64	177	87	73	14	110	123	144	152

After ㄩ	ㄩ	ㄩ	ㄩ	ㄩ	ㄩ	ㄩ	ㄩ	ㄩ	。	ㄩ	ㄩ	ㄩ	ㄩ	ㄩ	ㄩ
Expected	14	8	2	3	6	77	41	66	98	15	26	38	10	6	18
Actual	4	1	1	2	2	60	36	98	93	26	39	19	12	5	29
Percentage	29	13	50	67	33	78	88	148	95	173	150	50	120	83	161

After ㄩ	ㄩ	ㄩ	ㄩ	ㄩ	ㄩ	ㄩ	ㄩ	ㄩ	。	ㄩ	ㄩ	ㄩ	ㄩ	ㄩ	ㄩ
Expected	35	19	5	7	16	192	101	163	243	37	65	95	26	15	44
Actual	8	6	3	11	11	169	91	41	293	62	300	28	15	3	17
Percentage	23	32	60	157	69	88	90	25	121	168	462	29	58	20	39

After ㄩ	ㄩ	ㄩ	ㄩ	ㄩ	ㄩ	ㄩ	ㄩ	ㄩ	。	ㄩ	ㄩ	ㄩ	ㄩ	ㄩ	ㄩ
Expected	170	93	26	32	77	924	484	786	1167	176	312	456	125	75	210
Actual	321	102	20	21	68	1138	651	411	999	128	178	673	140	65	189
Percentage	189	110	77	66	88	123	135	52	86	73	57	148	112	87	90

After ㄩ	ㄩ	ㄩ	ㄩ	ㄩ	ㄩ	ㄩ	ㄩ	ㄩ	。	ㄩ	ㄩ	ㄩ	ㄩ	ㄩ	ㄩ
Expected	168	92	26	31	76	913	479	777	1153	174	308	450	123	74	207
Actual	51	27	12	35	80	1056	633	250	1365	221	940	234	73	18	51
Percentage	30	29	46	113	105	116	132	32	118	127	305	52	59	24	25

After ㄩ	ㄩ	ㄩ	ㄩ	ㄩ	ㄩ	ㄩ	ㄩ	ㄩ	。	ㄩ	ㄩ	ㄩ	ㄩ	ㄩ	ㄩ
Expected	10	5	1	2	4	52	27	44	66	10	18	26	7	4	12
Actual	4	1	0	0	7	72	33	25	82	11	13	26	15	0	2
Percentage	40	20	0	0	175	138	122	57	124	110	72	100	214	0	17

After ㄩ	ㄩ	ㄩ	ㄩ	ㄩ	ㄩ	ㄩ	ㄩ	ㄩ	。	ㄩ	ㄩ	ㄩ	ㄩ	ㄩ	ㄩ
Expected	172	94	26	32	78	935	490	796	1181	179	316	461	126	75	212
Actual	30	27	9	46	122	1308	667	387	1576	257	255	372	79	11	41
Percentage	17	29	35	144	156	140	136	49	133	144	81	81	63	15	19

Analysis

Many word break combinations show large differences between the expected and actual occurrences. For example, the combination **g.a** is expected to occur 780 times, but actually occurs 112 times. Similarly, the combination **∩.cz** is expected to occur 935 times, but actually occurs 1,308 times. The total expected and actual number of occurrences of any glyph will be equal across all word break combinations, so the deviation in one combination will be matched by opposite deviations in other combinations for the same glyph.

Following are histograms for two glyphs, **†** and **∩**, which show large discrepancies between the expected and actual occurrences for many of their combinations. The graphs only show the combinations with the highest counts, which make up the majority of occurrences. Less common combinations are not shown.

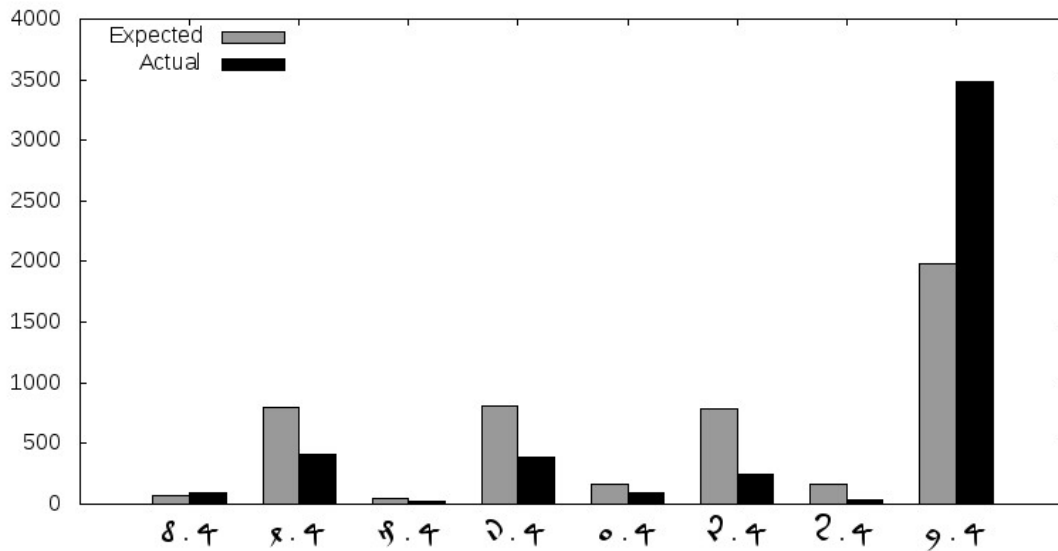


Chart 2: Actual and expected counts of word break combination with **†**.

