Rightward and Downward Grapheme Distributions in the Voynich Manuscript

Patrick Feaster

Griffonage.com, Bloomington, IN, USA

Abstract

A systematic approach is outlined for detecting patterns of uneven positional distribution of words and glyphs within lines and paragraphs in terms of "rightwardness" (distance towards the right end of a line) and "downwardness" (distance towards the bottom of a paragraph). Three case studies are developed to support an argument that such patterns are both more numerous and more pervasive in the Voynich Manuscript than generally supposed and that the specific combination of glyphs that make up a Voynichese word is significantly constrained by that word's position rightward and downward.

Keywords

Digital Humanities, Quantitative Text Analysis, Grapheme Distribution, Voynich Manuscript

1. Introduction

It has long been recognized that certain glyphs in the text of the Voynich Manuscript have uneven positional distributions within lines and paragraphs. Well-known examples include the tendency of paragraphs to begin with a "gallows" glyph, most often [p]; the preference of [p], [f], [cPh], and [cFh] for the first lines of paragraphs; the preference of [m] and [g] for the ends of lines; and differences between the relative frequencies of glyphs found respectively at the beginnings of line-initial words and of mid-line words (note that all transcriptions in this paper employ the widely used Extensible Voynich Alphabet, or EVA). The detection of such patterns famously led Prescott Currier to conclude in 1976 "that the line is a functional entity…and that the occurrence of certain symbols is governed by the position of a 'word' in a line" [1]. Patterning at the line and paragraph level may so far have received less methodical attention than the patterning within individual words, but it presents a similarly daunting hurdle for hypotheses about the nature of the text to need to surmount. In a recent essay, René Zandbergen highlights one pattern in particular—the preference of [p], [f], [cPh], and [cFh] for the first lines of paragraphs—as "a prohibitive feature for a large number of proposed solutions to the Voynich MS" [2]. For convenience, I will refer here to patterns of uneven positional distribution within lines and paragraphs simply as "distribution patterns."

At first, the identification of distribution patterns was restricted to distinctive behaviors at the extremities of the text: the first and last glyphs of the line, the first word in the line, or the first line of the paragraph. This limitation left open the possibility that the patterns might represent narrowly isolatable contextual modifications made to the edges of text that is otherwise uniform in character. During the past decade, however, distribution patterns have begun to be detected deeper inside lines as well. Elmar Vogt reports that the second words of lines tend to be shorter than average [3], Emma May Smith and Marco Ponzi find that the second words of lines in Quire 20 begin disproportionately with [Sh] and [ch] [4], and J. K. Petersen observes that [an], [ain], and [aiin] tend to appear later in lines than [on], [oin], and [oiin] [5]. Occasional findings such as these have hinted that distribution patterns might

International Conference on the Voynich Manuscript 2022, November 30--December 1, 2022, University of Malta.
EMAIL: pfeaster@gmail.com
ORCID: 0000-0001-7448-2290

be more pervasive than formerly suspected, implicating much or even all of the text, and that they are accordingly all the more deserving of methodical study.

In the present paper, I will discuss some recent efforts I've made to systematize the search for distribution patterns among both words and glyphs. My approach centers on the use of two convenient quantitative metrics, rightwardness and downwardness, which can be reported as numerical scores or plotted on graphs of various kinds. Rather than attempting to identify and list as many distribution patterns as possible, I will focus here on establishing the existence of three presumptively representative ones as persuasively as I can. My choice of case studies has been informed by several rhetorical considerations. On the one hand, I've selected distribution patterns that, to the best of my knowledge, have not previously been reported by others, and that are also strong and consistent enough that there can be little doubt as to their statistical significance. On the other hand, I've also prioritized cases that furnish evidence that distribution patterns permeate the whole text and are not confined to its edges. These "new" distribution patterns—just like the better-known but less pervasive ones—seem to affect the component glyphs of words rather than words as unitary wholes, suggesting that the internal composition of a Voynichese word is significantly constrained by its position rightward and downward.

