Observation of left and right entropy in Voynich MS

Akinori Ito

2002/12/9

1 Introduction

This paper describes a few attempts to estimate characteristics of ‘words’ used in the Voynich Manuscript[1]. The Voynich Manuscript (‘VMS’ hereafter) is an unciphered book with mysterious script and many figures of unknown plants. In this paper, I concentrated into estimation of statistical property of ‘word’ in VMS. On analyzing the sequence of ‘word,’ I made some assumptions:

- VMS is written in some kind of natural language[1].
- Space-separated strings of Voynich character are (roughly) correspond to ‘word’ in ordinary natural languages. Note that the role of ‘word’ differs from language to language. For example, as Germany has compound word system, variation of words in Germany is much larger than English. This assumption simply excludes (as a working hypothesis) so-called ‘dain-daiin hypothesis[3].’

Under these assumptions, I made a word-based analysis on VMS.

2 Context analysis using left entropy and right entropy

The basic idea of the analysis is to investigate the variation of word at the left (or right) of a certain word. For example, an English word ‘am,’ when used as a verb, always comes after ‘I.’ Therefore variation of words comes at the left of ‘am’ will be very limited. Following that idea, I define *left entropy*H_L and *right entropy* H_R. Let $< w, w' >$ be a part of word sequences and w' appears just after w. Now,

$$H_R(w) = \sum_{w'} P(< w, w' > |w) \log_2 P(< w, w' > |w) \tag{1}$$

$$H_L(w') = \sum_{w} P(< w, w' > |w') \log_2 P(< w, w' > |w') \tag{2}$$

If a word w is used in limited left-context, it has smaller $H_L(w)$ than $H_R(w)$. The relationship between $H_L(w)$ and $H_R(w)$ shows a kind of deviation of the context the word w appears.

3 Word-based Experiment

3.1 Experimental conditions

I carried out an experiment to measure H_L and H_R for high frequency words in VMS. Used data are listed in table 1. I use the VMS script transcribed by Takahashi[2].

3.2 Result for English

Results for the Book of Genesis is shown in figure1. In this figure, words appear more than 50 times are plotted at (H_L, H_R) point. From this result, a couple of property of H_L and H_R can be observed. First, a frequent word has higher H_L and H_R. In fact, ‘the’ and ‘and’ are two of the most frequent words. Next, most words have similar left and right entropies. However, a few words have differences between the two entropies. For example, ‘am’ has small left entropy as explained above. As the word ’s’ always comes after apostrophe, it has left entropy of 1.0. Words begin with capital letter (‘When’, ‘He’, ‘Now’,...) also have smaller left entropy because it tend to appear at the beginning of a sentence. On the other hand, verbs (‘said’, ‘come’,...) and unit (‘years’) have smaller right entropy.

[1] In my feeling VMS looks like a kind of conlang. I believe the following discussion is valid for both natual and constructed languages.

Table 1: Corpora for the experiment.

Corpus	# of word	# of character	# of distinct words
Voynich MS	37973	198859	8495
Book of Genesis	41834	150449	3150