The chart for word break combinations of **X.†** shows a clear and simple bias. The actual occurrence of the combination **g.†** is 177% of its expected occurrence. As **†** is almost always word-initial, this combination accounts for nearly two thirds of the glyph's occurrences. In compensation the combinations **x.†**, **2.†**, and **∩.†** are around 50% or less as common than expected.

Chart 3 (below) for word break combinations of **∩.X** shows a more complex picture. Several combinations, **∩.o**, **∩.g**, **∩.cz**, and **∩.Zc** show higher occurrences than expected. Some combinations show numerically small but proportionately large negative deviations. The combinations **∩.ll** and **∩.x** occur less than one third of the expectation. Overall, the lower than expected combinations compensate for the higher than expected combinations for **∩**.

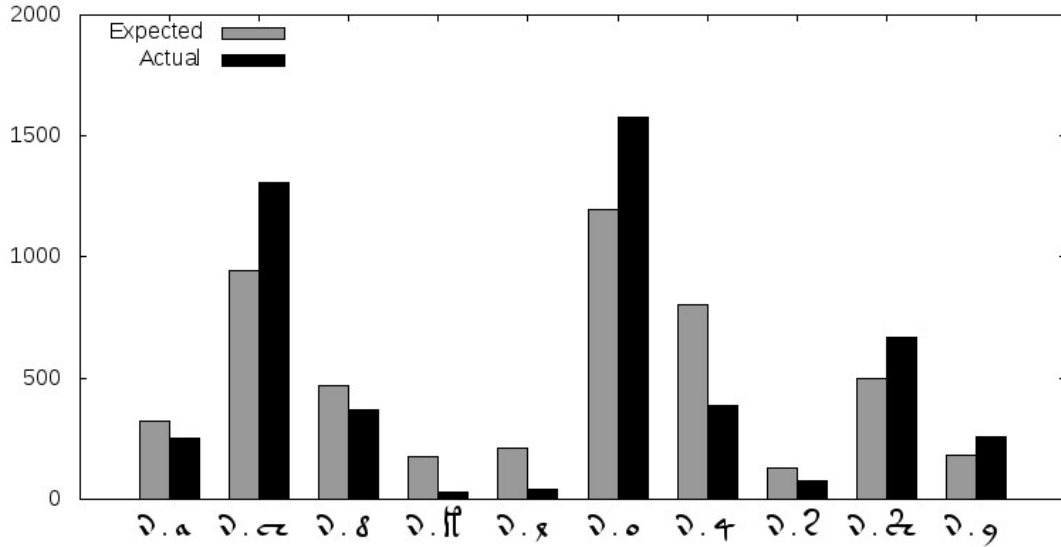


Chart 3: Actual and expected counts of some word break combination with 𐌆.

The combination 𐌆.𐌾 also appears in this chart with a numerically large difference between the expected number of occurrences and the actual. Here it evens out the difference of four other combinations with 𐌆, while in Chart 2 it was one of three combinations with 𐌾 which evened out 𐌾.𐌾. Thus 𐌆 and 𐌾 behave in opposite ways when in combination with 𐌾.

Glyph Groups

All glyphs which occur at the end of words can be divided between those which are more common than expected before 𐌾 and those which are less common. Only 𐌾 and 𐌺 belong in the former group, the other glyphs in the latter.

Likewise, we can divide the glyph set into two groups according to how they behave in combination with 𐌆, 𐌾, or any glyph. We find that across multiple such divisions glyphs are regularly in the same or opposite groups as other glyphs. For example, 𐌸 and 𐌺 always deviate the same way in combination with the same glyph. Other pairs may match more often than not, or always deviate in opposite ways, when combined with the same glyph.

Over the whole glyph set these similarities interact in a series of interlocking relationships. As with 𐌾, 𐌾 and 𐌆 deviate in opposite directions before 𐌺, as well as before 𐌸 and 𐌺. However, the direction of deviation before 𐌺 and before 𐌸 and 𐌺 is different: 𐌾 is positive before the former and negative before the latter; 𐌆 the opposite. This suggests that not only are 𐌆 and 𐌾 in different groups according to how they work, but that there is also a difference between 𐌺 on one hand and 𐌸 and 𐌺 on the other.

The presence of groups of glyphs which can be distinguished according to how they act in word break combinations opens up new possibilities for understanding the structure of the glyph set. Using the word break combination data we can construct a series of key divisions for the whole glyph set.

The following dendrograms¹³ (Charts 4.1-2) show the relationships between glyphs according to the correlation of the ratio between actual and expected occurrences in word break combinations. The first chart shows the relationship between glyphs at the beginning of words, and the second chart shows the relationships between glyphs at the end of words.

Note that the Y-axis shows correlation distance. Nodes connecting glyphs which appear low on the Y-axis represent closer correlation than nodes which appear high on the Y-axis. The ordering of glyphs on the X-axis is irrelevant, and adjacent glyphs are not necessarily related to the same degree as other adjacent glyphs.