2. Mean Relative Rightwardness Scores

The term "rightwardness" will refer in this paper to how far towards the right end of a line a word or glyph appears. In the past, this parameter has typically been discussed in terms of *absolute* positions, such as the second word in a line or the second-to-last word in a line. Absolute positions might well be significant, but they're cumbersome to analyze: any word could be associated with either of two numbers, depending on whether positions are counted from the left or the right, and lines containing different numbers of words can be difficult to compare with each other in these terms, as Vogt discovered in his efforts to study word lengths [3]. Thus, I've chosen instead to analyze *relative* rightwardness, or how far towards the right a token appears *proportionally* within a line—a metric that permits easier comparison across disparate contexts. Relative rightwardness can be calculated by dividing a token's ordinal position within its line, counted from left to right starting at zero, by the line's total token count minus one. If we assume for purposes of argument that Voynichese text runs from left to right, a token at the beginning of a line will always score 0, a token at the end of a line will always score 1, and a token in the exact middle of a line will always score 0.5. To demonstrate the process of calculation more concretely, let's turn to f103v.2 as an example:

[daiin.Shey.qokal.Shedy.qokeedy.qoteor.Shey.qoty.chcKhy.qotain.chalr]

There are two tokens here of [Shey], the second and seventh out of eleven words, so their individual relative rightwardness scores are 1/10 and 6/10, and the mean of these two scores is 7/20 or 0.35. We can also calculate relative rightwardness at the glyph level, but this requires that we first make a working decision about what counts as an individual glyph. In my own experiments with glyph-level analysis, I've followed the lead of EVA except that I've counted all structures incorporating "benches" (e.g., [ch], [Sh], and [cKh]) and all quantities of [i] and [e] (e.g., [ii] and [ee]) as single glyphs. By this measure, f103v.2 contains three tokens of [Sh] in positions 5, 13, and 29 out of 48, so their individual rightwardness scores are 4/47, 12/47, and 28/47, and the mean of these scores is 44/141 or about 0.312.

One object I've found it especially productive to study in terms of relative rightwardness is the graphemic minimal word pair. By this I mean a pair of words that are formally identical except for a single grapheme, such as [chol] versus [chal], much as a "minimal pair" in phonology refers to two words that differ only by a single phoneme or other phonological element. Strictly speaking, of course, the identification of graphemes—a writing system's smallest meaningful contrastive units—should still be regarded as tentative at this point, since we don't yet know with any certainty which units are meaningful, if any. However, I consider it worthwhile to preserve the analogy between phonemes and graphemes, which is why I'm opting to invoke the latter term anyway. For convenience, I will refer to graphemic minimal word pairs below simply as "word pairs," and to words belonging to such pairs as "homologous" to each other, and I will further define a word pair "set" as consisting of all attested word pairs that differ in the same way, such as by beginning alternately with [Sh] and [ch].

If we compare mean rightwardness scores either for individual word pairs or for whole sets of them, we find that words that contrast in the same way sometimes display consistent positional deviations relative to one another, with one type routinely scoring rightward of the other. In some cases, the deviations only confirm the existence of distribution patterns we would already have known to expect. For example, words ending with [g] or [m] will predictably score rightward of homologous words ending in any other way because of the well-known and easily noticed preference of [g] and [m] for the ends of lines. In other cases, though, the deviations can reveal subtler distribution patterns that would not have been apparent from casual observation. The deviations themselves might be very slight, perhaps corresponding to an average difference in position of only one twentieth of a line. What is remarkable is not their magnitude, but their consistency. The statistics for the three patterns that follow were obtained by applying custom Python scripts to the Zandbergen-Landini ("ZL") transcription of the Voynich Manuscript, version 1.7, omitting commas that represent uncertain spaces, resolving any glyphs noted as ambiguous in favor of the first option provided other than "?", and limiting analysis to text in paragraphs as opposed to circles, labels, radii, and so forth.

First, words beginning with [ch] routinely score rightward of homologous words beginning with [Sh]. If we calculate mean relative rightwardness for both of the pertinent word pair sets, the [Sh] set scores 0.434 with 2456 tokens, while the [ch] set scores 0.514 with 4050 tokens, showing a difference of about 0.08. These figures include the first and last words of lines, and since words in those two positions have long been recognized as displaying anomalous behavior, it might be tempting to assume that the difference in score results from them alone. However, if we exclude all scores of 0 and 1—limiting our analysis to what I will call "mid-line" words—the [Sh] set scores 0.433 with 2167 tokens, while the [ch] set scores 0.481 with 3600 tokens, which still shows a difference of about 0.05. If we compare specific word pairs, we find that the pairs with the highest total token counts conform individually to the same pattern (see Table 1). There are 21 word pairs that alternate between initial [Sh] and [ch] with 60 or more total tokens, and in every one of these cases the [ch] word scores rightward on average from the [Sh] word. If we limit our analysis to mid-line words, 19 of these word pairs—roughly 90%—still conform to the same pattern.