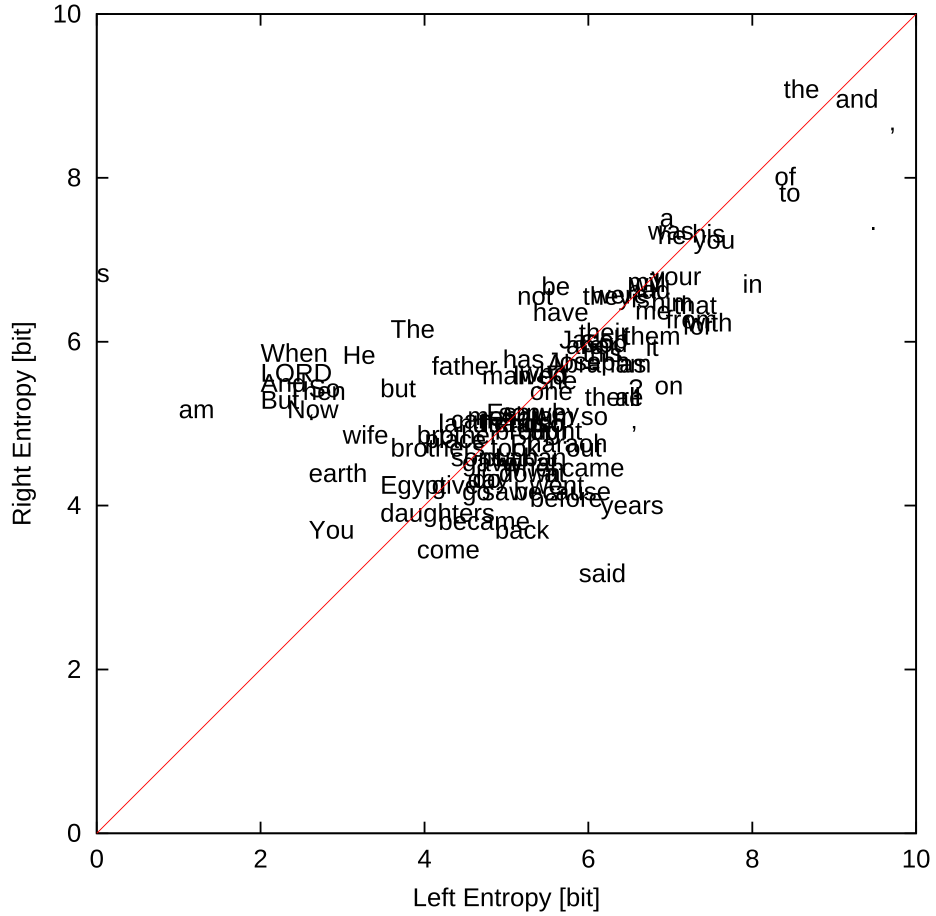


Figure 1: H_L and H_R distribution for Genesis.

3.3 Result for Voynich MS

Now, the result for VMS is shown in figure 2. In this figure, words appear more than 50 times are plotted at (H_L, H_R) point. Compared to figure 1, the distibution looks very different. Almost all words have almost same left and right entropy. It looks that many words have smaller left entropy, but it is not true because each word is ploted for its leftmost point to be (H_L, H_R). This result suggests that words in VMS are quite context-independent. Figure 3 is detail of figure 2. I can find nothing meaningful from this result.

Next, according to the experimental result, I investigated the relationship between the word frequency N and H_L. According to the information theory, when a word occurs N times, its entropy has upper limit of $\bar{H}$ where

$$\bar{H} = \log_2 N \tag{3}$$

Then I plotted word frequency and H_L to compare its theoretical upper limit. The results are shown in figure 4 and 5. The solid lines show the upper limit. From these results, it is found that H_L of Voynich words are very close to its upper limit. This result means that most words, except very high frequency words, meet different word on its left side each time it appears. It looks quite abnormal for me,

There seem to be several possibilities to explain the result. One might say VMS word order is completely

