Analyzing Voynich Manuscript

By Manish Rajkarnikar

05/5/2004

Problem Description:

Voynich Manuscript is one of the most mysterious manuscripts written in late medieval or early modern age. It is written in a script that has never been understood or seen elsewhere. Nobody knows what it means. Lots of people have tried to decipher it, but failed. Many theories have been proposed to explain the Voynich manuscript. Some say the Voynich Manuscript is just a hoax, while others say that it is genuine. It consists of pictures of stars which can be seen by modern telescope only. It also has other sections such as herbal, astronomical, chemical etc which makes it very interesting. The objective of this project is to analyze Voynich manuscript and to find out if it is just a hoax or a real language. There are different approaches that can be used to perform this task.

[1] explains how Voynich Manuscript text is converted into computer readable characters. Many characters in the Voynich Manuscript cannot be represented exactly by any of the existing alphabets. There are some rare characters and there are what appear to be ligatures of several characters **[1]**. However, few transcriptions have been created which try to map majority of manuscript's characters into computer readable text and EVA is one of them. The transcription that has been adopted for this project is also based on EVA. It assumes that spaces between words are truly word separators. It uses some combination of letters for rare characters.

[3] reports on an experiment using character based Entropy for comparison between human language and Voynich manuscript. Entropy is a numerical measure of the uncertainty in a sequence or string of characters. There are two types of entropy: conditional and unconditional entropy. Conditional entropy is amount of uncertainty about the next event, given that the current one is known while unconditional entropy is the entropy calculated without prior knowledge of any events. This concept of entropy has also been adopted for this project. But, instead of character based entropy, word based unconditional entropy is used here.

Zipf law is another tool which is very useful in analysis of text. Zipf's law concerns the frequency of words in a piece of text. It says that if one orders the words according to decreasing frequency and label the most frequent word as number one, the second most frequent word as number two, etc, and then make a plot of the frequency of this word according to the rank, the result should show a straight line with a slope of -1 **[4]**. Both scales have to be logarithmic in this case. This holds true for all the human languages and this project intends to test if it holds true for Voynich manuscript too.

[2] points out the fact that there is some kind of relation between the letter occurring in any type of text. This relation is maintained to convey some kind of meaning. The order in which all text elements are needed to be placed is decided by the rule of grammar and specific content. Therefore, texts written in any language are highly structured. This relation does not exist in gibberish text since they do not follow any grammatical rules. This property of text is used in this project to analyze if voynich is really a gibberish or true human language.

Approach

There have been lots of analyses that have been performed on the Voynich Manuscript. Analyses used for this project are as follows:

1. Statistical Characteristics of Text
    a. Zipf Law: It is related to the frequency of words in text. According to this law, if all the words found in a text are listed against their frequency and ranked in order of decreasing frequency, than the product of the rank and frequency should be same for all the words. Such kind of table can be made for the Voynich manuscript and tested if they satisfy the Zipf law

b. Token vs. Word Types: Token is any string found in a text. Its count concerns with the total number of words found in the text while words type concerns with distinct token. It is counted by number of distinct tokens occurring in the text. In human language, there is a relation between word types and tokens. This relation can also be tested on MS. d. It is found that few words e.g. “the”, “is” etc have more frequency than any other words. A test can be done on manuscript to see if such words exist in manuscript too.

2. Entropy :

It is the quantity by which the receiver’s uncertainty is reduced when the message is received. It is low if one can predict what word/character is coming next. But if one does not know what is coming next and probability of all the upcoming events are same, then Entropy is maximum. Entropy can be calculated by following formula

$$H(x) = -\sum p(x) * \log2\, p(x) \;\ldots\ldots\ldots\ldots(1)$$

where x is an upcoming event and P(x) is its probability

A text containing 1000 words will have a word entropy less than 2log(1000) depending on the distribution of the word frequencies. Entropy of manuscript can be calculated by finding the probability of each word and then plugging the value calculated in equation 1. Entropy of different human languages can also be calculated using the same process. These values can then be compared to find if there are any kinds of relation between them.