Table 1

Mean rightwardness scores for the most frequent graphemic minimal word pairs beginning with [Sh] and [ch], with anomalous scores highlighted

	All		Mid-Line			All		Mid-Line	
	Tokens	**Score**	**Tokens**	**Score**		**Tokens**	**Score**	**Tokens**	**Score**
Shedy	359	0.483	337	0.467	**chedy**	416	0.551	380	0.525
Shey	214	0.451	198	0.432	**chey**	272	0.470	251	0.441
Shol	165	0.373	145	0.397	**chol**	316	0.416	298	0.427
Shor	87	0.320	66	0.391	**chor**	192	0.421	179	0.424
Sheey	119	0.427	112	0.427	**cheey**	133	0.449	128	0.443
Sheol	94	0.380	86	0.369	**cheol**	138	0.427	129	0.425
Shy	74	0.484	65	0.443	**chy**	116	0.500	108	0.481
ShcKhy	50	0.404	50	0.404	**chcKhy**	123	0.541	114	0.513
Shdy	40	0.470	35	0.480	**chdy**	124	0.582	108	0.520
Sho	98	0.321	70	0.450	**cho**	40	0.437	37	0.446
Shody	52	0.464	37	0.382	**chody**	81	0.591	71	0.548
Sheedy	67	0.412	64	0.432	**cheedy**	53	0.580	49	0.546
Sheody	43	0.466	40	0.451	**cheody**	75	0.572	62	0.498
Sheor	42	0.361	37	0.409	**cheor**	72	0.452	66	0.447
ShcThy	30	0.516	29	0.499	**chcThy**	70	0.598	63	0.553
Shar	23	0.290	19	0.299	**char**	60	0.473	58	0.489
Sheky	30	0.504	29	0.486	**cheky**	53	0.519	48	0.469
ShecKhy	29	0.413	28	0.392	**checKhy**	46	0.522	45	0.511
Shodaiin	24	0.424	15	0.412	**chodaiin**	42	0.539	33	0.534
Shear	20	0.497	20	0.497	**chear**	41	0.523	39	0.498
Shaiin	19	0.290	17	0.325	**chaiin**	41	0.551	33	0.442

Second, words beginning with [o] routinely score rightward of homologous words beginning with [qo]. If we calculate mean relative rightwardness for both of the pertinent word pair sets, the [qo] set scores 0.468 with 4461 tokens, while the [o] set scores 0.534 with 5161 tokens, showing a difference of about 0.07. If we limit our analysis to mid-line words, the [qo] set scores 0.487 with 3735 tokens, while the [o] set scores 0.517 with 4256 tokens, which still shows a difference of about 0.03. If we compare specific word pairs, we again find that the pairs with the highest total token counts conform individually to the same pattern (see Table 2). There are 23 word pairs that alternate between initial [qo] and [o] with 100 or more total tokens, and in every one of these cases the [o] word scores rightward on average from the [qo] word. If we limit our analysis to mid-line words, 19 of these word pairs—roughly 83%—still conform to the same pattern.

Table 2

Mean rightwardness scores for the most frequent graphemic minimal word pairs beginning with [qo] and [o], with anomalous scores highlighted

	All		Mid-Line			All		Mid-Line	
	Tokens	Score	Tokens	Score		Tokens	Score	Tokens	Score
qokaiin	259	0.457	218	0.488	**okaiin**	199	0.481	171	0.484
qol	104	0.430	84	0.449	**ol**	353	0.505	298	0.481
qokeey	287	0.442	247	0.473	**okeey**	146	0.457	129	0.478
qokeedy	291	0.440	256	0.484	**okeedy**	99	0.555	85	0.541
qokain	255	0.491	222	0.519	**okain**	126	0.520	108	0.514
qokedy	269	0.473	240	0.488	**okedy**	105	0.527	100	0.503
qokal	181	0.499	156	0.508	**okal**	116	0.557	100	0.516
qor	16	0.244	9	0.323	**or**	248	0.452	213	0.461
qokar	140	0.467	130	0.480	**okar**	107	0.527	99	0.539
qotedy	84	0.537	78	0.527	**otedy**	131	0.549	121	0.544
qotaiin	76	0.518	65	0.513	**otaiin**	136	0.599	120	0.579
qoky	128	0.606	97	0.531	**oky**	83	0.661	60	0.565
qoty	73	0.596	55	0.536	**oty**	93	0.697	65	0.612
qotal	55	0.582	50	0.560	**otal**	104	0.634	92	0.597
qotar	57	0.595	49	0.569	**otar**	99	0.611	94	0.590
qoteedy	70	0.489	57	0.530	**oteedy**	85	0.543	78	0.541
qokey	95	0.435	80	0.454	**okey**	52	0.471	46	0.445
qokol	88	0.432	78	0.487	**okol**	56	0.537	48	0.522
qotain	53	0.549	46	0.546	**otain**	89	0.622	80	0.592
qoteey	36	0.364	32	0.409	**oteey**	100	0.510	89	0.540
qokeol	46	0.505	42	0.529	**okeol**	57	0.511	51	0.473
qotol	39	0.403	31	0.507	**otol**	61	0.515	51	0.557
qotchy	60	0.352	39	0.464	**otchy**	40	0.551	34	0.618

