Effects of line context and prefix qo upon contextual deviation of Voynich 'words'

Akinori Ito

2002/12/17

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 my previous report[2], I found that Voynich 'words' have quite context-independent property, and it looks as if they were meaningless word sequence. In this paper, I carried out further investigation, mainly on the effect of line context and prefix morphemes. The methodology of the investigation is almost same as the previous report[2], but the next two processes were applied:

- paragraph tags (,) and line tags (<l> and </l>) are given to the text.
- The most popular prefix morpheme[4], o ('o' in EVA) and qo ('qo' in EVA) are split from the rest of the word, and the split fragments are treated as individual 'words'. For example, the word qokaiin is split into qo- kaiin.

The reason why I chose only o and qo is that these suffixes seem to play important role on context-dependency of Voynich words. In [2], qo-prefixed words is shown to have distinct context-dependency.

2 Context analysis using left entropy and right entropy

The tool of the analysis is same as [2]: *left entropy*$H_L(w)$ and *right entropy* $H_R(w)$. They indicate the variation of words that come left/right of the word w. H_L and H_R are very rough index of context-dependency of the word.

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 Analysis

3.1 Effect of line

First, the effect of 'line' upon the contextual deviation was investigated. The experiment was carried out by inserting (beginning of paragraph, or BOP), (end of paragraph, EOP), <l> (beginning of line, BOL) and </l> (end of line, EOL) tags into the voynich text corpus. Table 1 shows the corpus. This corpus is transcribed by Takahashi[3]. I did the following two experiments:

- Calculate H_L and H_R for the text with paragraph tags.
- Calculate H_L and H_R for the text with paragraph and line tags.

Table 1: Used corpus

number of paragraphs	238
number of lines	3906
number of words	33165
number of characters	167860
number of distinct words	6917

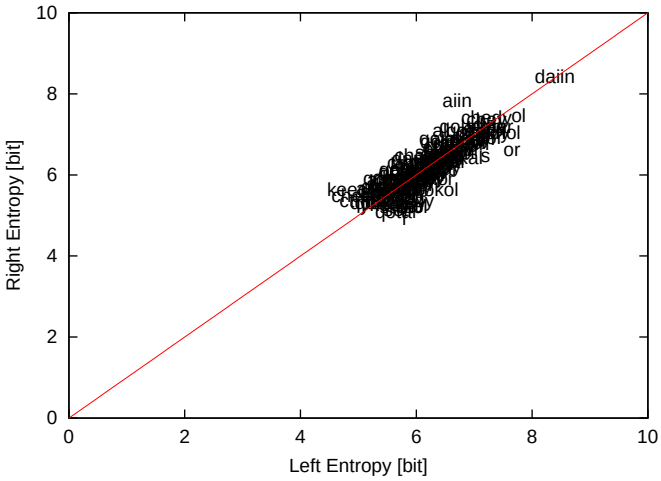

Figure 1: L&R Entropy distribution for paragraph-tagged text

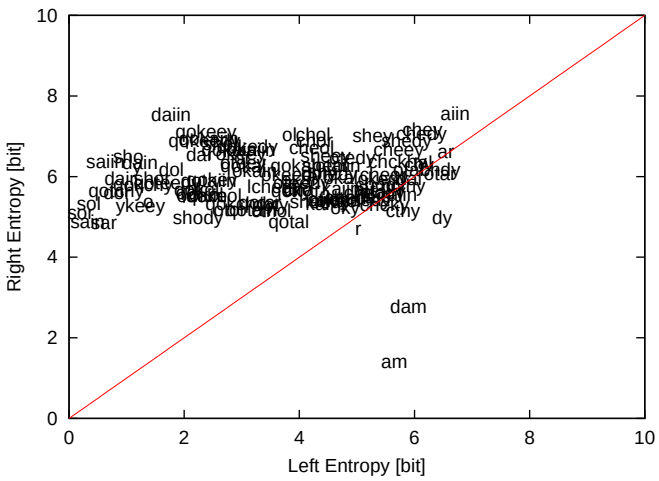

Figure 2: L&R Entropy distribution for line-tagged text

The results are shown in figure 1 and 2. In these figures, such words that occurs more than 50 times are plotted. As for paragraph-tagged text, the result was almost same as that from non-tagged text[2]. However, the result from line-tagged text was very different from the previous one.

From figure 2, sereral phenomena are observed:

- H_R of most words don't change very much, while H_L of most words reduced much. It means that many high-frequency words tend to appear at the beginning of line.
- On the contrary, the end of line doesn't affect most of words, except ꞩaiin(dam) and aiin(am). These two words are strongly affected by EOL. In fact, 56 times out of 69 aiin appears at EOL, and 58 out of 91 ꞩaiin appears at EOL.

Do these phenomena occur merely by chance? To see this, I carried out another experiment using English version of the Book of Genesis. I wrapped the lines of Genesis so that they don't exceed 80 characters, and observed the (H_L, H_R) distributions from that text with/without line tags. The results are shown in figure 3 and 4. These two results are not very different, that suggests the phenomena observed in VMS is not a chance.