3. Letter Serial Correlation (LSC)

LSC is one of the orders found in texts. It can be determined in two ways: first, by actual measurement and second by mathematical calculation. In both the methods the whole text of length L is divided into equal k chunks of size N each.

a. Actual Measurement

In this method, if the numbers of occurrences of letter *x* in any two adjacent chunks, *i* and *i+1,* be *Xi* and *Xi+1*, we will be measuring the following sum taken over all letters of the alphabet, *x* which varies between 1 and *z* where *z* is the number of letters in the alphabet and over all chunks (i.e. for *i* varying between 1 and *k*)

$$S_m=\sum(X_i - Xi+1)^2$$

b. Mathematical Calculation

In this method following formula is used.

$$S_e=(1-1/k) \sum 2M_x(L-M_x)/(L-1)$$

where M_x is total number of occurrences of a specific letter *x* in the entire text

Here calculation method is based on assumption that once a letter is picked for a word, stock of available letters are unlimited which is in fact only true for random text. i.e gibberish text. Hence if we find ΔS , where $\Delta S= S_m - S_e$, of different types of text, then ΔS should be small for gibberish text and higher for human text **[2]**.

Evaluation Plan:

All the analysis tests were also carried out in different types of texts – English, non-English and gibberish to compare the results. Texts which were used for evaluation purpose are as follows:

- Unigram model (gibberish) text generated from above text.
- Character based random text generated using perl's Silly::Werder module.
- Conan Doyle, "The adventures of Sherlock Holmes" (104507 tokens)
- Spiros Doikas, "Sangharakshita" (Greek text 54100 tokens)
- Francois-Alphonse Aulard, "Les grands orateurs de la Revolution"
  (French text, 62559 tokens)

Here text written in other languages were also used as gold standard along with English text because there are chances that VMS can be something written in language different from English. So we wanted to make sure that a property of such languages was not excluded in experiment.

All the analysis was done using the same programs as used for Voynich manuscript. It was expected that if Voynich was a human language, then the results should be similar to that we get from English or non-English texts and different from gibberish text. Specific evaluation plans were as follows:

Zipfs law:

Here frequency of most common words for all sample texts was measured. A graph was plotted between rank and frequency of these common words for all the text to see if they followed zipfs law. If the graph showed a straight line with negative slope of 1, then it was to be concluded that the text is human language.

Entropy:

Here entropy for text in different languages, voynich manuscript and gibberish text was measured. Entropy for human language should be less than that of Gibberish as prediction of next event occurring can be done easily in human language than in gibberish text. Here if entropy of voynich manuscript is found closer to gibberish it was to be concluded that it was gibberish and opposite otherwise

Letter Serial Correlation

For this method S_m and S_e for different type of text was calculated for different values of n. A graph was drawn between S_m / S_e against n for different types of text. The pattern in the graph for all the human language was expected to match while the pattern for the gibberish text was expected be different one. For voynich text, if it was human language than it's pattern should have inclined more towards real text than from gibberish text. If it was just gibberish, then it should have been opposite.

Experimental Observation:

1. Zipfs' Law:

Table 1.1-6 show frequency and rank obtained for different types of text. Graphs are plotted from these tables in Figure 1.1-6 to show relation between rank(x-axis) and frequency(y-axis)

Rank	Word	Freq	Percent	K
1	DE	3049	4.51	3049
2	LA	2276	3.37	4552
3	IL	1718	2.54	5154
4	A	1628	2.41	6512
5	ET	1591	2.35	7955
6	LE	1388	2.05	8328
7	L	1333	1.97	9331
8	LES	1258	1.86	10064
9	QUE	1005	1.49	9045
10	DES	792	1.17	7920
11	D	779	1.15	8569
12	QU	748	1.11	8976