Third, words containing [t] routinely score rightward of homologous words containing [k]**.** If we calculate mean relative rightwardness for both of the pertinent word pair sets (including cases of [cTh] and [cKh]), the [k] set scores 0.488 with 7177 tokens, while the [t] set scores 0.523 with 4423 tokens, showing a difference of about 0.035. If we limit our analysis to mid-line words, the [k] set scores 0.490 with 6048 tokens, while the [t] set scores 0.541 with 3467 tokens, which shows a difference of about 0.05. In this case, then, the difference in mean scores actually *increases* when we consider only mid-line tokens. If we compare specific word pairs, we find that the pairs with the highest total token counts follow this same pattern (see Table 3). There are 22 word pairs that alternate between [k] and [t] with 100 or more total tokens. Overall, the [t] word scores rightward on average from the [k] word in just 16 of these cases (about 73%), but that figure increases to 20 cases (about 91%) if we limit our analysis to mid-line tokens.

Table 3
Mean rightwardness scores for the most frequent graphemic minimal word pairs containing [k] and [t], with anomalous scores highlighted

	All		Mid-Line			All		Mid-Line	
	Tokens	**Score**	**Tokens**	**Score**		**Tokens**	**Score**	**Tokens**	**Score**
qokeedy	291	0.440	256	0.484	**qoteedy**	70	0.489	57	0.530
qokedy	269	0.473	240	0.488	**qotedy**	84	0.537	78	0.527
okaiin	199	0.481	171	0.484	**otaiin**	136	0.599	120	0.579
qokaiin	259	0.457	218	0.488	**qotaiin**	76	0.518	65	0.513
qokeey	287	0.442	247	0.473	**qoteey**	36	0.364	32	0.409
qokain	255	0.491	222	0.519	**qotain**	53	0.549	46	0.546
okeey	146	0.457	129	0.478	**oteey**	100	0.510	89	0.540
okedy	105	0.527	100	0.503	**otedy**	131	0.549	121	0.544
qokal	181	0.499	156	0.508	**qotal**	55	0.582	50	0.560
okal	116	0.557	100	0.516	**otal**	104	0.634	92	0.597
okain	126	0.520	108	0.514	**otain**	89	0.622	80	0.592
okar	107	0.527	99	0.539	**otar**	99	0.611	94	0.590
qoky	128	0.606	97	0.531	**qoty**	73	0.596	55	0.536
qokar	140	0.467	130	0.480	**qotar**	57	0.595	49	0.569
chcKhy	123	0.541	114	0.513	**chcThy**	70	0.598	63	0.553
okeedy	99	0.555	85	0.5405	**oteedy**	85	0.543	78	0.5409
oky	83	0.661	60	0.565	**oty**	93	0.697	65	0.612
cKhy	31	0.586	23	0.442	**cThy**	98	0.657	72	0.533
qokol	88	0.432	78	0.487	**qotol**	39	0.403	31	0.507
qokchy	60	0.393	45	0.479	**qotchy**	60	0.352	39	0.464
okol	56	0.537	48	0.522	**otol**	61	0.515	51	0.557
qokey	95	0.435	80	0.454	**qotey**	20	0.441	18	0.490

[Sh]→[ch]

[k] ↓ [t]	**ShcKhy** All: 0.404 Mid-Line: 0.404	**chcKhy** All: 0.541 Mid-Line: 0.513
	ShcThy All: 0.516 Mid-Line: 0.499	**chcThy** All: 0.598 Mid-Line: 0.553

[qo]→[o]

[k] ↓ [t]	**qokedy** All: 0.473 Mid-Line: 0.488	**okedy** All: 0.527 Mid-Line: 0.503
	qotedy All: 0.537 Mid-Line: 0.527	**otedy** All: 0.549 Mid-Line: 0.544