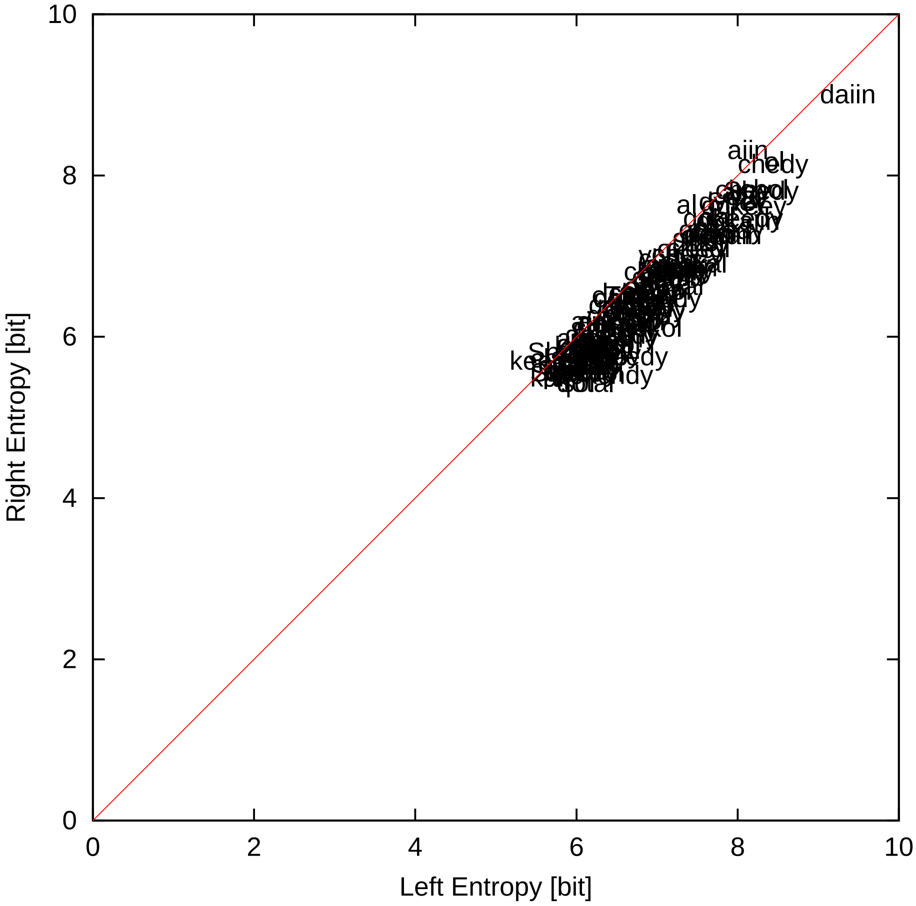


Figure 2: H_L and H_R distribution for VMS.

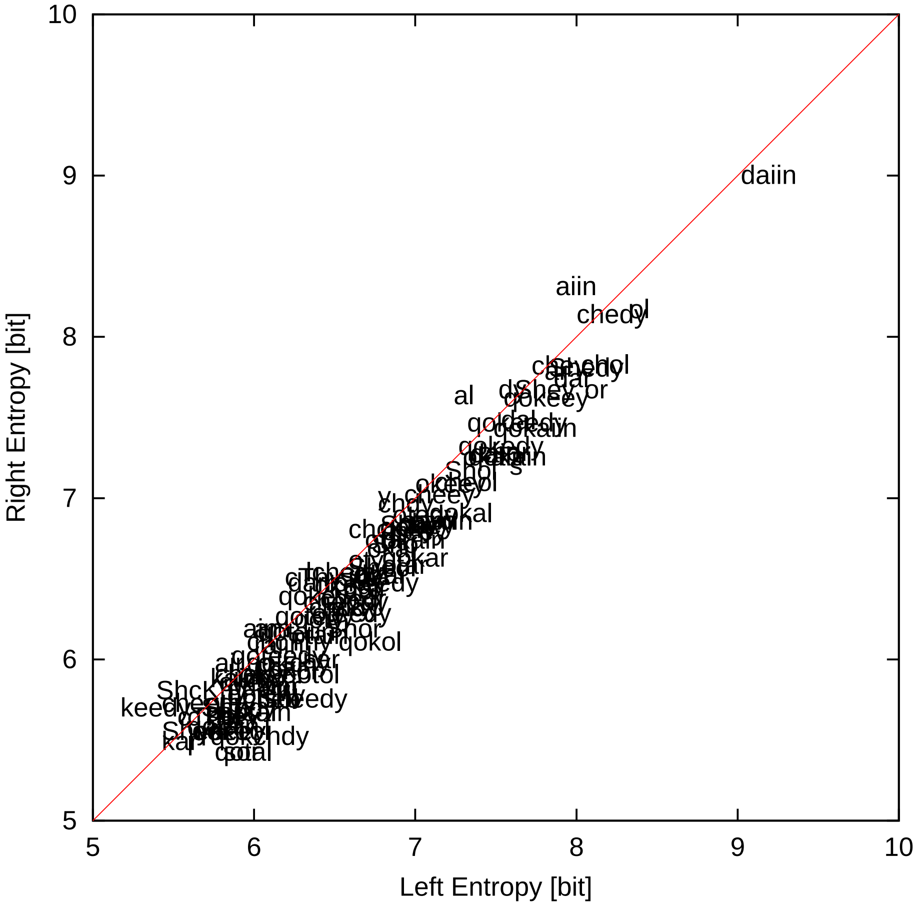


Figure 3: H_L and H_R distribution for VMS (zoomed).