Table 1.1 French

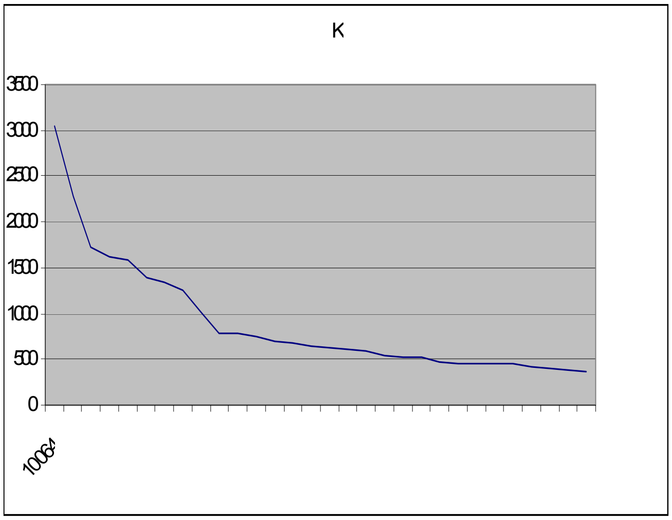

Fig 1.1 French

Rank	Word	Freq	Percent	K
1	TO	285	0.32	285
2	AND	267	0.3	534
3	I	255	0.29	765
4	THE	222	0.25	888
5	IT	221	0.25	1105
6	YOU	219	0.24	1314
7	THAT	208	0.23	1456
8	A	196	0.22	1568
9	WELL	195	0.22	1755
10	HE	158	0.18	1580
11	THERE	157	0.18	1727
12	GOOD	153	0.17	1836

Table 1.2 Unigram Gibberish

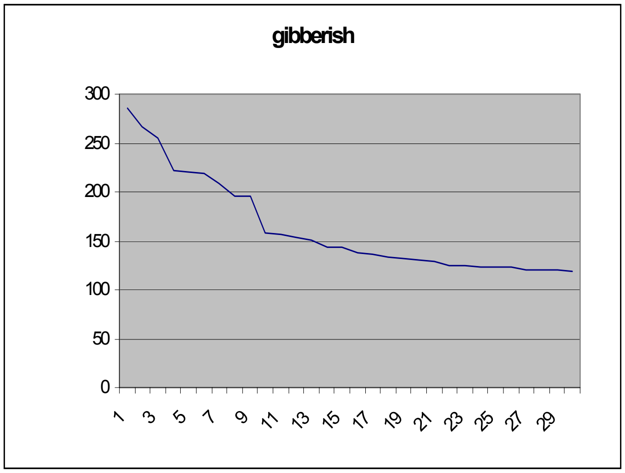

Fig 1.2 Unigram Gibberish

Rank	Word	Freq	Percent	K
1	DAIIN	805	2.17	805
2	OL	525	1.42	1050
3	CHEDY	495	1.34	1485
4	AIIN	456	1.23	1824
5	SHEDY	424	1.14	2120
6	CHOL	380	1.03	2280
7	OR	348	0.94	2436
8	AR	344	0.93	2752
9	CHEY	339	0.92	3051
10	QOKEEY	308	0.83	3080
11	QOKEEDY	301	0.81	3311
12	DAR	297	0.8	3564

Table 1.3 Voynich

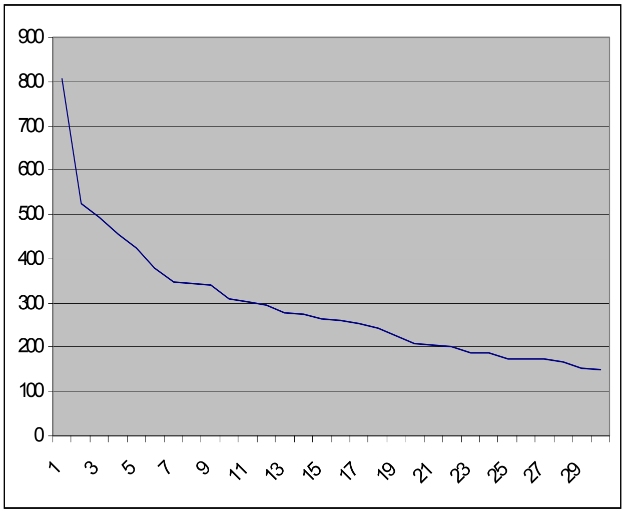

Fig 1.3 Voynich

Rank	Word	Freq	Percent	K
1	THE	5631	5.31	5631
2	I	3031	2.86	6062
3	AND	3017	2.85	9051
4	TO	2743	2.59	10972
5	OF	2658	2.51	13290
6	A	2642	2.49	15852
7	IN	1765	1.66	12355
8	THAT	1751	1.65	14008
9	IT	1731	1.63	15579
10	YOU	1503	1.42	15030
11	HE	1483	1.4	16313
12	WAS	1410	1.33	16920