Figure 1: Examples of homologous words conforming to two distribution patterns simultaneously

It's worth emphasizing that a single word can contain more than one of the elements associated with the three patterns I've just outlined; for example, it might begin with [ch] and also contain [k]. When this happens, the word's distribution will necessarily conform to both patterns simultaneously (see Figure 1) unless it is anomalous with respect to one or the other of them. Particularly when we bear this situation in mind, it appears that positional constraints must be operating independently on component parts of words rather than on words as unitary wholes. Indeed, if we switch from word-level analysis to glyph-level analysis, we find that the word-level patterns we've been examining recapitulate glyph-level patterns in both cases to which this other mode of investigation is applicable (it's not applicable to the alternation of [o] with [qo], which involves adding or subtracting a glyph rather than substituting one glyph for another). The mean relative rightwardness scores for [Sh] in all positions and in mid-line positions only (excluding the first and last glyphs of lines) are 0.373 and 0.391 respectively, while those for [ch] are 0.458 and 0.464. Thus, word pairs that begin alternately with [ch] and [Sh] tend to deviate in the same direction as [ch] and [Sh] do more generally. The scores for the glyph [k] (including [cKh]) are 0.479 and 0.484, while those for the glyph [t] (including [cTh]) are 0.505 and 0.541, so word pairs

that alternately contain [k] and [t] also tend to deviate in the same direction as the glyphs themselves do more generally. To a point, of course, there's nothing surprising about these correlations. The distribution of individual glyphs is obviously linked to the distribution of words that contain them. However, the existence of these patterns implies that the affected words are behaving to a surprising degree as nothing more than the sums of their parts.

3. Graphs of Prevalence by Line Position

The scoring method on which I've focused here so far can be used to detect differences in the average positions of words, glyphs, and other textual units, but it tells us frustratingly little *about* those differences. In particular, it reveals nothing about the specific patterns of distribution within the line that must be responsible for the discrepancies we've observed between whole-line and mid-line scores. If we want to draw out such additional details for study, one option we have is to create a graph in which the *x* axis represents relative rightwardness quantized into some number of bins and the *y* axis represents the ratio of pertinent tokens to total tokens within each bin. Each point on the graph can then show the proportion of tokens in some particular part of the line that satisfy a given criterion on a scale ranging from 0 (none of them) to 1 (all of them). If we organize and display our data in this way, we find that distribution patterns can in fact have distinct shapes.

For the three graphs presented in Figure 2, I've assigned a rightwardness score of 0 to the first bin, a rightwardness score of 1 to the last bin, and eight equal divisions of the intervening range of scores to eight bins in the middle. Thus, bins 1 and 10 represent the two extremities, while bins 2-9 represent the mid-line. In every case, the sharpest "jumps" all occur between the extremities and the nearest parts of the mid-line, which reinforces the perception that behavior at the extremities is especially anomalous and, I believe, justifies the decision to separate out those positions analytically. However, distinct patterns can be seen in the mid-line as well, and those are the ones on which I want to focus here.

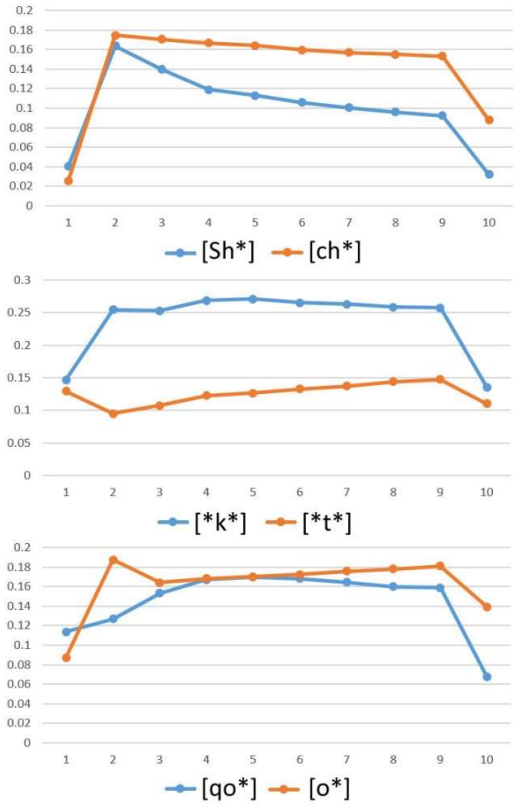

Figure 2: Graphs of prevalence by line position. The word pair set with the higher relative rightwardness score is plotted in orange and the word pair set with the lower relative rightwardness score is plotted in blue.