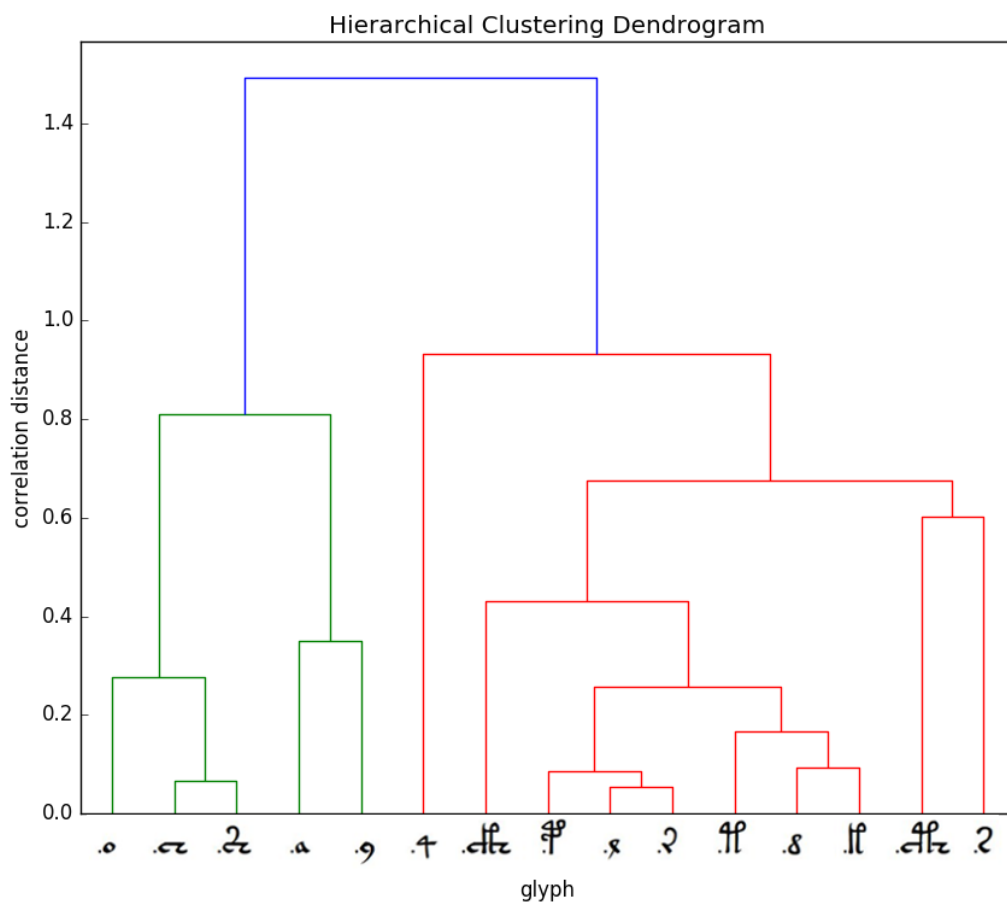


Chart 4.1: The relationships between glyphs at the beginning of words.

¹³ These graphs were produced with the Scipy open-source software (<https://www.scipy.org/>). A similar technique is described in Knight, Kevin, Beáta Megyesi and Christiane Schaefer (2011) *The Copiale Cipher* in *Proceedings of the 4th Workshop on Building and Using Comparable Corpora, 49th Annual Meeting of the Association for Computational Linguistics*, Portland, Oregon.