Table 1.4 English

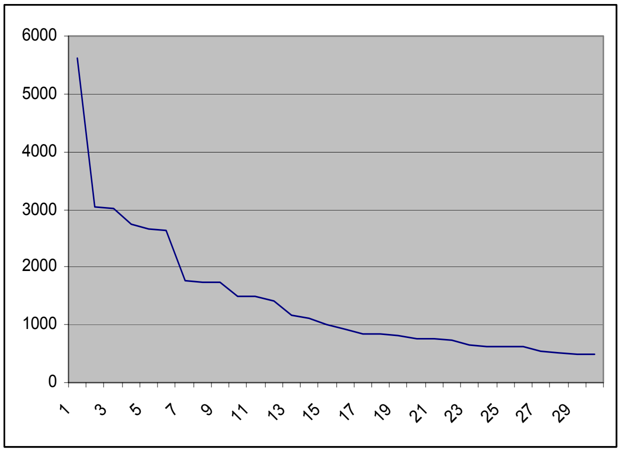

Fig 1.4 English

Rank	Word	Freq	Percent	K
1	íá	1660	3.06	1660
2	êáé	1612	2.97	3224
3	ôï	1179	2.17	3537
4	ôïõ	985	1.82	3940
5	ôçò	903	1.66	4515
6	åßíáé	854	1.57	5124
7	ôçí	807	1.49	5649
8	ðïõ	740	1.36	5920
9	ìå	689	1.27	6201
10	ç	658	1.21	6580
11	ìáò	643	1.18	7073
12	áðü	573	1.06	6876

Table 1.5 Greek 1

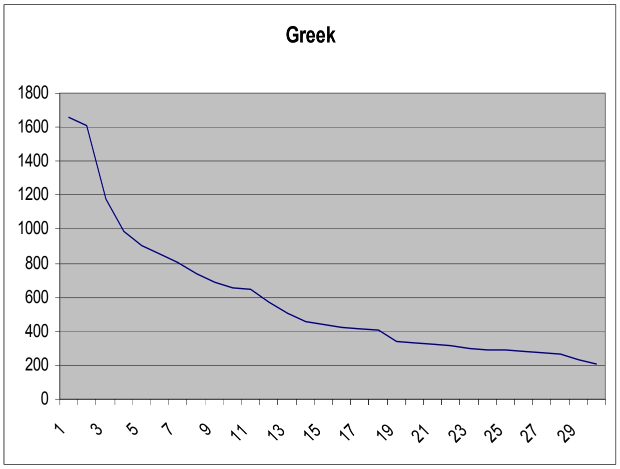

Fig 1.5 Greek 1

Rank	Word	Freq	Per	K
1	O	18	0.06	18
2	NGFTEETKARDMIEX	1	0	2
3	ACFINKIPMIOLICIKIE	1	0	3
4	UNDCIVDIGOTHLADJUL	1	0	4
5	URPACCNON	1	0	5
6	ROXTHUICODIPCOU	1	0	6
7	SPRUMFALVAKSILULT	1	0	7
8	DUMROXZYPETTZ	1	0	8
9	HAZMIRRANPELNONAND	1	0	9
10	HOENSTALDIIRRBEL	1	0	10
11	GOBNOYEBERLT	1	0	11
12	NOIPITICTYROSO	1	0	12

Table 1.6 Character Based Gibberish

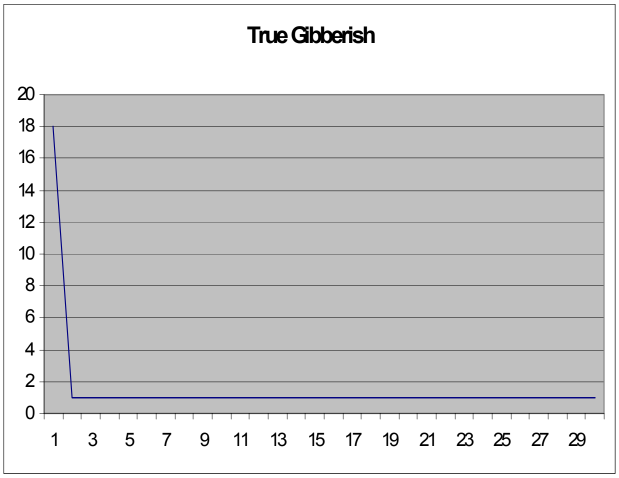

Fig 1.6 Character Based Gibberish 1

Data in Tables 1.1, 1.4 and 1.5 are obtained from real human language and have the typical characteristics of Zipfs Law. On their graphs, we observe a curve which is approximately a straight line with slope of -1. This demonstrates the fact that product of rank and frequency for all languages are indeed approximately constant.