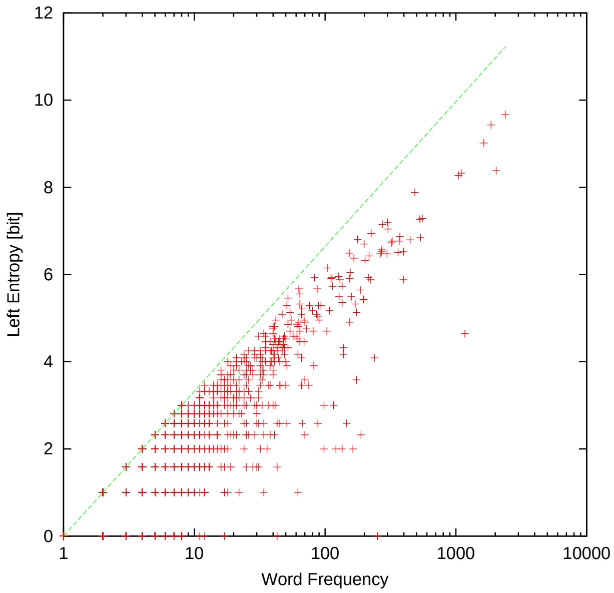


Figure 4: Word frequency vs. H_L for Genesis.

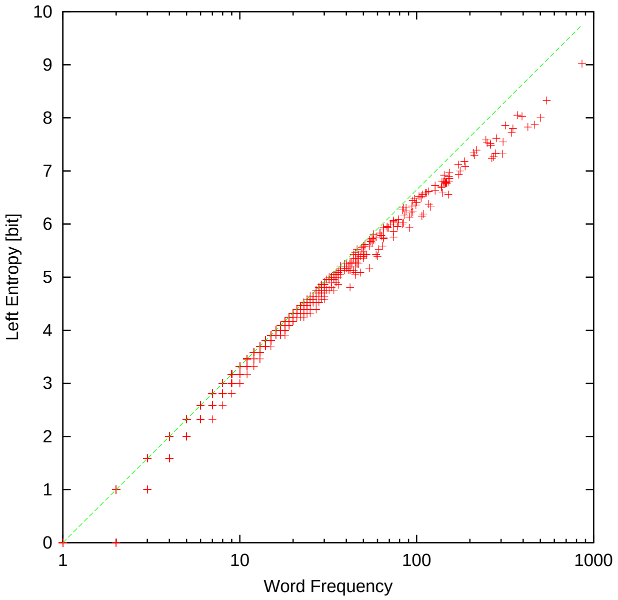


Figure 5: Word frequency vs. H_L for VMS.

Table 2: 2-character context of qokchdy

left context		right context	
dy	35	qo	17
in	4	ot	8
hy	4	ch	6
ey	4	da	3
ar	2	ol	3
l	1	op	3
y	1	ka	2
or	1	ok	2
al	1	dy	2
ol	1	s	2
ed	1	or	1
sd	1	ra	1
		r	1
		of	1
		lk	1
		dc	1
		do	1
		Sh	1

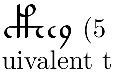
Table 3: 2-character context of Sheody

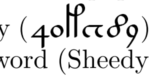
left context		right context	
in	9	qo	18
ol	8	Sh	8
dy	6	ch	7
ar	5	ol	2
al	4	pc	2
ey	3	ok	2
os	2	yk	2
ty	2	to	1
ko	1	yt	1
ky	1	ro	1
ho	1	lf	1
or	1	or	1
ir	1	sa	1
ay	1	ai	1
od	1	po	1
op	1	op	1
eo	1		

meaningless, but a couple of character-based research tells us that VMS is very similar to a kind of natural language[4]. It seems quite impossible for me to make nonsense word strings in 13∼16th century whose character-based statistics is similar to natural language. Another possibility is that a Voynich word is not a really word, but a kind of phrase. This hypothesis explains the fact that VMS has much larger number of distinct words than ordinary documents(see table 1). Word-internal grammar[5] might give us a hint about it.