These observations on VMS strongly support Currier's claim[5] that line in VMS is a functional entity.

3.2 Effect of prefix o and qo

From the result previously reported[2], qo-prefixed words show left context dependency. From that result I got an intuition that this prefix might be a kind of preposition or article, attached to its next words. Then I carried out an experiment that treat prefix o and qo as individual morphemes. As I made no distinction between the long word like qokeedy and the short word like qol, these words are equally separated as qo- keedy and qo- l. With this separation, number of words in the corpus became 44957 and number of distinct words became 5842.

Is it sure that qo and o are detachable prefix? To prove this, I investigated if a prefix-removed word is in the original vocabulary or not (I call it 'reproductivity' of a prefixed word). Table 2 shows the result. Note that this

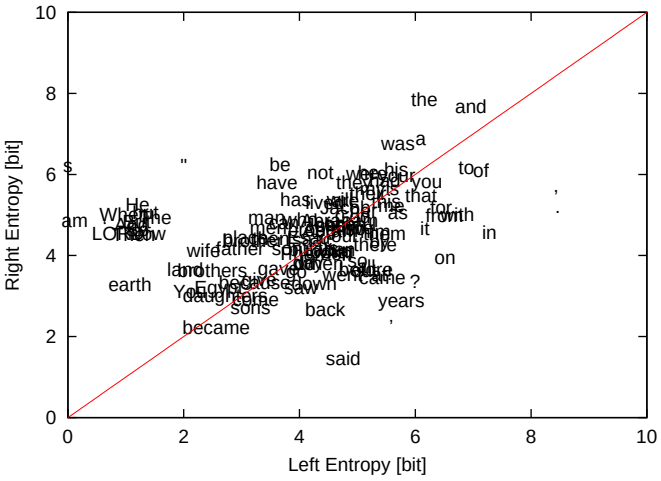

Figure 3: L&R Entropy distribution for Genesis, without line tag

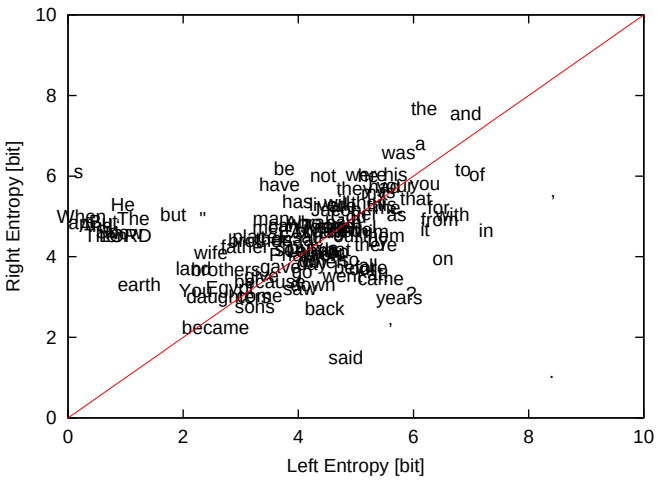

Figure 4: L&R Entropy distribution for Genesis, with line tag

Table 2: Reproductivity of prefix-removed words

prefix	reproduced	not reproduced
4o	57.5%	42.5%
o	47.3%	52.7%

result doesn't take the possibility into account that the prefix-removed word exists in the vocabulary prefixed by another prefix morpheme (for example,). The result shows that about half of the prefixed words are reproducable.

The result without line tag is shown in figure 5, and that with line tag is shown in figure 6. Words occur more than 50 times are plotted in these figures. The results were big surprise to me, because the words are split into two groups very clearly. One group has smaller H_L, while the others has larger H_L and a little bit smaller H_R. Insertion of line tag makes a few words ((ly), (tam), (kam), etc.) split from the two groups, but the fundamental structure of the groups doesn't seem to be changed.

I named these two groups 'red' and 'green.' There is no reason why I call them with color name, but it is convenient to display them on screen and remember. Moreover, as color names don't have grammatical meanings, a misunderstanding to associate its name with grammatical function will be avoided. Now, I named those words with smaller H_L 'red', and others 'green'. Several words and prefixes might be better treating separately.

3.3 Word group distribution and 'language'

I noticed that certain pages contain much larger 'red' words than other pages. So I investigated the relationship between frequencies of 'red' and 'green' words and Voynich 'languages' discovered by Currier[5]. I observed relative frequency of red words, green words and others (low frequency words) for each page. Figure 7 shows the 2-D plot of each page with respect to relative frequency of red and green words. The strings 'HA', 'BB' etc. denote the section and Currier language of each page. For example, 'HA' means 'Harbal section, language A', 'BB' for 'Biological section, language B' etc. From the result, it is found that relative frequency of red word is strongly related with language. On the other hand, the relative frequency of green words seem to have nothing to do with its language.