Data obtained for Voynich Manuscript in Table 1.3 matches close to that of human language. It is clearly different from graph obtained from true gibberish. However, we can also note that graph of unigram based gibberish in Fig 1.2 matches with graph of Fig 1.1,1.4 and 1.5. But this can be explained with the fact that value of K in first row and last row are substantially different which is a not characteristic of Zipfs' law.

Hence, on base of the data and graphs obtained, it can be asserted that VMS is not entirely a gibberish text.

Entropy

It was found in the experiment that the value of entropy for regular English text is 9-10 and that of gibberish is 12. In gibberish, once a word is chosen, the following word does not depend on the earlier word chosen. That means the following word can be anything and probability of words occurring next is same for all. But for human language, once a word is selected from pool of words, probability of word occurring next is not same for all the words in pool. This is the reason human languages have low entropy and gibberish has high entropy.

In the experiment, it was found that value of entropy for Voynich Manuscript was found somewhere in between 10 and 11 which is inclined more towards property of regular human text. Hence this characteristic of VMS suggests that it is more likely to be human language. However, this property cannot be used as concrete proof. A carefully written gibberish text as in ASAAA ASAAA CCC CCC CCC EEEE has low entropy but still is gibberish.

Letter Serial Correlation:

Table 2.1-4 show different values of S_m and S_e calculated for different types of text. By looking at all the figures we can find some kind of similar pattern in fig 2.1, 2.2 and 2.4. However fig 2.3 differs from other graph as curve for the S_m in it, is pointing upwards. It is trying to get near to S_m curve. However in other figures distance between S_m and S_e continues to grow.

This is occurring probably because S_m and S_e in random text are supposed to be close as Se is only calculated value which is based on concept of randomized text as described above. But it also seems reasonable to assume that the shape of the experimental curve for the LSC is affected by a number of various factors. One of such factors is that when text is divided into equal k chunks of size N each and if text length L, is not exactly divisible by k then there is some text where is truncated at the end. Length of this truncated text can vary from 1 to N-1. This number can be as big as 9999 when N=10000 and can have a great impact on the result.

	S_m	S_e	S_m/S_e
1	426108	412983.6	1.03178
3	849882	757227.5	1.12236
5	1199438	1032734	1.161421
7	1518288	1239504	1.224916
20	1821434	1377472	1.322302
50	2136764	1446591	1.477103
100	2490730	1446854	1.72148
200	2905670	1378121	2.10843
500	3553698	1239696	2.866587