4 Prefix/suffix based approach

As word-based context analysis gave no useful information, I carried out character-based analysis. The basic approach is the same as the above-mensioned one, but left and right entropies are calculated using suffix and prefix of the words at each side respectively. For example, let's think of the following context

("cKhey kodaiin cPhy" in EVA). When two letters for suffix/prefix are considered, the word has left context and right context . Then H_L and H_R are calculated using these context letters. Let the number of letters considered be k. In this experiment, I didn't consider contexts which exceed word boundary. For example, when $k = 5$ at the above example, the left context of is  (5 letters in EVA) while the right context is (4 letters in EVA). Therefore, using large number of k is equivalent to the word-based context analysis.

Correlation coefficient between H_L and H_R as a function of k is investigated for Genesis and VMS. The result is shown in figure 6. It is obvious that the correlation coefficient is smaller for small k, and it satulates for larger k. Distributions of each word in (H_L, H_R) plane for $k = 1, 2$ are shown in figure 7 and 8. These results tells us several interesting points. First, most words have smaller H_L than H_R, that means the variation of suffix is smaller than that of prefix. This may related to the word-internal grammar of Voynich language[5]. Next, qo-prefixed words (qokchedy, qotal,...) have smaller H_L. For example, 2-character context of the word qokchdy () is shown in table2. From this table, it is found that a qokchdy tends to follow -dy. Third, Sh-prefixed word (Sheedy, Sheody,...) have smaller H_R. 2-character context of the word Sheody ( ) is shown in table3. From this result, it is found that Sheody tends to precede qo-prefixed word.

References

[1] M. I. D'Imperio: "The Voynich Manuscript: An Elegant Enigma", Aegean Park Press (1978)

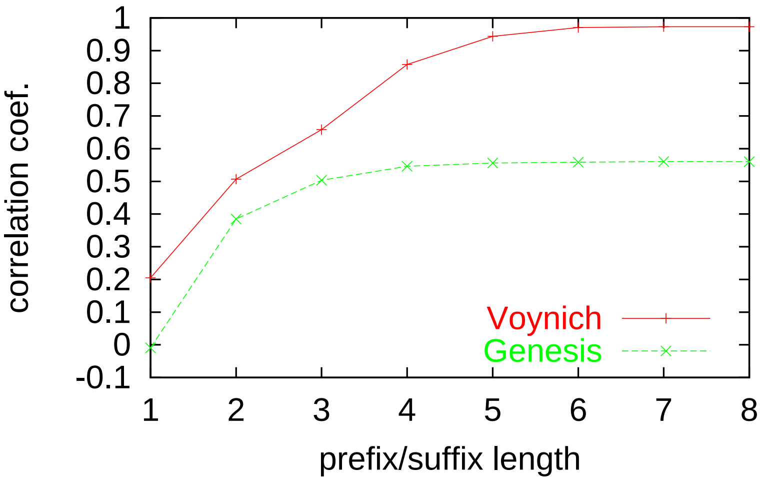


Figure 6: prefix/suffix length vs. correlation between H_L and H_R.

