【論　文】

On the Directionality of Text in the Voynich Manuscript: An Edge-Based Approach

Jun UTSUMI

Abstract

In this paper, we examine some linguistic properties of the Voynich manuscript. Our approach focuses on the properties of 'word' edges. It is shown that there are clear differences of the symbol distribution between the left edges and the right edges. We argue that these differences indicate the use of capitalization. Since the capital letters are placed at the right-edge, we conclude that the directionality of the text is right-to-left.

Keywords

Voynich manuscript, Text directionality, Capitalization

Introduction

The Voynich manuscript is notorious for its rejection of several attempts of decoding. Although a lot of researches have been done on this strange document, serious considerations about the directionality of the text are few. We admit that its hand-written strokes indicate that the text was written from left to right. And we also accept that the text lines are left-aligned. These facts, however, do not necessarily suggest the text should be read from left to right. Many people have considered this manuscript as a cipher. The purpose of using a cipher is deceiving someone. When you deceive, you would use whatever means available, including misleading the directionality when you write a message.

In this article, we analyzed the text directionality problem of the document. We assumed that the document was not a cipher, but a simple natural language text written using strange symbols. Text directionality is very important for the reconstruction of unknown languages. We focused on the properties of the word edges, because, in general, word edges show some important linguistic properties. Phonologically, for example, English words can start with /str/ sequence, while they cannot end with the same sequence. And orthographically, capital letters can appear in word-initial positions, while they cannot in word-final positions.

About the Text and Tools

We used the original transcription of Mary D'Imperio and Prescott Currier.[1] We assumed that a blank space, which is represented by”/” in this transcription, and a new line, which is represented by “-“ or “#”, indicated some kind of word-boundaries.

And we used the tools provided by NLTK（Natural Language Toolkit）[2], to analyze the distributional frequencies of the symbols in the text.

Data

We analyzed the following frequencies of each symbol used in the document: The total frequency, the frequency as a single symbol word, the frequency in the left-edge positions, the frequency in the right-edge positions, and the frequency in the non-edge positions. Figure 1 shows the distributional frequencies of each symbols.（The full result is presented as Table.1.）LEF stands for the left-edge frequency, REF for the right-edge frequency, and NEF for non-edge frequency.