1000	4673722	1032225	4.527814
5000	7541976	745468.8	10.11709

Table 2.1 French

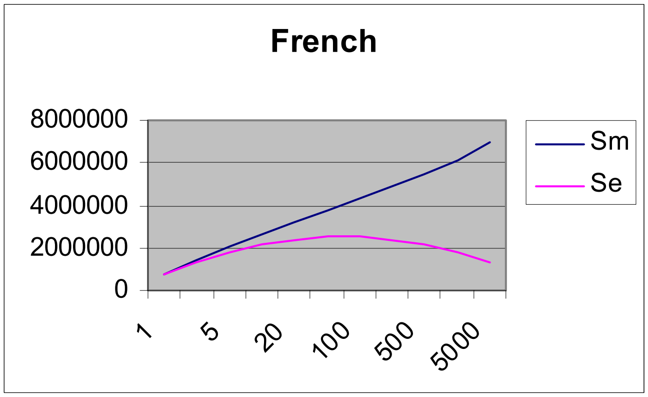

Fig 2.1.1 Graph of S_m and S_e VS n in French Text

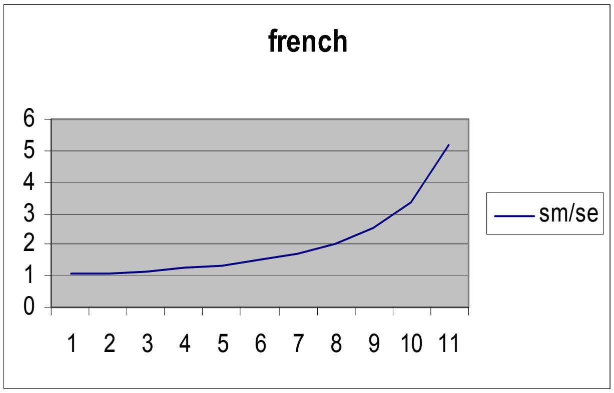

Fig 2.1.2 Graph of S_m/S_e VS n in French Text

	S_m	S_e	S_m/S_e
1	751868	708384.8	1.061384
3	1413552	1299992	1.087354
5	2053144	1774825	1.156815
7	2665680	2132883	1.249801
20	3221428	2374102	1.356904
50	3760836	2498433	1.505278
100	4293070	2505868	1.713207
200	4849268	2396263	2.02368
500	5432498	2168902	2.504723
1000	6098784	1824413	3.342874
5000	6964660	1351999	5.15138

Table 2.2 English

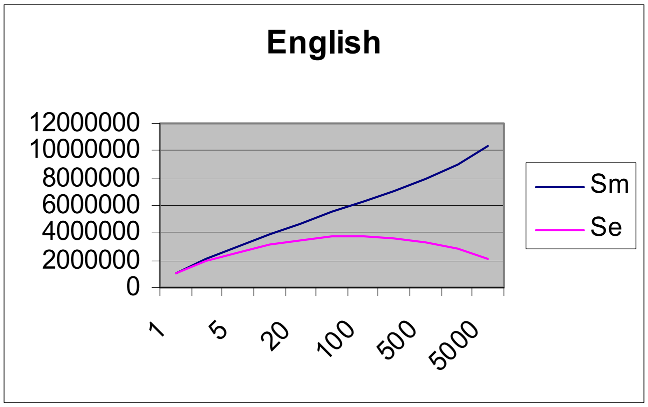

Fig 2.2.1 Graph of S_m and S_e VS n in English Text

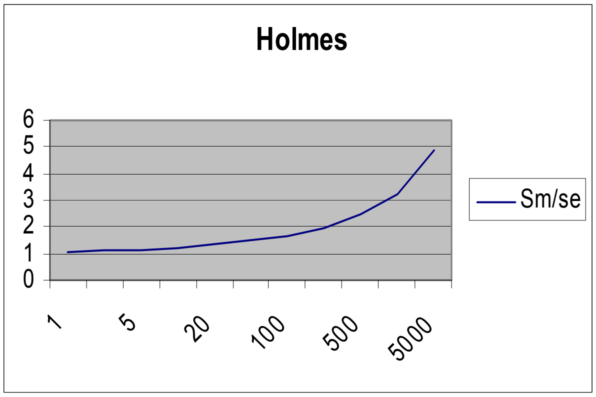

Fig 2.2.2 Graph of S_m/S_e VS n in English Text

	S_m	S_e	S_m/S_e
1	1098176	1039539	1.056406
3	2082304	1910066	1.090174
5	2999288	2611583	1.148456
7	3888652	3144091	1.236813
20	4710538	3507524	1.342981
50	5497598	3701832	1.485102
100	6281616	3727006	1.685432
200	7099734	3582895	1.981564
500	7990810	3268756	2.444603
1000	8963762	2785192	3.218364
5000	10392708	2120605	4.900821