From the top graph, we learn that the proportions of homologous words beginning with [Sh] and [ch] both peak at the far left of the mid-line and then decrease progressively towards the right. The lower mean relative rightwardness score of the [Sh] set turns out to be due to its proportion decreasing more steeply from left to right than that of the [ch] set. The graph in the middle reveals that, among homologous words that alternate between [k] and [t], the [k] set peaks just to the left of center, whereas the [t] set increases gradually in proportion from the left to right of the mid-line by a factor of about 1½. The bottom graph shows that, among homologous words that begin alternately with [qo] and [o], the [o] set peaks at the left and right of the mid-line with a dip in the middle, whereas the [qo] set shows the opposite trajectory, peaking in the middle with dips at the left and the right. These patterns aren't quite symmetrical, however, and the [o] set rises above the [qo] set for longer on the right than it does on the left, which I assume is the source of its higher mean relative rightwardness score. If the patterns had

been more symmetrical, their left and right halves might instead have cancelled each other out in scoring.

Graphs such as these can provide more nuanced information about the distribution patterns detected by our earlier scoring method, and they also have the potential to reveal further distribution patterns to which our earlier scoring method wouldn't be sensitive, such as horizontally symmetrical ones.

4. Combined Displays of Rightwardness and Downwardness

The graphs presented in Figure 2 still have a serious limitation. Some of the best-known distribution patterns—such as the preference of [p], [f], [cPh], and [cFh] for the first lines of paragraphs—are tied not to rightwardness, but to a second dimension perpendicular to it which we can call "downwardness." Relative downwardness complements, and is analogous to, relative rightwardness: it measures how far proportionally towards the bottom a token appears within a paragraph, and it can be calculated by dividing the number of the line in which a token occurs, counted from top to bottom starting at zero, by the total quantity of lines in its paragraph minus one. Any token in the top line will score 0, any token in the bottom line will score 1, and any token in a line at the exact center of the paragraph will score 0.5. In theory, it should be easy enough to display relative rightwardness data and relative downwardness data together on a Cartesian plane by assigning them respectively to the x and y axes. In practice, this process is complicated by the need to reckon simultaneously with lines of different lengths and paragraphs containing different numbers of lines. One option would be to quantize relative rightwardness and relative downwardness into a two-dimensional array of bins, expanding on the same strategy used to create the graphs in Figure 2. However, I'm concerned that quantizing positional data in two dimensions at once might exacerbate quantization error beyond tolerable limits.

With that in mind, I've devised an alternative approach to visualization that combines several common digital image processing techniques. This approach entails (1) generating a row of pixels for each line of text, with a bright pixel representing each token of the target element and a dark pixel representing anything else; (2) stretching the row horizontally to the width of a square; (3) stacking the rows representing each paragraph and stretching the stack vertically to the height of a square; and then (4) overlaying and averaging the stretched images for all paragraphs. The result is a square grayscale image in which brighter or darker areas correspond to relative positions rightward and downward where the target element is more or less prevalent. One reason for mapping quantity to brightness rather than darkness is that we can then assign the results for multiple target elements to different color channels in a single image so that contrasts of color will correspond vividly to mutual differences in distribution, making these easier to discern. I also find that it can once again be advantageous to separate out the extremities, which in this case means assigning them to their own rows and columns and treating only mid-line tokens and mid-paragraph lines "relatively." As it happens, the first line, mid-paragraph, and last line often display conspicuously different patterns from one another, whereas any differences by position *within* the mid-paragraph tend to be comparatively subtle.