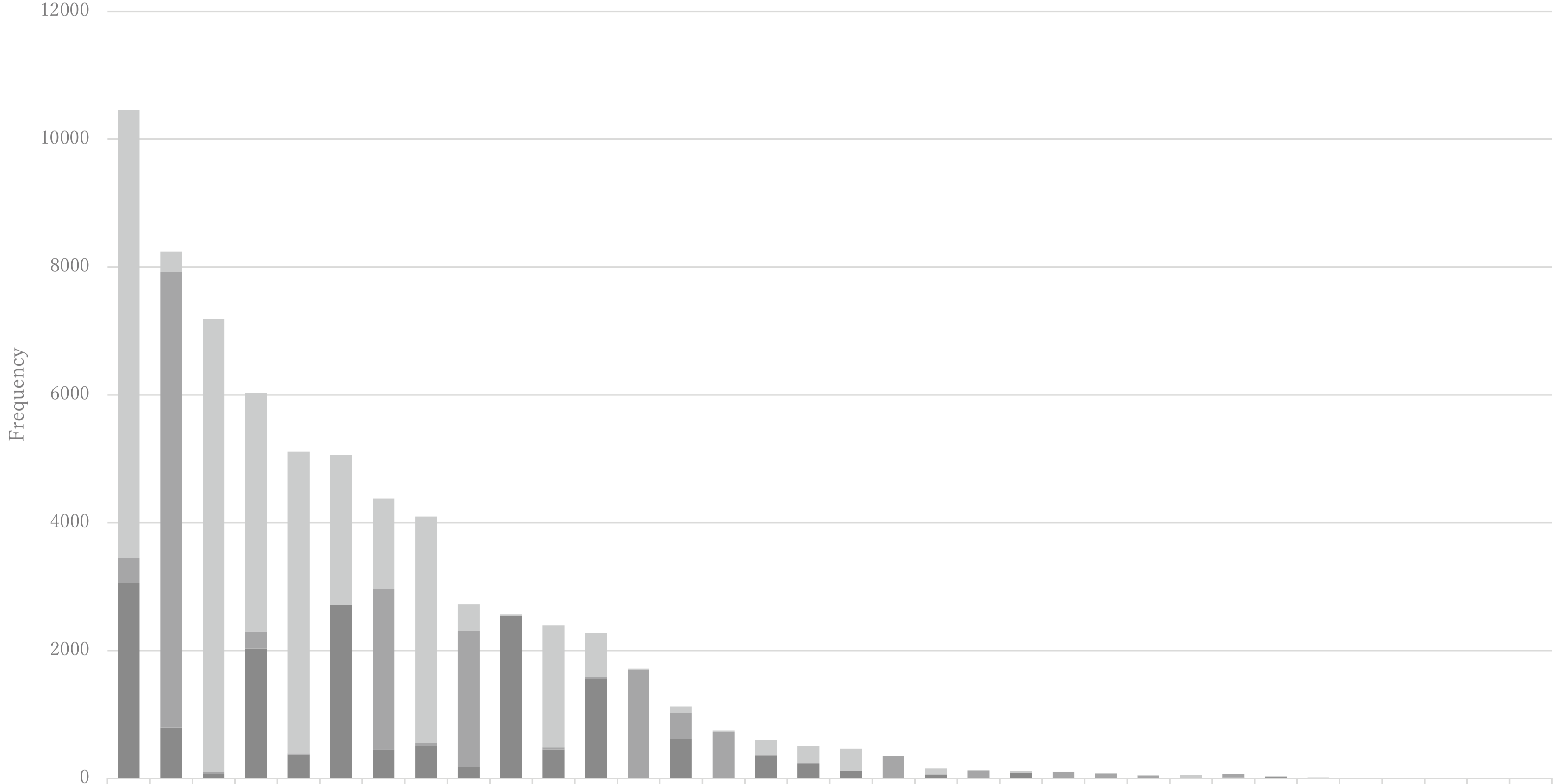
The result shows a clear discrepancy in the distribution of the symbols. According to their edge-distribution patterns, these symbols can be divided into the following three groups:

- Group A consists of “M”, “N”, “T”, “I”, “K”, “7”, “H”, “L”, and “5”. The symbols in Group A do not occur in the left-edge.
- Group B consists of “Y”. That symbol never appears in the right-edge.
- Group C consists of the other symbols. The symbols in this group appear in either edge.

In addition to the occurrence/nonoccurrence distinction, there are other differences, which are not so clear-cut. The symbols in Group A and Group B tend to have lower total frequencies than those in Group C. And the symbols in Group A, except “I” and “7”, tend to show quite low frequencies in non-edge positions. In fact, the symbols that show rather high right-edge frequencies tend to show quite low frequencies in the non-edge and left-edge positions. The symbol “J”, for example, shows the high frequency of 341, while its non-edge frequency is 7 and its left-edge frequency is 5. Another example is the symbol “D”, whose right-edge frequency is 88, while its non-edge frequency is 7 and its left-edge frequency is 6. The symbols “3”, “U”, and “G” show similar tendencies.

[1] The transcription was obtained from http://www.ic.unicamp.br/~stolfi/voynich/mirror/gillogly/voynich.orig.

[2] The tools are introduced by Bird, Klein, and Loper（2009）.



	O	9	C	8	A	S	E	F	R	4	P	Z	M	2	N	Q	B	X	J	V	T	W	D	3	U	I	6	Y	K	G	7	H	L	5
▨ NEF	7005	318	7089	3734	4729	2343	1410	3542	418	27	1912	699	20	100	21	238	265	350	7	94	20	40	7	14	11	52	4	14	2	2	7	3	1	0
▨ REF	395	7123	37	269	17	7	2518	44	2127	7	32	25	1697	408	728	9	14	5	341	8	114	2	88	61	44	3	48	0	14	10	6	4	6	2
■ LEF	3059	798	63	2030	370	2708	449	508	176	2534	450	1555	0	618	0	358	226	108	5	54	0	80	6	8	1	0	16	19	0	1	0	0	0	0

Figure 1. Distributional frequencies of the Voynich symbols in LEF, REF and NEF.