Table 2.3 Character Based Gibberish

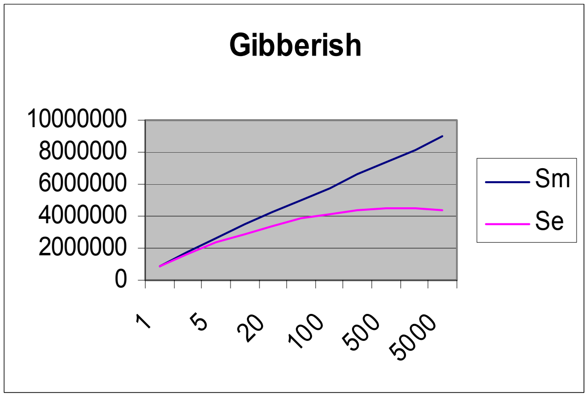

Fig 2.3.1 Graph of S_m and S_e Vs n for Gibberish

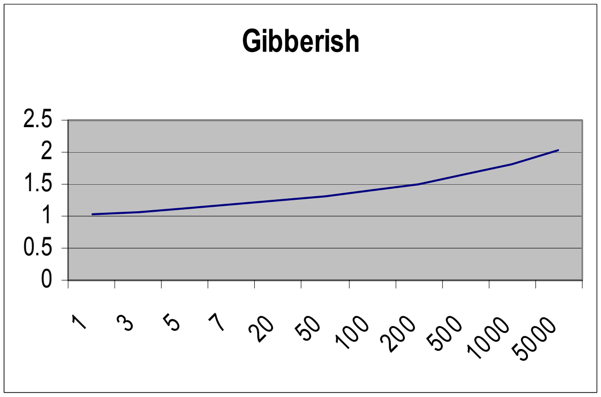

Fig 2.3.2 Graph of S_m/S_e VS n for Gibberish Text

	S_m	S_e	S_m/S_e
1	893914	874271.7	1.022467
3	1753594	1654844	1.059674
5	2607476	2341717	1.113489
7	3444836	2934893	1.173752
20	4244212	3434291	1.235833
50	5028388	3839833	1.309533
100	5807406	4151465	1.398881
200	6594216	4368893	1.509356
500	7372222	4490642	1.641686
1000	8157590	4516733	1.806082
5000	8957556	4415445	2.028687

Table 2.4 Voynich Text 1

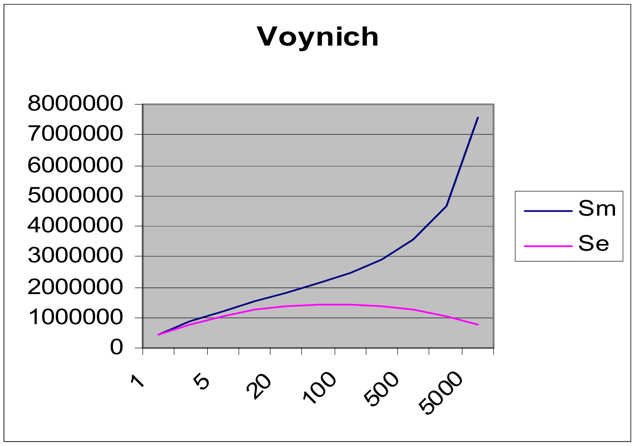

Fig 2.4.2 Graph of S_m/S_e VS n in Voynich Text

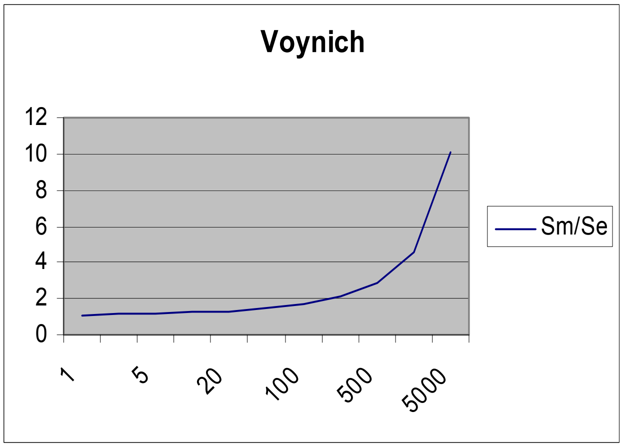

Fig 2.4.2 Graph of S_m/S_e VS n in Voynich Text

Conclusion:

Based on above observation, it can be concluded that VMS has property similar to that of human text and is bit different from gibberish. Properties of text which are analyzed above hold true for any language i.e they are language independent. However, although they give strong indication of VMS being a human language, they cannot prove it. They also show VMS has properties which come in between human text and gibberish. Hence, VMS cannot be definitely defined as a human language or gibberish. It can also be some kind of carefully written gibberish which matches close to real human text.

References:

[1] Landini, Gabriel(1998): *A Well-kept Secret of Mediaeval Science: the Voynich manuscript.* Journal of the University of Birmigham Medical and Dental Graduates Society

[2] McKay, Brendan & Perakh, Mark(1999). *Study of Letter Serial correlation (LSC) in some English Hebrew, Aramaic and Russian texts*. Retrieved April 15, 2004 from http://www.nctimes.net/~mark/Texts/Serialcor1.htm

[3] Zandbergen, Rene (2002). *From digraph entropy to word entropy in the Voynich Manuscript*. Retrieved April 15, 2004 from http://www.voynich.nu/extra/wordent.html

[4] Zandbergen, Rene (2002). *The Voynich Manuscript*. Retrieved April 15, 2004, from http://www.voynich.nu/a_intro.html#intro