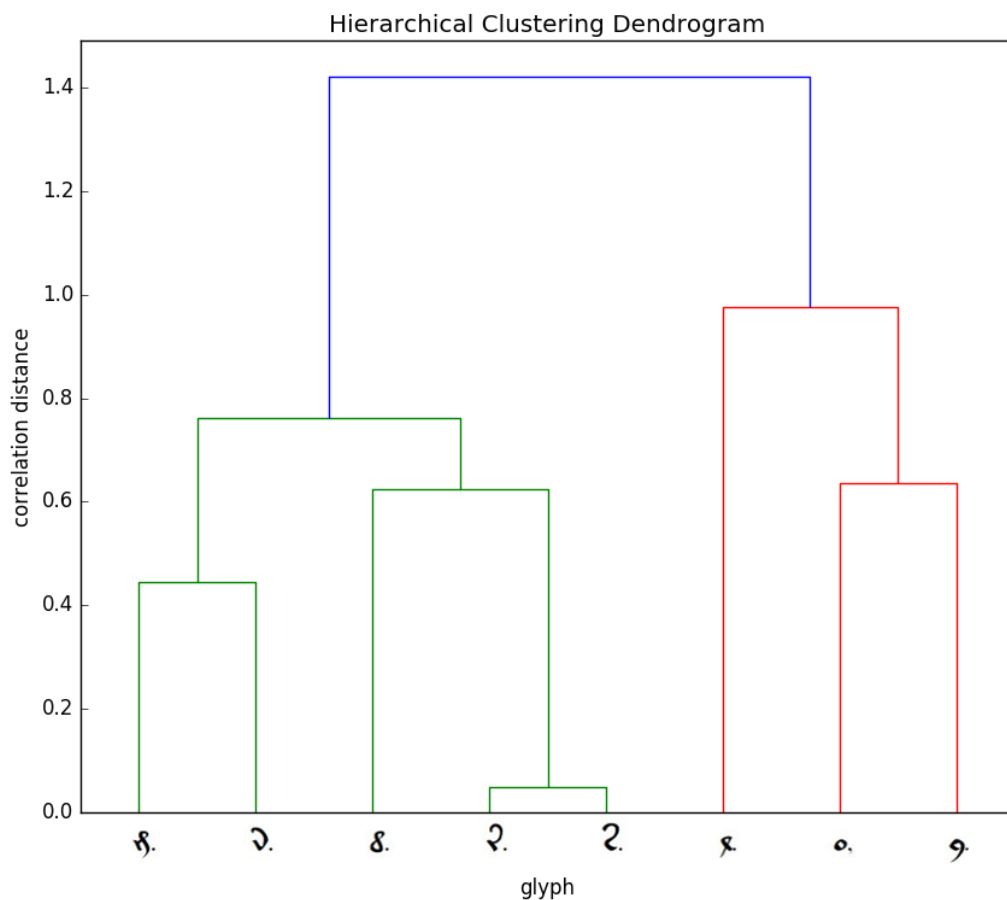


Chart 4.2: The relationships between glyphs at the end of words.

The glyphs fall into clear divisions based on these dendrograms. For glyphs at the beginning of words those on either side of the root node (marked out with green and red lines) form two groups: ɔ, ɔ, ɔ, ɔ, ɔ and ʃ, ʃ, ʃ, ʃ, ʃ, ʃ, ʃ, ʃ, ʃ. For glyphs at the end of words the two groups are: ɔ, ɔ, ɔ and ʒ, ʒ, ʒ, ʒ, ʒ.

Between the beginning and end of words the two groups are nearly exclusive. The same glyphs appear in groups with one another if they appear on both charts. The exception is ʒ. The table below shows all four groups together.

	Group 1	Group 2
Beginning	ɔ, ɔ, ɔ, ɔ, ɔ	ʃ, ʃ, ʃ, ʃ, ʃ, ʃ, ʃ, ʃ, ʃ
End	ɔ, ɔ, ʒ	ʒ, ʒ, ʒ, ʒ, ʒ

Table 4: Glyph groups identified from Charts 4.1-2.

The existence of these two groups has been informed by the similar biases of the word break combinations in which they participate. However, those biases may be positive or negative, with higher or lower levels of actual to expected occurrences of such combinations.

We can use the fact that some of the glyphs in each group occur at both the beginning and end of words to compare the groups internally. In tables 5.1–4 below each row corresponds to a word-ending glyph and each column to a word-initial glyph. Each cell represents the actual-expected ratio for the end-start combination corresponding to the row and column. For instance, the top-left cell of the first table states that the $\mathfrak{z}.\mathfrak{ll}$ combination only occurs 23% of the expected times.

	$\mathfrak{x}.\mathfrak{ll}$	$\mathfrak{x}.\mathfrak{ff}$	$\mathfrak{x}.\mathfrak{ff}$	$\mathfrak{x}.\mathfrak{ffz}$	$\mathfrak{x}.\mathfrak{ffz}$	$\mathfrak{x}.\mathfrak{s}$	$\mathfrak{x}.\mathfrak{z}$	$\mathfrak{x}.\mathfrak{z}$	$\mathfrak{x}.\mathfrak{z}$	$\mathfrak{x}.\mathfrak{f}$
$\mathfrak{z}.\mathfrak{x}$	23%	32%	60%	157%	69%	29%	58%	39%	20%	25%
$\mathfrak{s}.\mathfrak{x}$	29%	13%	50%	67%	33%	50%	120%	161%	83%	148%
$\mathfrak{z}.\mathfrak{x}$	30%	29%	46%	113%	105%	52%	59%	25%	24%	32%
$\mathfrak{s}.\mathfrak{x}$	40%	20%	0%	0%	175%	100%	214%	17%	0%	17%
$\mathfrak{z}.\mathfrak{x}$	17%	29%	35%	144%	156%	81%	63%	19%	15%	49%

Table 5.1: The actual-expected ratios for word break combinations from Group 2

	$\mathfrak{x}.\mathfrak{o}$	$\mathfrak{x}.\mathfrak{g}$	$\mathfrak{x}.\mathfrak{a}$	$\mathfrak{x}.\mathfrak{z}$	$\mathfrak{x}.\mathfrak{z}$
$\mathfrak{o}.\mathfrak{x}$	45%	89%	75%	60%	47%
$\mathfrak{g}.\mathfrak{x}$	87%	73%	14%	72%	64%
$\mathfrak{s}.\mathfrak{x}$	86%	73%	57%	123%	135%

Table 5.2: The actual-expected ratios for word break combinations from Group 1

We can also compare across the two groups. The next two tables show the same actual-expected ratios as above but for word break combinations with pairs from different groups.

