The Line as a Functional Unit in the Voynich Manuscript: Some Statistical Observations

Elmar Vogt, Fürth*

November 27, 2012

This paper has a closer look at several of the attributed properties in the Voynich Manuscript (»VM«), especially regarding the changes of average word length depending on the position of a word in its manuscript line. We[1] examine the actual word-length distributions and give examples how these could appear. Especially we point out that these effects could be the result of simple handwriting, and not be an artefact of any enciphering method which may process information in line-sized chunks.

We put our ideas to the test by checking whether the same effects can be observed in »natural« text which has not been enciphered or otherwise manipulated.

1 Preliminaria

1.1 Claims

Over time, a certain amount of »lore« has developed around the Voynich Manuscript to complement the lamentable dearth of hard facts in connection with this booklet. This »lore« are findings which have once been observed, but perhaps never fully properly explored, and are thus a questionable basis for further arguments. For example, supposedly the VM was written with hardly any corrections in it, which may or may not hint at particular enciphering schemes. Also, every now and then there is the announcement that the VM was written from an alphabet with 17 characters.[2]

*http://voynichthoughts.wordpress.com

[1] Yes, I'm writing this on my own, but talking about »us« is also academic tradition and give me a twinkle of cheap self-aggrandization.

[2] This of course is a half-truth, since while it is true that a large amount of the volume of the VM can be written with such a small number of different letters, there is a considerable amount of rarer, but still recognizable glyphs other than that. Also, there is no universal agreement about the fundamental character set of the VM. EVA <ch> could be a single letter, or a ligature of <ee>, or something completely different.

One of these often-touted claims is the fact that »the line is a functional unit« for the encipherment of the VM. This is usually based on a number of observations:

- Several characters, especially the gallows, tend to occur line- and paragraph-initial,
- Word length varies considerably over the lines of the VM.

We would like to examine especially claims of the second group in this paper.

1.2 Selected Materials

1.2.1 Sources

As material for the analysis, we use different transcriptions of the VM, namely those provided by Currier, Takahashi, and – to a lesser extent – Stolfi. The transcriptions were collated by Elias Schwerdtfeger and downloaded through his *Voynich Information Browser*[3].

Please note that Takahashi and Currier offer the most complete transcriptions with 68,000 letters in hand »A« and 145,000 letters in hand »B« for Takahashi, and 45,000 and 64,000 letters for Currier »A« and »B«, resp., while Stolfi's transcription is smaller and thus statistically less significant.[4]

1.2.2 Text Fragments Used

Not all text available was used for the calculations, but we restricted ourselves to lines of at least 3 and at most 12 words (with the exception of figure refall). Shorter lines are mostly labels and in any case can't shed much light on the distribution of word length along the line. Longer lines usually are found on fold-out pages and not in »regular« text; the behaviour they exhibit may or may not be representative of the general enciphering method. Furthermore, such long lines are also fairly rare, and thus not well suited to statistical analyses.

1.2.3 Transcription

The transcriptions were encoded in »Basic EVA«, and all subsequent character counts were done based on this transcription scheme. Doing statistics with other transcriptions may result in different actual numbers, but the general behaviour should be the same, since the differences are systematic.[5]

[3] http://voynich.freie-literatur.de/index.php?show=extractor

[4] Beside compiling a fairly complete transcription of the VM, Currier was also the first to note and designate the hands »A« and »B« in the VM; a classification which hasn't seriously been challenged to my knowledge. Thus, when we talk about »Takahashi hand A«, this is to mean »the Currier hand A in Takahashi's transcription«, rather than any classification invented by Takahashi.

[5] For example, if one transcription treated <iin> as a single character, this would result in a generally reduced word lengths compared to EVA, but since this reduction would affect *all* word positions – presumably in the same way – this would not invalidate the general trends.

1.3 Notation

We use l to denote the length of a word, and $\bar{l}$ for its »expectation value«, or average. All lengths are given in numbers of EVA characters, no physical measurement is implied.

Indices i and k denote word numbers on a line, and line numbers, respectively. (Thus, l_i would be the length of the ith word on a line.)

The length of a line is denoted n, and measured in words. This does not mean the actual physical length of the line, or the amount of space which is available, but the actual number of words which were written on this line. Accordingly, l_n is the length of the last word on a line, n^k is the true number of words on line k.

2 Theory

»The line is a functional unit« makes sense only if we assume a »traditional« way of enciphering the VM. This is opposed to, for example, David Suter's suggestion that the VM is actually a graphical encoding of landscape features or maps.