4 Conclusion

I investigated the effect of 'line' and 'prefix' on context-dependency of Voynich words. The experiment of 'line' revealed the following facts:

- BOL affects many words, i.e. many words tend to appear at BOL.

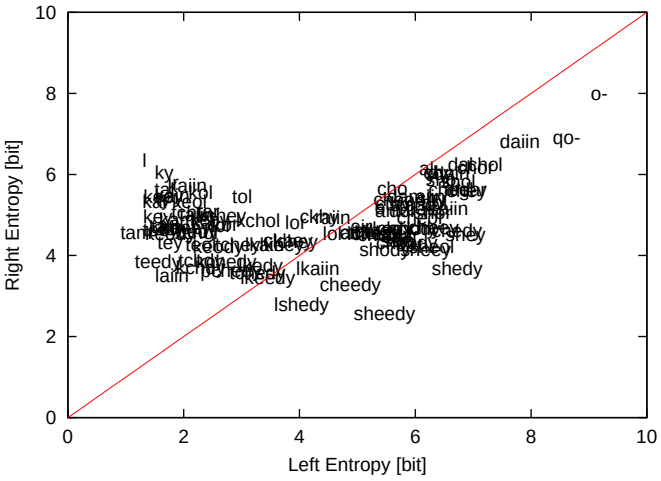

Figure 5: L&R Entropy distribution for o/qo- separated data (paragraph-tagged text)

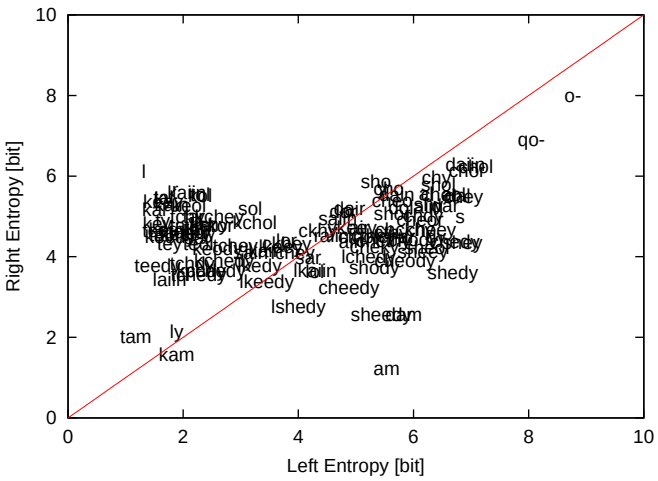

Figure 6: L&R Entropy distribution for o/qo- separated data (line-tagged text)

- Limited number of words (𝔬𝔦𝔣, 𝔔𝔬𝔦𝔣) tend to appear at EOL.

The next experiment showed that prefix-removed words (or word stems) are of two distinct groups. Then it is found that one of the two groups appears frequently in language B.

It is very interesting for me that my experiments validated two of Currier's discoveries: 'line as functional unit' and 'existence of two languages'. I didn't intend to validate Currier's discoveries but tried to find grammatical (in statistical meaning) characteristic of specific words. I think these results shows that Currier's two discoveries are true.

References

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

[2] Akinori Ito: *Observation of left and right entropy in Voynich MS*, `http://www.geocities.co.jp/Technopolis/7220/voy/paper021209.pdf`

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

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

[5] P. Currier: *Papers on the Voynich Manuscript* `http://www.dcc.unicamp.br/~stolfi/voynich/mirror/reeds/currier.html`

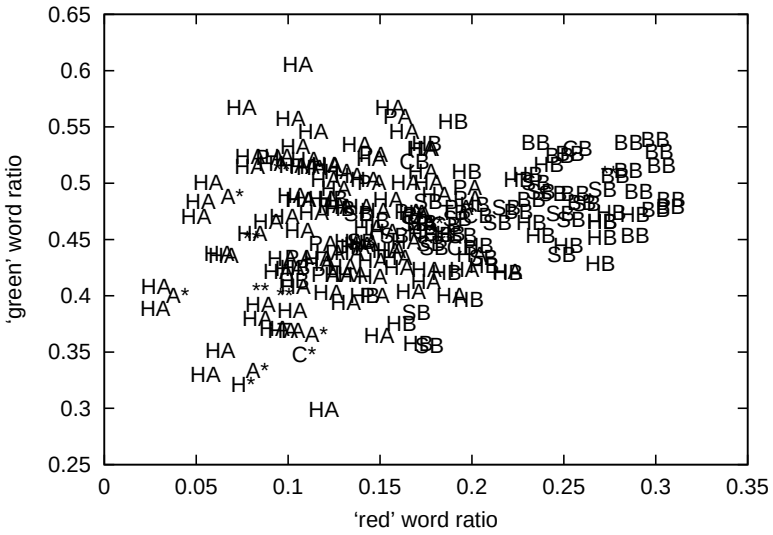

Figure 7: Distribution of each folio based on the ratio of 'red' and 'green' words