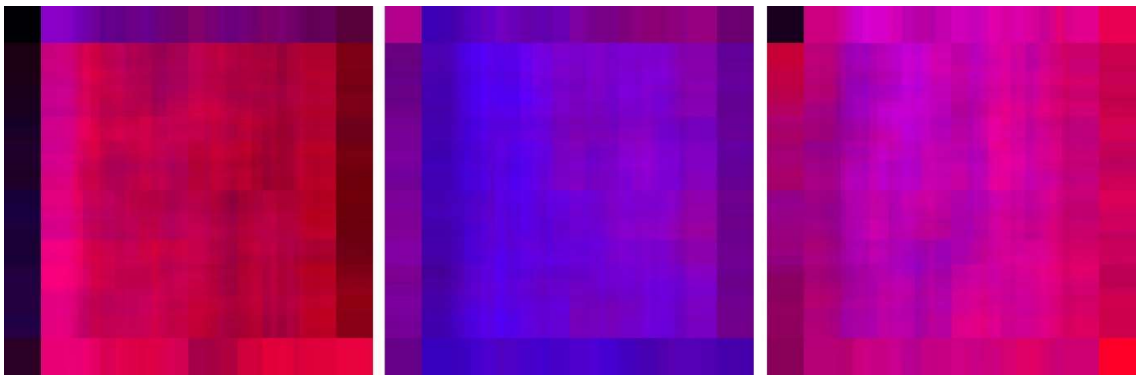

Figure 3: Rightward and downward distribution of words beginning [Sh] in blue and words beginning [ch] in red (left); words containing [k] in blue and words containing [t] in red (center); words beginning [qo] in blue and words beginning [o] in red (right)

The visualizations in Figure 3 each present two sets of results using the red and blue channels of a standard RGB image. They cover the same three contrastive word categories as before, although this time I've included all words that fit the criteria in question, such as beginning with [ch] or containing [k], rather than limiting the scope of analysis to homologous word pairs as before. These visualizations reinforce some of our earlier observations about distribution patterns, but they also enable us to make some new ones. Words beginning with [Sh] turn out to be distinctly more prevalent in the first lines of paragraphs than words beginning with [ch], even though they're consistently less prevalent in lower lines, and the average positions of words containing [k] and [t] also contrast more conspicuously there than elsewhere. Words beginning with [qo] appear noticeably more often as the first words of lines towards the center of paragraphs than they do further upward or downward. Finally, in all three cases, the proportion of the words being studied doesn't decrease sharply at the last position in the last line of the paragraph as it does at the last positions in all other lines, and sometimes it even increases there.

5. Discussion

It's commonly recognized that the sequence of glyphs within a Voynichese word is constrained by rather rigid morphological rules. From the foregoing evidence, however, it seems that the presence of *specific* glyphs within words is simultaneously constrained by additional forces operating throughout the text at the line and paragraph levels. Numerous glyph types can be shown to appear preferentially in certain parts of lines and paragraphs as opposed to others, perhaps causing words containing them to show similar distributions as a secondary effect.

Space doesn't permit me to address some important ancillary questions here, such as whether distribution patterns are affected by different line lengths; whether they play out differently in different sections of the manuscript; whether the second words of lines are more anomalous than other mid-line words, as some previous work has suggested [3, 4]; and whether text outside of paragraphs shows any comparable tendencies. However, I would like to touch at least briefly on the potential implications of these patterns for two other current lines of inquiry: namely, anomalous frequency counts of word-break combinations and the self-citation hypothesis.

Emma May Smith and Marco Ponzi define "word-break combinations" as combinations of the last glyph of one word with the first glyph of the next word within the same line; for example, the word-break combination for the successive words [okal.dam] is [l.d]. They find that frequency counts of actual word-break combinations differ significantly from the counts that would be expected from the set of words involved based on the frequencies of their beginning and ending glyphs if they had been shuffled into a sequence at random. Some combinations turn out to be significantly more frequent than expected, such as [y.q], while others turn out to be significantly less frequent than expected, such as [n.q] [6]. Distribution patterns of the kind we've been examining here offer one possible explanation for these observations. If words that end with one glyph and words that begin with another glyph occur preferentially in the same parts of lines and paragraphs, the corresponding word-break combination should be more common than we would expect from a random distribution, whereas if they tend to occupy non-overlapping positions, the corresponding word-break combination should instead be less

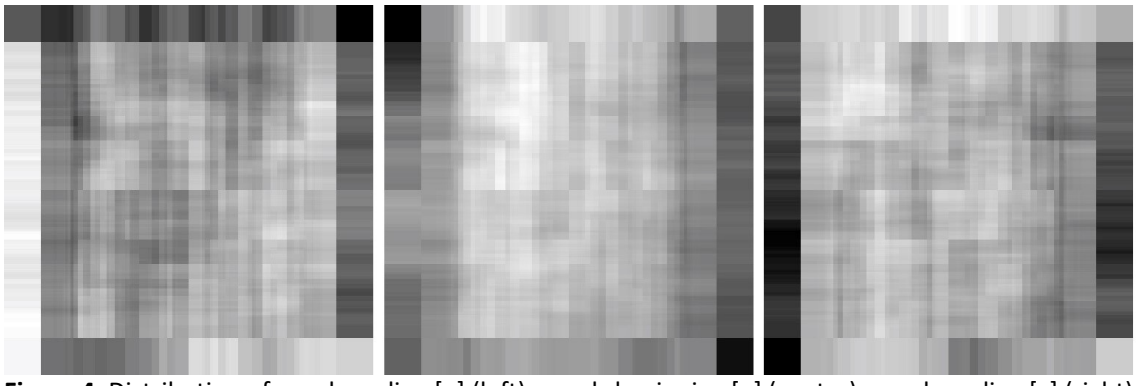

Figure 4: Distribution of words ending [n] (left); words beginning [q] (center); words ending [y] (right)