	$\mathfrak{x}.\mathfrak{ll}$	$\mathfrak{x}.\mathfrak{ff}$	$\mathfrak{x}.\mathfrak{ff}$	$\mathfrak{x}.\mathfrak{ffz}$	$\mathfrak{x}.\mathfrak{ffz}$	$\mathfrak{x}.\mathfrak{s}$	$\mathfrak{x}.\mathfrak{z}$	$\mathfrak{x}.\mathfrak{z}$	$\mathfrak{x}.\mathfrak{z}$	$\mathfrak{x}.\mathfrak{f}$
$\mathfrak{o}.\mathfrak{x}$	246%	158%	260%	366%	200%	165%	156%	335%	513%	56%
$\mathfrak{g}.\mathfrak{x}$	122%	159%	157%	70%	76%	110%	123%	152%	144%	177%
$\mathfrak{s}.\mathfrak{x}$	189%	110%	77%	66%	88%	148%	112%	90%	87%	52%

Table 5.3: The actual-expected ratios for word break combinations with the last glyph of the preceding word from Group 1 and the first glyph of the following word from Group 2.

	X.๐	X.๑	X.๒	X.๓	X.๔
๒.X	121%	168%	462%	88%	90%
๘.X	95%	173%	150%	78%	88%
๒.X	118%	127%	305%	116%	132%
๘.X	124%	110%	72%	138%	122%
๑.X	133%	144%	81%	140%	136%

Table 5.4: The actual-expected ratios for word break combinations with the last glyph of the preceding word from Group 2 and the first glyph of the following word from Group 1.

A broad pattern is noticeable in the statistics of Tables 5.1-4. Combinations with glyphs from the same group tend to have lower ratios and combinations with glyphs from different groups tend to have higher than expected ratios. The pattern is not absolute and there are a number of exceptions. Many of the exceptions can be accounted for with either **๘**, which switches groups depending on whether it is at the beginning or end of a word, and the glyphs **๓๓**, **๓๔**.

The patterns of glyph group preferences appear to be stronger for those glyphs in Group 2 than those in Group 1. We have therefore chosen to name Group 2 ‘Strong’ because of the strong preferences, and Group 1 ‘Weak’ because of the relatively weaker preferences.

Nature of Glyphs

The Weak and Strong glyph groups show some interesting behaviour outside of word break combinations. Strong glyphs are uncommonly found adjacent within words, the main exception being **๘**, which can be followed by another Strong glyph. However, this exception fits with that glyph being variously Strong and Weak, depending on position. Weak glyphs, on the other hand, will often occur adjacently, especially when in a particular order (typically **๓๓**, **๓๔** before **๐**, **๑**, **๒**).

The Weak glyph group is also notable for its essential role in words. Almost all words contain at least one Weak glyph, and many contain multiple. A number of words contain only Weak glyphs, whereas almost none contain only Strong glyphs.

What we can say about the nature of the glyph groups is limited but suggestive. We have noted that word break combinations where the two glyphs are from the same group tend to occur less commonly than expected. This suggests that the ‘similarity’ of the glyphs is a negative factor on the occurrence of word break combinations. But as the groups are defined by the actual-expected ratios for word break combinations, this is circular reasoning.

However, word break combinations with the same glyph in both positions are also less common than expected. The table below shows all such combinations which commonly appear and their ratios with respect to the expected number of occurrences.

o.o	9.9	8.8	2.2	8.8	2.2
45%	73%	90%	24%	50%	58%

Table 6: the ratios of word break combinations with the same glyph in both positions.

On the assumption that glyphs have the same or similar value at the beginning and end of words (which is reasonable though not assured), the above figures show that glyph pairs with similar characteristics are selected against in word break combinations. We can thus tentatively suggest that the Weak and Strong glyph groups, although defined by word break combinations, represent some characteristic similarity for the glyphs in each group.

Other researchers have attempted to separate the Voynich glyph set into sub-groups based on different characteristics than the ones used in this paper. Guy¹⁴ applied Sukhotin’s vowel-identifying algorithm to two pages of the Voynich text. The algorithm identified the characters o, 9, a, c, cc, v, z as being potential vowels. Of this set only o, 9, a have been analysed in word break combinations but all belong to the Weak glyph group. However c, 2, which are also Weak glyphs, were identified as potential consonants.

A similar outcome was found by Zandbergen¹⁵, who applied a two-state Hidden Markov Model to four different transcriptions of the Voynich text. In all cases o, 9, a, c were identified as vowels, with c being identified for at least one transcription¹⁶. Other glyphs such as w2, d, p were identified as potential vowels by one transcription each.

Although the split suggested in these analyses do not coincide fully with the Strong/Weak groups they are largely coincident. All three place o, 9, a in a the same group, with ll, ll, p, ff, ff, 2, 8, 2, 4 in another group. The main difference is their position for c, 2. Given that the other analyses approached the text from a different angle than word break combinations, they provide some support that a fundamental split in the glyph set can be identified.

14 Guy, J (1991) ‘Statistical properties of two folios of the Voynich Manuscript’, *Cryptologia*, 15:3, 207-218

15 Zandbergen, R (2018) ‘What we may learn from the MS text entropy’, *The Voynich Manuscript*. [Online: www.voynich.nu/extra/sol_ent.html. Retrieved 8 May 2018.]

16 Due to specifics of the transcription the outcome for EVA is ambiguous. c, 2 were considered as two glyphs, with both c, 2 being identified as vowels, though not z. It is not safe to assume that were they considered as one glyph the outcome would have identified them as vowels. The opposite is possible.