Hypotheses

The symbols' distributional characteristics suggest two possible hypotheses. One is that the symbols in Group A are some kinds of punctuation symbols. And the other one is that the symbols in Group A are capital letters.

Punctuation Hypothesis

Let us consider the possibility of punctuation symbols first. The punctuation symbol hypothesis does not conflict with the result mentioned above. Punctuation symbols are usually placed at the end of words. In a language with left-to-right text directionality, they appear in the right-edge, which explains Group A's non-occurrence in the left-edge positions.

And the hypothesis does not conflict with chronological evidence. According to Zyats, *et al.* (2016), the physical properties of the Voynich manuscript suggest that the document was produced in early fifteenth century. Clemens and Graham (2007) illustrate the various use of "punctus" and other symbols for punctuation in medieval era. Therefore, the use of various symbols as punctuation signals in the manuscript is not unreasonable.

The punctuation hypothesis, however, seems to be implausible, because it assumes too many punctuation types, though we have no concrete evidence against the claim.

Capitalization Hypothesis

Let us turn to the capitalization possibility. Capitalization is, in other words, using special symbols, which have the same phonetic values with the normal ones, in edge-specific positions. In Latin and other alphabetical scripts, special symbols, or capital letters, are placed in the word-initial positions when the words are in the sentence-initial positions, or the words represent proper names. Using edge-specific symbols, however, is not limited to the word-initial positions. In Greek script, when you use the letter "σ" in a word-final position, it should be replaced by the special form "ς".

The capitalization hypothesis explains the non-occurrence of Group A symbols in the left-edge positions if we assume the text direction is right-to-left. It also explains the lower total frequencies of Group A symbols.

One might argue, however, that Group A symbols do appear in non-edge positions, and that it contradicts the claim that Group A symbols are capital letters. This argument is untenable, because we do have word-medial occurrences of capital letters, such as *McIntosh* and *DiCaprio*.

One might also argue that if the Voynich symbols are divided into the uppercase group and the lowercase group, the number of Group A symbols is too small to form the uppercase group. The claim sounds quite convincing. If we look into the result closely, however, the putative problem can be solved. As we mentioned above, some symbols in Group C show high frequencies in right-edge positions, while they show quite low frequencies in left-edge and non-edge positions.

These symbols are “J”, “D”, “3”, “U”, and “G”. If we include these symbols into the uppercase class, the number of the uppercase letters is closer to the half. Transcribing a hand-written manuscript, which is full of unfamiliar symbols, is troublesome and prone to be misled. In order to decide whether the inclusion is right or not, we need to reevaluate the transcription with the insight provided by our finding.