common. Positional factors of precisely this kind do in fact appear to underlie the anomalous frequency counts of the word-break combinations [y.q] and [n.q]. In Figure 4, we can see that words ending with [n] are especially prevalent as the first words of lines, at the far right of the mid-line, and towards the end of the last line of a paragraph. The positions following next after each of those positions are all ones where words beginning with [q] occur relatively infrequently. By contrast, words ending with [y] tend to appear most often in parts of the line where words beginning with [q] are especially likely to follow them. Whether this finding can be generalized to other word-break combinations remains to be seen.

Torsten Timm observes that words which resemble each other tend to appear near each other both horizontally and vertically on the same pages [7], and largely on this basis, he and Andreas Schinner have developed what they call the "self-citation hypothesis" [8], according to which the bulk of the text of the Voynich Manuscript was generated by iteratively copying strings of nearby text, usually with slight changes such as adding or removing a glyph or substituting one glyph for another. Timm further suggests that a tendency to copy a given string to the same position in a new line could account for the existence of distribution patterns, citing the preference of [m] for line-final position as an example [7]. However, repeated arbitrary changes made to strings during copying should have caused the distributions of any variable elements to converge rather than to diverge. As currently formulated, then, the self-citation hypothesis doesn't account satisfactorily for distribution patterns of the kind described above. On the other hand, those distribution patterns could arguably account for one of the observations on which the self-citation hypothesis rests, in that similar words should predictably appear near each other if positions near each other are subject to similar constraints on word composition.

When it comes to *explaining* distribution patterns, there are various possibilities we might entertain. One is that each line of text corresponds to some unit of meaningfully patterned content, such as a grammatical sentence, a line of poetry, or an entry in a list. Exploratory studies of a few well-known works of poetry show that similar patterns can be detected in them, presumably due to a complex interplay of grammatical, metrical, and stylistic factors. In Virgil's *Aeneid,* for example, if we compare homologous word pair sets ending [es] and [ibus], the [es] set has a midline average rightwardness score of 0.399 with 343 tokens, while the [ibus] set scores 0.708 with 390 tokens—a difference as stark as any presented above. Alternatively, we might hypothesize that distribution patterns arose as a byproduct of some method of encoding meaningful content rather than from the content itself. Here I'll cite just one representative scenario. Fifteenth-century ciphers often sought to increase security by providing multiple options for encoding each plaintext character, and for this ploy to work as intended, a writer needed to alternate repeatedly among those options. One strategy for ensuring that happened would have been to favor different options in different areas of the page. Thus, there's more than one angle from which we could try to *explain* distribution patterns, but the methods outlined above for *identifying* such patterns should be equally applicable to any and all prospective interpretations of them.

6. References

[1] P. Currier, Papers on the Voynich Manuscript, 1976 [1992]. URL: http://www.voynich.nu/extra/curr_main.html.

[2] R. Zandbergen, The Cardan grille approach to the Voynich MS taken to the next level, 2021. URL: https://arxiv.org/abs/2104.12548.

[3] E. Vogt, The Line as a Functional Unit in the Voynich Manuscript: Some Statistical Observations, 2012. URL: https://voynichthoughts.files.wordpress.com/2012/11/the_voynich_line.pdf.

[4] E. M. Smith, Word Position in Quire 20, 2017. URL: https://agnosticvoynich.wordpress.com/2017/01/17/word-position-in-quire-20/.

[5] J. K. Petersen, Medieval Padding, 2020. URL: https://voynichportal.com/2020/07/31/medieval-padding/.

[6] E. M. Smith, M. Ponzi, Glyph combinations across word breaks in the Voynich manuscript, Cryptologia 43.6 (2019) 466–485. doi: 10.1080/01611194.2019.1596998.

[7] T. Timm, How the Voynich Manuscript was created, 2014. URL: https://arxiv.org/abs/1407.6639.

[8] T. Timm, A. Schinner, A possible generating algorithm of the Voynich manuscript, Cryptologia 44.1 (2020) 1-19. doi: 10.1080/01611194.2019.1596999.