Further, the analyses above were intended to discover vowels in a plainly written text. Although it is outside the scope of this paper to question whether the analyses succeeded in their goal, they do provide a potential identification of the Strong/Weak groups. If these analyses are correct, then the Weak group would more closely align with those glyphs identified as vowels. The Strong group, by default, would be associated with consonants. These conclusions, however, are dependent on several assumptions we cannot answer here.

Dependence

Most glyphs have multiple combinations which deviate from the expected level. The text as a whole thus contains many thousands of individual word break combinations which must not occur by chance. Although whole-word phrases are not common in the text, it is not unstructured above the level of the word (which is known to be highly structured).

One glyph in any word break combination is predictive of the other, as can be seen when compared with texts from natural languages. The following table presents the percentage of last-first X.Y combinations deviating from the number expected under the assumption that X and Y were independent. The comparison texts are excerpts of about 20,000 words each; uppercase characters have been converted to lowercase and punctuation marks have been considered as spaces.

Source	Number of Couples	Number of Couple Types	% deviation
Dante ¹⁷ (Italian)	17,427	230	14.5
Mattioli ¹⁸ (Latin)	18,666	309	8.1
Shakespeare ¹⁹ (English)	17,642	450	10.7
Entire Voynich ms	31,767	285	20.0
Voynich Currier A	9,021	224	15.6
Voynich Currier B	20,326	213	23.0

Table 7: the Root Mean Square Deviation for the Voynich text and two comparators.

In the sample from Dante's *Divina Commedia*, the single word break combination that contributes most to the deviation is 'e.l'. Most of the combinations that are more frequent than expected include a vowel and a consonant (in either order). The elision of the preceding vowel before a word starting with a vowel is a common phenomenon in Italian (e.g. "l'erta" / "le lusinghe" - "qual è" / "quali sono") resulting in a

¹⁷ Dante Alighieri, *Divina Commedia*.

¹⁸ Pietro Andrea Mattioli, *Commentarii in libros sex Pedacii Dioscoridis de medica materia*, 1554.

¹⁹ William Shakespeare, *The Sonnets* and several theatrical plays.

noticeable correlation between the last character of a word and the first character of the following one. In the English and Latin samples, deviations from the expected values appear to be caused by frequently repeated sentences.

In the Voynich text, word break combinations are more predictable than for Latin and English, and also more predictable in Currier A than for Italian.

Alongside the question of predictability is that of direction: whether one side of the word break combination is causing the other. The differences in word break combinations between Currier A and B offer a partial answer.

The combination $\mathfrak{D}\circ$ appears in Currier A at nearly the expected level, but there is a clear preference for the combination in Currier B. The frequency of \circ at the start of a word is much higher in Currier B and this increase is concentrated in combinations with \mathfrak{D} . This suggests that \mathfrak{D} is causing the presence of \circ .

A similar situation is found with the combinations $\mathfrak{D}\mathfrak{a}$, where the \mathfrak{D} appears to cause the presence of \mathfrak{a} . Yet a contrary situation is found with other combinations and the overall picture may be too complex for a definitive answer. We can thus only tentatively say that the second glyph in a combination depends on the first.

Discussion

The meaning of word break combinations can only be understood through knowledge of the cause. There are several possible explanations which may be considered plausible. Below we set out some explanations and provide arguments for and against them.

1. Word breaks are not real and the patterns found in word break combinations are a continuation from the patterns found inside words.

Although there is an obvious structure within words the bigrams which are preferred by the structure are not the same as those preferred in word break combinations. Combinations such as $\mathfrak{g}\mathfrak{f}$ and $\mathfrak{D}\circ$ are much more common than expected at word breaks but are practically absent within words, with 7 tokens for $\mathfrak{g}\mathfrak{f}$ and 18 for $\mathfrak{D}\circ$.

This could simply point to the existence of a process for inserting word spaces into the text at predetermined points. However, single word labels in diagrams and next to illustrations follow the same general structure as the main text. They often begin with \circ and end with \mathfrak{g} and other parts of word structure, such as words beginning \mathfrak{of} , \mathfrak{ff} or ending \mathfrak{dg} are shared between labels and the main text. As labels are strings of glyphs which have no immediate relationship with other text, they must be regarded as complete in themselves and not divisions of longer text.

The sharing of word structure between labels and the main text suggests that word breaks are a real phenomenon and that word structure operates between such breaks but not across them.

2. Word break combinations are the result of procedural generation.

Several researchers, such as Rugg²⁰ and Timm²¹, have proposed that the text of the Voynich manuscript was procedurally generated. The researchers claim that they are able to create a text with the same statistical features as the Voynich text using a basic set of rules. In these theories the text is potentially meaningless though its creation is definitely structured.

The word break combinations in a piece of text generated by Timm²² show that the actual and expected occurrences are very similar. They also differ significantly from the actual word break combinations found in the Voynich text, as shown below. Only the frequencies of **ꞥꞥ** and **ꝥꝥ** are close matches to text itself.