Comparison with Another Text

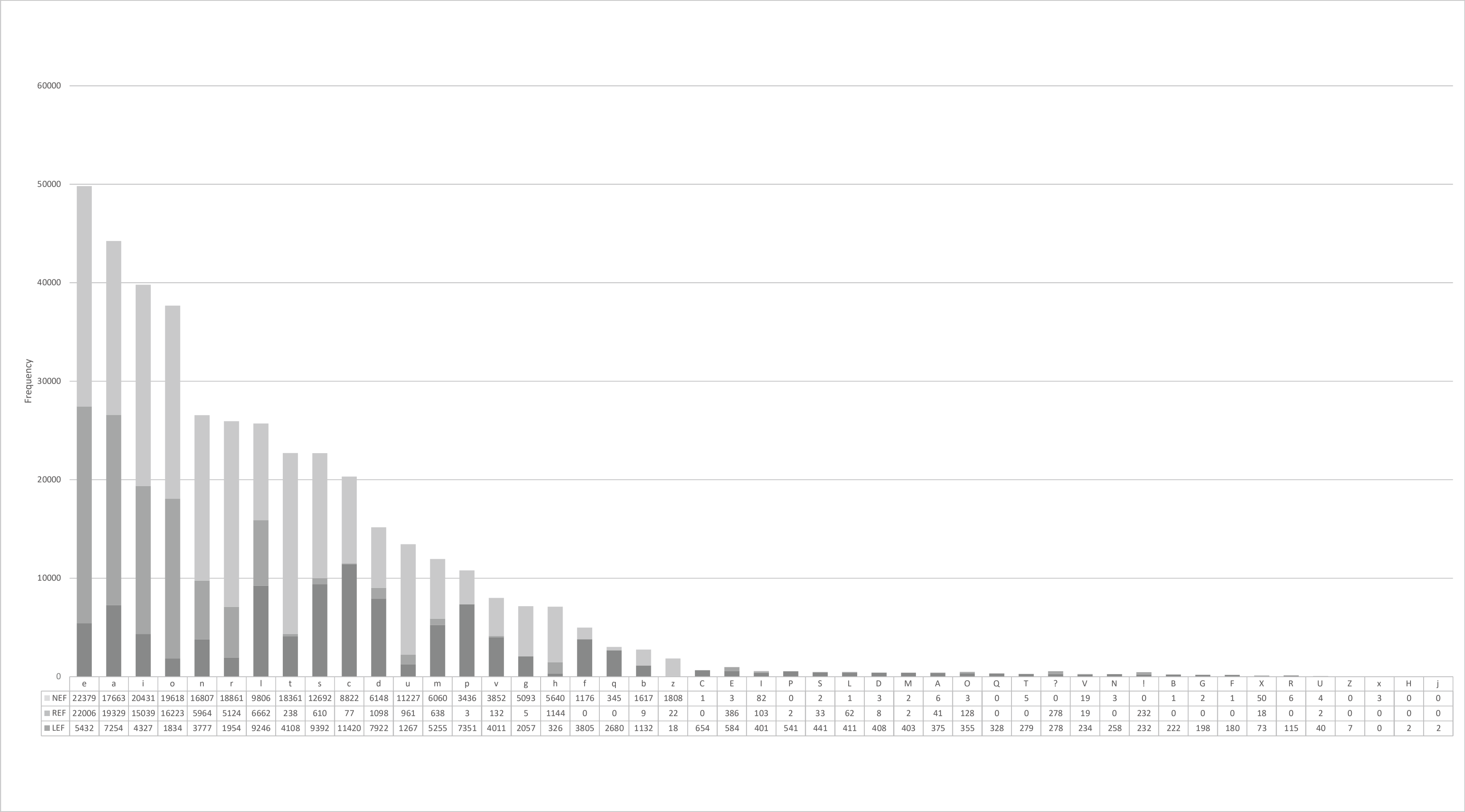
To confirm that our claim about capitalization is on the right track, let us compare the distributional frequencies of the Voynich manuscript with those of other texts. We chose Dante’s *La Divina Comedia*[3] to compare with the Voynich manuscript, because Dante’s work seemed to be chronologically closer to the Voynich manuscript. We analyzed the text in the same way as the Voynich manuscript. The result is shown as Figure 2.[4]

As expected, most uppercase letters show no frequency in right-edge positions. Their behavior is parallel with Group A symbols in the Voynich manuscript, if we exchange the right-edge positions with the left-edge positions. Some uppercase letters, however, show high frequencies in the right edges. They are the uppercase letters, “E”, “S”, “L”, “A”, and “O”, most of which appear as one symbol words. And the uppercase letters, “I”, “V”, and “X”, show high frequency in the non-edge positions. These symbols are not only used as normal letters, but also used as Roman numerals. And in this text, most of these uppercase letters do appear as Roman numerals. Therefore, these cases pose no problem against our claim.

The total frequency difference between uppercase letters and lowercase letters is similar to the difference between Group C symbols and Group A symbols, which reinforces our claim. In fact, we observe further similarities between the uppercase letters and Group C symbols. They show the similarities only when the other one’s directionality is reversed. For example, in the Voynich manuscript, the symbol “R” in Group C shows very high right-edge frequency of 2127, while it shows very low left-edge frequency of 176. In *La Divina Comedia*, the lowercase “c” shows very high left-edge frequency of 11420, while it shows very low right-edge frequency of 77.

[3] The text is taken from Project Gutenberg（http://www.gutenberg.org/ebooks/1012）.