While rarely used in practice, there are actual enciphering methods which take their plaintext information line-by-line, manipulate it, and spit out packets of ciphertext the size of lines. To my knowledge, there exists no contemporary line-based scheme that had actually been in use and statisfies the other statistical features of the VM, especially the »grammar« which governs word composition, nor has in modern times a system been suggested which would result in the structure of the VM text as it is observed.

Besides, while such a line-based approach might serve well with a document which consists mostly of rectangular blocks of plaintext, the VM exhibits pages which are often strongly segmented, and where text is interrupted and must flow around »obstacles« like plants, nymphs and other objects. Any scheme which would try to, for example, encipher a 60-character plaintext line could *not* rely on there being actually the same amount of space on the ciphertext vellum to accomodate the ciphertext. This would be a nuisance at the least, and a killing blow to the technique in a worst case scenario when the deciphering depended on a certain line length of ciphertext.

Note that there are no indications that the VM scribe was forced to put a certain amount of information into a certain amout of space: There are hardly any characters »crammed« into an amount of space; rather, the document appears as if the scribe was free to set line breaks where they came naturally.

Note also that the VM shows now signs of »word-breaking« at the end of lines. In general both line-terminal and line-initial words obey the usual grammar rules[6], despite some changed frequencies. This indicates that whenever the scribe ran out of space on his line, he simply started a new line with the next word, rather than writing the first part of the next word on the old line, and the rest of it on the subsequent line.

[6] Whatever they are ...

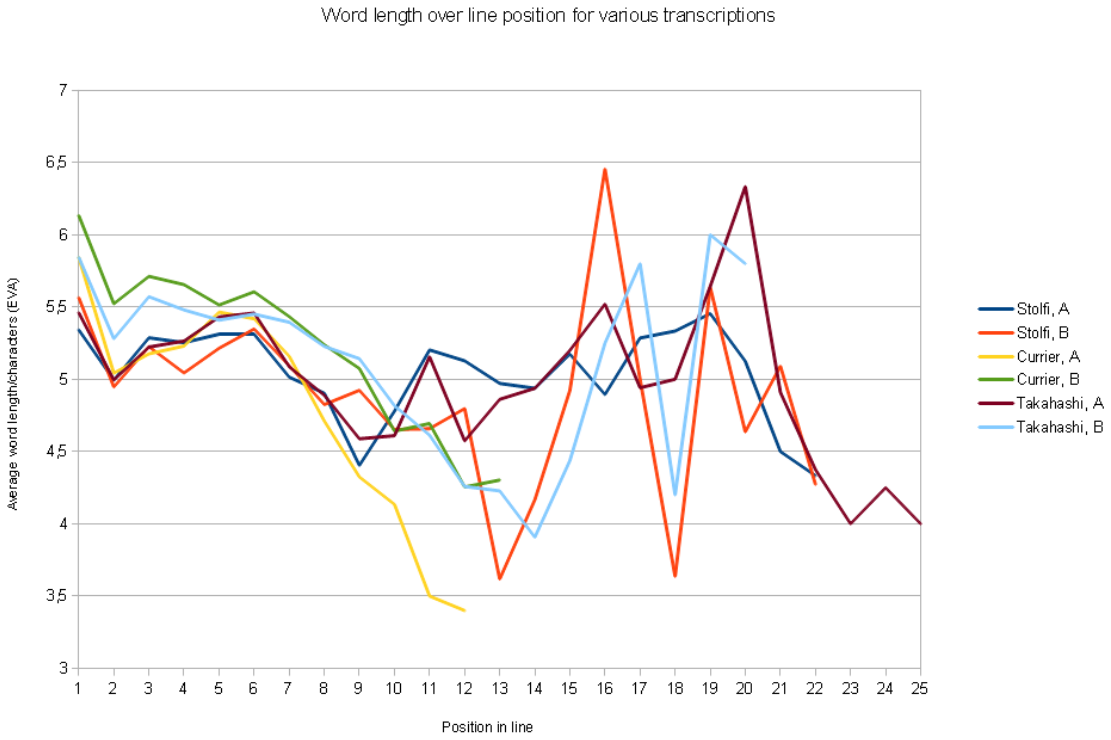

Figure 1: Overview over the average wordlength $\bar{l}_i$ according to the word's position i on the line for various transcriptions

3 A First Look

Figure 1 shows an overview over the wordlength distribution of words along the VM lines.

Each drawn line shows the average word length $\bar{l}_i$ for one of the transcriptions used, depending on the position i of the word on the current line, with $i = 1$ being the line-initial position. The transcriptions (Currier, Takahashi, Stolfi) were divided again for the »hands« »A« and »B«, resp.