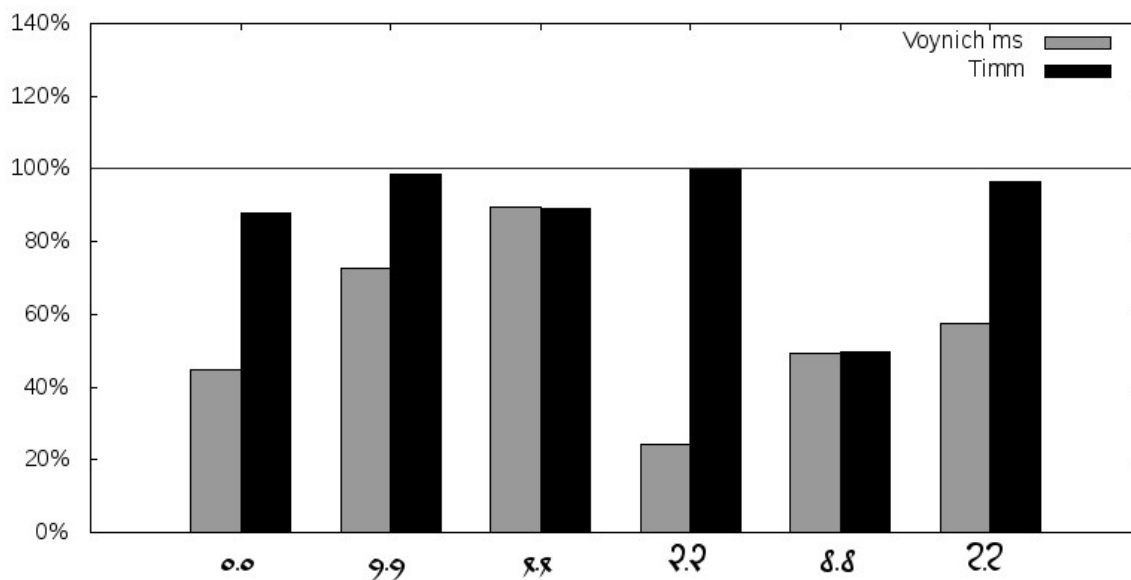


Chart 5: The ratio between actual and expected counts of word break combinations for identical glyph pairs. Values corresponding to the Voynich manuscript are compared with a piece of text generated by Timm.

A sufficiently complex procedural approach could reproduce the distribution of word break combinations observed in the manuscript. The fact that current procedural models fail to do so is likely due the limited knowledge of the phenomenon. The same is true for any feature of the text. Current theories of procedural generation may thus be disproven because of their gaps in modelling word break

20 Rugg G (2004). 'An Elegant Hoax? A Possible Solution to the Voynich Manuscript'. *Cryptologia*, vol. XXVIII(1), 31-46.

21 Timm, Torsten (2014). *How the Voynich Manuscript was created*.

[Online: <https://arxiv.org/abs/1407.6639>. Retrieved 26 March 2018.]

22 <https://www.voynich.ninja/thread-1385-post-12230.html#pid12230>

combinations. Any future textual generation process which does model word break combinations will only be more sophisticated and not necessarily more correct.

3. Glyphs within words are arranged to create certain word break combinations.

This explanation supposes that a glyph may be moved to or from the start and end of words. So the combination $\mathfrak{D}\cdot\mathfrak{O}$ could involve a word such as $\mathfrak{A}\mathfrak{L}\mathfrak{C}\mathfrak{O}\mathfrak{S}\mathfrak{G}$ being rearranged to become $\mathfrak{O}\mathfrak{A}\mathfrak{L}\mathfrak{C}\mathfrak{S}\mathfrak{G}$ when occurring after a word ending \mathfrak{D} .

The rigidity of the known word structure is suggestive that glyphs are not mobile within words and there is not a wide diversity of possible structures. As in explanation 1 above, the presence of the same structures in labels which occur alone shows that this structure is normal even for words with no neighbours.

4. The arrangement of whole words creates certain word break combinations.

This explanation could work in two possible ways to create word break combinations: certain word phrases are preferred and word break combinations result from these preferences, or words are arranged with the goal of creating word break combinations.

In the first, phrases such as $\mathfrak{S}\mathfrak{A}\mathfrak{U}\mathfrak{D}\ \mathfrak{O}\mathfrak{R}$ or $\mathfrak{S}\mathfrak{A}\mathfrak{U}\mathfrak{D}\ \mathfrak{O}\mathfrak{X}$ might be preferred throughout the text for a reason such as meaning or grammar, unrelated to the glyphs themselves. This would cause the combination $\mathfrak{D}\cdot\mathfrak{O}$ to occur more often than expected. In the second, any two words, one ending \mathfrak{D} and one beginning \mathfrak{O} , would be brought together in order to create the combination $\mathfrak{D}\cdot\mathfrak{O}$.

We can definitely rule out the first as being the sole contributor to the cause of word break combinations. If we measure only those combinations in which one word is a hapax legomenon²³ (thus excluding completely the possibility of a repeated phrase) the combinations are similar as for the whole text with repeated phrases included.

The chart below shows common word break combinations against their expected value. Dark bars represent the combinations for the whole text, light bars for only those combinations including a hapax legomenon. An observed count identical to the expected results in a 100% ratio. Many combinations are similar or near-identical between the two measures. Although some combinations with only hapax legomena deviate much less from the expected value, this partial influence of repeated phrases cannot explain the pervasive presence of word break combinations.

²³ A *hapax legomenon* is a word occurring only once in the text.