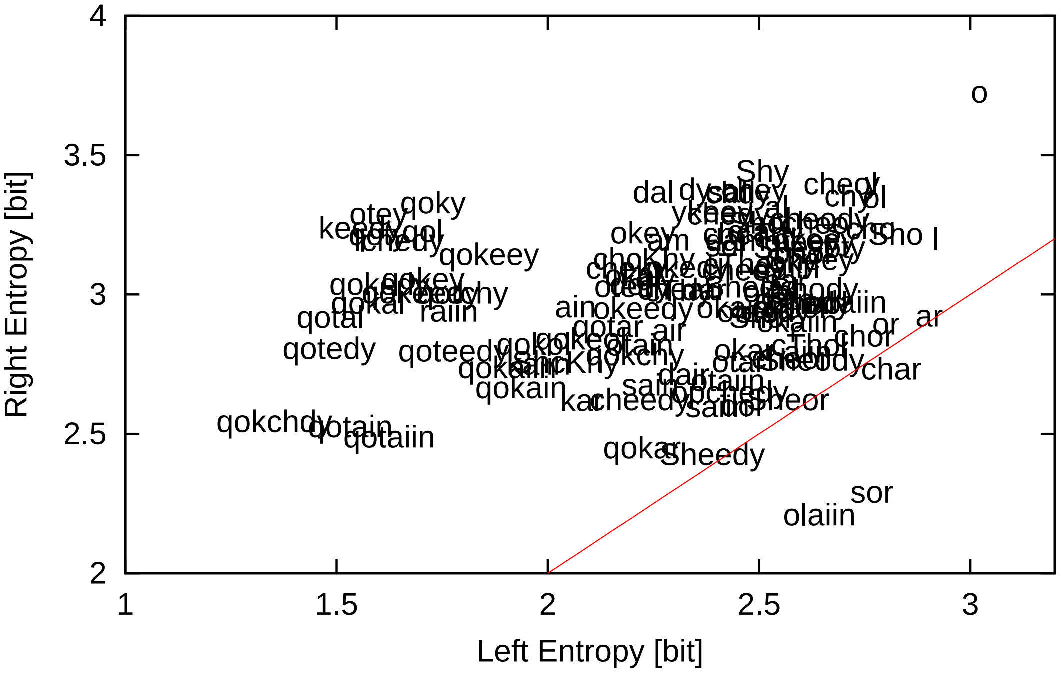


Figure 7: H_L and H_R distribution for VMS (1 character suffix/prefix).

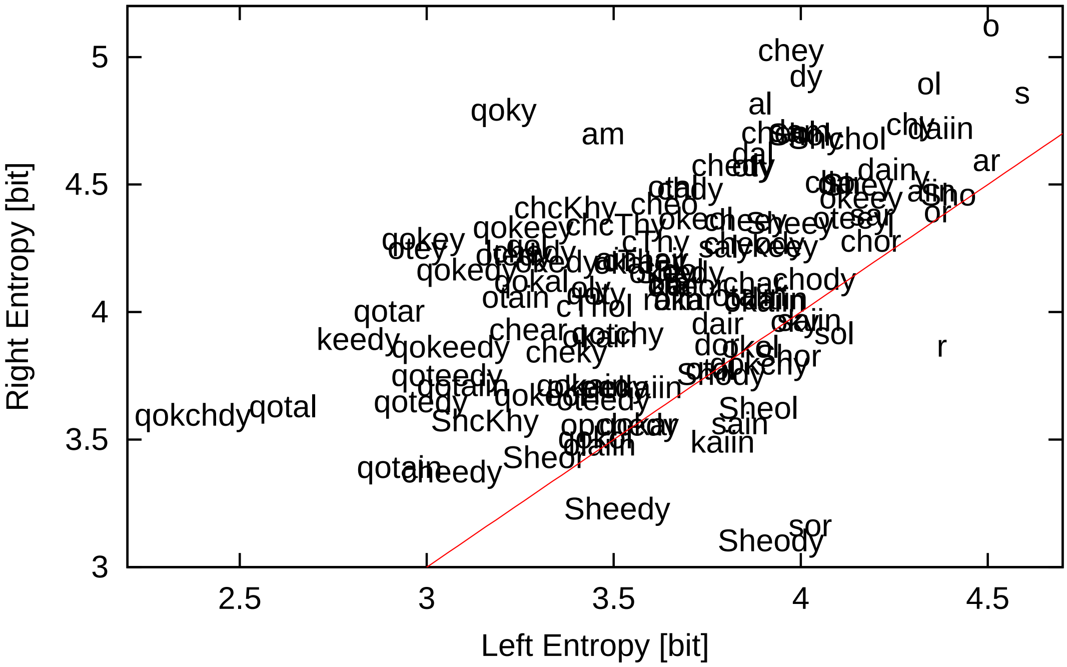


Figure 8: H_L and H_R distribution for VMS (2 character suffix/prefix).

[2] T. Takahashi: `http://www.voynich.com/`

[3] G. Landini: “The ”dain daiin” hypothesis”, `http://web.bham.ac.uk/G.Landini/evmt/daindaiin.htm`

[4] Mark Perakh: `http://www.nctimes.net/~mark/Texts/`

[5] Jorge Stolfi: “A Grammar for Voynichese Words”,
`http://www.dcc.unicamp.br/~stolfi/voynich/00-06-07-word-grammar/`