[4] For the sake of brevity, we removed alphabets with diacritics（“ù”, “Ë”, …）, punctuation marks, and other non-alphabetical symbols.



	e	a	i	o	n	r	l	t	s	c	d	u	m	p	v	g	h	f	q	b	z	C	E
NEF	22379	17663	20431	19618	16807	18861	9806	18361	12692	8822	6148	11227	6060	3436	3852	5093	5640	1176	345	1617	1808	1	3
REF	22006	19329	15039	16223	5964	5124	6662	238	610	77	1098	961	638	3	132	5	1144	0	0	9	22	0	386
LEF	5432	7254	4327	1834	3777	1954	9246	4108	9392	11420	7922	1267	5255	7351	4011	2057	326	3805	2680	1132	18	654	584

	I	P	S	L	D	M	A	O	Q	T	?	V	N	!	B	G	F	X	R	U	Z	x	H	j
NEF	82	0	2	1	3	2	6	3	0	5	0	19	3	0	1	2	1	50	6	4	0	3	0	0
REF	103	2	33	62	8	2	41	128	0	0	278	19	0	232	0	0	0	18	0	2	0	0	0	0
LEF	401	541	441	411	408	403	375	355	328	279	278	234	258	232	222	198	180	73	115	40	7	0	2	2

Figure 2. Frequency Distribution of Alphabets in Dante's *La Divina Comedia*

Conclusion

The edge-specific distributional discrepancies observed in the Voynich manuscript indicate that the use of capitalization in the right-edge positions, which strongly suggests that the text directionality of the manuscript is right-to-left.

If our capitalization hypothesis is correct, it entails that the symbols in the Voynich manuscript is a variation of European alphabets. The use of capitalization is limited to European alphabets, such as Latin script, Greek script, and Cyrillic script. Therefore, the Voynich script is written in either one of these systems or some variation related to them.

Our analysis also implies that the number of the symbols with concrete phonetic value is significantly reduced. Uppercase letters and lowercase letters share phonetic values. If our approach is on the right direction, the phonetic repertoire of Voynich symbols is reduced by half. It also entails that the writing system of the document is some kind of an alphabet, not a syllabary, because the symbols are too few to represent sufficient syllable patterns. Even if the punctuation symbols hypothesis, which we dismissed in a less convincing way, was correct, the same implication holds. If the symbols in Group A are punctuations, their phonetic values are “silence”, which reduces the manuscript’s phonetic repertoire significantly.

REFERNCES

Bird, Steven, Ewan Klein, and Edward Loper（2009）*Natural Language Processing with Python*, O'Reilly Media.

Clemens, Raymond, ed.（2016）*The Voynich Manuscript*, Yale University Press.

Clemens, Raymond, and Timothy Graham（2007）*Introduction to Manuscript Studies*, Cornell University Press.

Zyats, Paula, Erin Mysak, Jens Stenger, Marie-France Lemay, Anikó Bezur, and David D. Driscoll（2016）“Physical Findings”, in Clemens, ed.（2016）, 23–37.

Symbols	Total Freq.	One Symbol Word Freq.	Left-Edge Freq.	Right-Edge Freq.	Non-Edge Freq.
O	10442	17	3059	395	7005
9	8186	53	798	7123	318
C	7188	1	63	37	7089
8	6001	32	2030	269	3734
A	5115	1	370	17	4729
S	5058	0	2708	7	2343
E	4362	15	449	2518	1410
F	4090	4	508	44	3542
R	2686	35	176	2127	418
4	2564	4	2534	7	27
P	2394	0	450	32	1912
Z	2270	9	1555	25	699
M	1717	0	0	1697	20
2	1005	121	618	408	100
N	749	0	0	728	21
Q	603	2	358	9	238
B	503	2	226	14	265
X	463	0	108	5	350
J	350	3	5	341	7
V	156	0	54	8	94
T	134	0	0	114	20
W	122	0	80	2	40
D	98	3	6	88	7
3	82	1	8	61	14
U	56	0	1	44	11
I	55	0	0	3	52
6	53	15	16	48	4
Y	33	0	19	0	14
K	16	0	0	14	2
G	13	0	1	10	2
7	13	0	0	6	7
H	7	0	0	4	3
L	7	0	0	6	1
5	2	0	0	2	0
Total	66593	318	16200	16213	34498

Table 1. Frequencies of the Symbols in the Voynich Manuscript