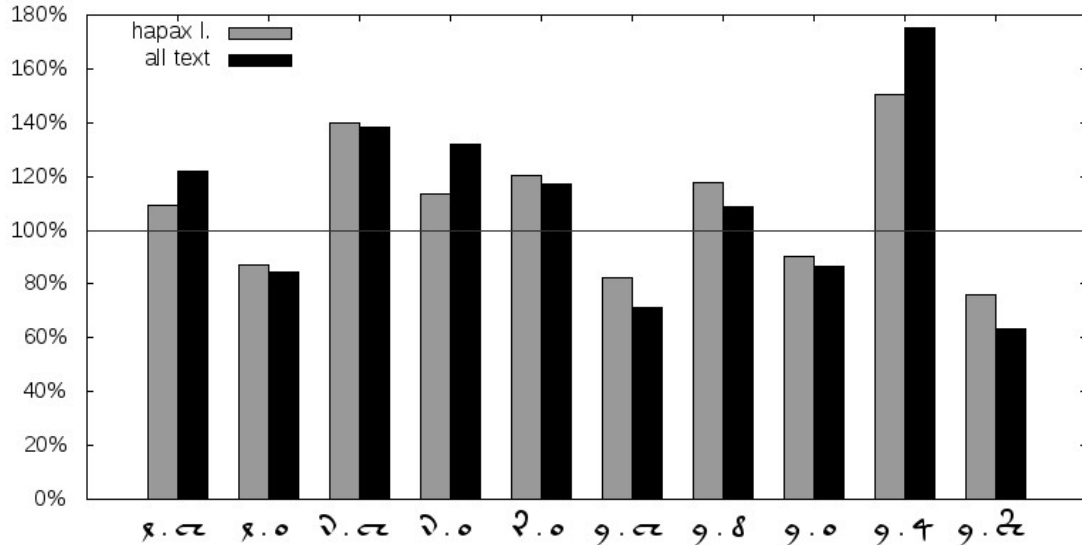


Chart 6: Comparison between all word break combinations and only those including a hapax legomenon.

For the second possibility, that whole words are arranged to create word break combinations, the evidence for the direction of causation outlined above makes this unlikely.

We saw that the combination ɀ.ο appears in Currier A at nearly the level expected and there was no preference for the combination. The frequency of ο at the start of a word is much higher in Currier B and this increase is concentrated in combinations with ɀ. The increase in words beginning ο would not alone cause the unexpected levels of the combination ɀ.ο as the expected levels of ɀ.ο in Currier B would also be higher. There would need to be a separate cause for the increase in the number of words beginning ο and an additional process to bring them into the combination ɀ.ο.

This split between the causes of the number of words beginning ο and their placement is unsatisfactory. It proposes two different processes to explain a single observation and would potentially need to be replicated across multiple word break combinations. The unnecessary complexity of the solution is a mark against it.

5. Words are altered to create certain word break combinations.

This explanation proposes that, whatever process orders the words within the text, alterations are subsequently made to words to increase or decrease the occurrence of certain word break combinations. Glyphs might be added, removed, or altered in order to satisfy unknown preferences. There are many potential reasons why some combinations are preferred.

Glyphs at the start and end of words could be nulls intended to obscure the underlying text. They might be added to a meaningful text but without having meaning themselves. Were the glyphs not to have intrinsic meaning there would need to be a regular procedure for adding them in order to create the

biases in the word break combinations that have been observed.

However, it would be difficult to explain the reasons for such glyphs if they are regularly applied to the text. Any regular process would be reversible and defeat an attempt to use such glyphs for obfuscation.

Alternatively, glyphs might be added or altered for a reason intrinsic to the glyphs. Certain combinations could have been visually appealing to the author or easier to write yet by being regularly applied made no difference to readability. It is difficult to prove or disprove such a hypothesis at our current level of knowledge.

Another option is that alterations were made for reasons intrinsic to the text. Either the content or the medium (whether language, code, or cipher) promoted certain combinations and not others. In some languages words can exhibit sound changes due to the interaction or influence of nearby sounds or morphemes.

Such phenomena appear in the written form of some languages. We mentioned above the Italian language in which a final vowel is commonly elided before a word beginning with a vowel. A more extensive feature is initial consonant mutation in the Welsh language which is part of the orthography. Some words have up to four different forms, each differing in the initial consonant, which are conditioned by the preceding word. An example is the definite article *y* which causes the soft mutation: *merch* ('daughter') becomes *y ferch*, and *draig* ('dragon') becomes *y ddraig*.

The authors admit a preference for this last hypothesis. The presence of a phonological process causing word break combinations would overcome a number of problems. Firstly, the process would be regular enough to create a preference for certain combinations which is observed throughout the text. Second, by being based on the intrinsic knowledge of a language, word break combinations would place no extra burden on either the composition or interpretation of the text.

Third, it could explain why some combinations have a much higher preference than others, as the phonological influence of some sounds is more salient. Fourth, it would also explain, as proposed earlier, why the glyphs appear to fall into groups based on their behaviour. The glyph groups identified, and the relationships between different glyphs in general, could be reflective of phonological relationships, such as between vowel and consonants.

We would like to suggest that the existence of word break combinations provides some support for the possibility that the Voynich text is written phonetically in a language. We also believe that information gleaned from further study of word break combinations could help to identify the sound values of the glyphs in which the text is written.