Please note that this diagram can't show line-terminal effects, because the lines considered were of differing lengths, so while one line might have ended with the fourth word, another may have ended with the eighth, and so on.

Three effects are immediately obvious for virtually all distributions considered:

1. The first word with $i = 1$ of a line is *longer* than average, $\bar{l}_1 > \bar{l}$,
2. The second word with $i = 2$ is *shorter*, $\bar{l}_2 < \bar{l}$,
3. Over the course of the line, the average word length $\bar{l}_i$ drops: $\bar{l}_i > \bar{l}_{i+1}$.

It should be noted that the average word length begins to vary more wildly, the higher the word position i on the line is. This is only to be expected, since with growing

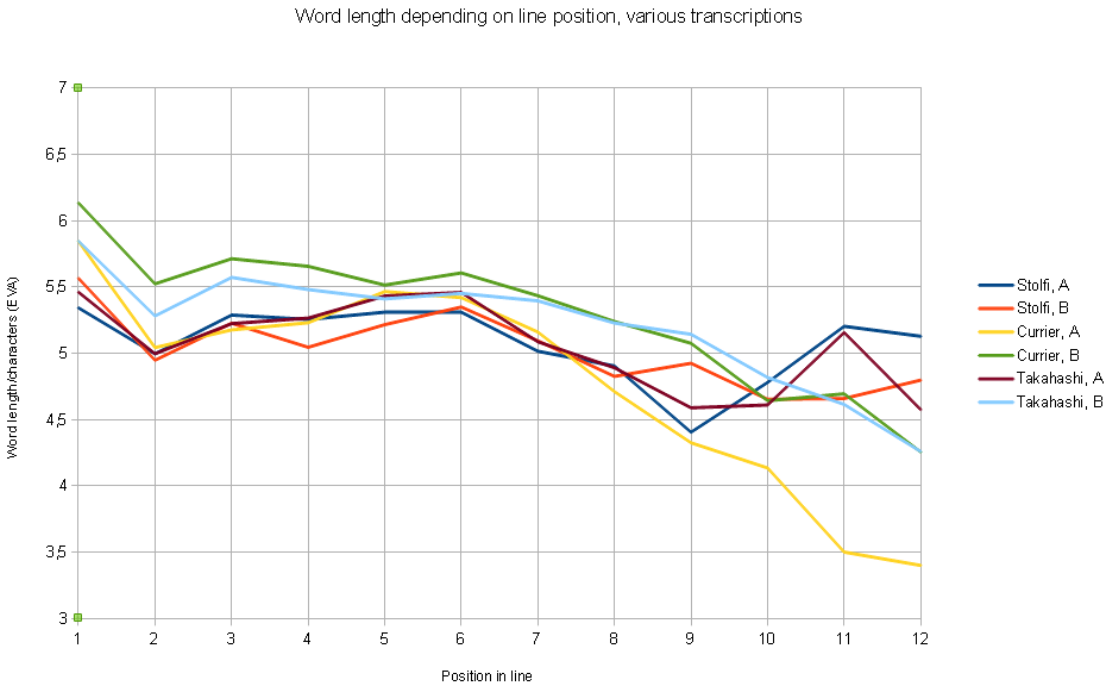

Figure 2: Same as Figure 1, but showing only positions up to word $i = 12$ on the line

i, the number of lines with sufficient length $n \geq i$ naturally shrinks. For example, the Currier transcription, hand »B« has 1154 lines of $n \geq 3$, but only 904 lines with $n \geq 7$, which constantly drops to 10 instances of 14-word lines or longer. To avoid the frantic wiggles due to the low number of small samples for very large lines, figure 2 shows the same distribution as before, but only the section with line lengths up to $n = 12$ words.

Overall, all transcriptions show the same behaviour, in accordance with the three above claims.

4 Is That So?, or: Examining Claim 3

Let us first examine the claim of the dropping average character length $\bar{l}_i$. A look at figure 2 shows that from the beginning to the end of the line, the average wordlength drops roughly between 0.5 and 1 characters ($\bar{l}_1 - \bar{l}_{12} \approx 0.7$), which is not drastic, yet significant.

But there is one problem with this observation, namely, it doesn't take into account the respective total line lengths.

Consider for a minute that word lengths can only be calculated as long as $i \leq n$. Now, for small values of i, almost all lines will statisfy this condition, and all text will be considered. But for growing values of i, only lines with large n are eligible. At the same time a line can only achieve a large value of n if many short words occur

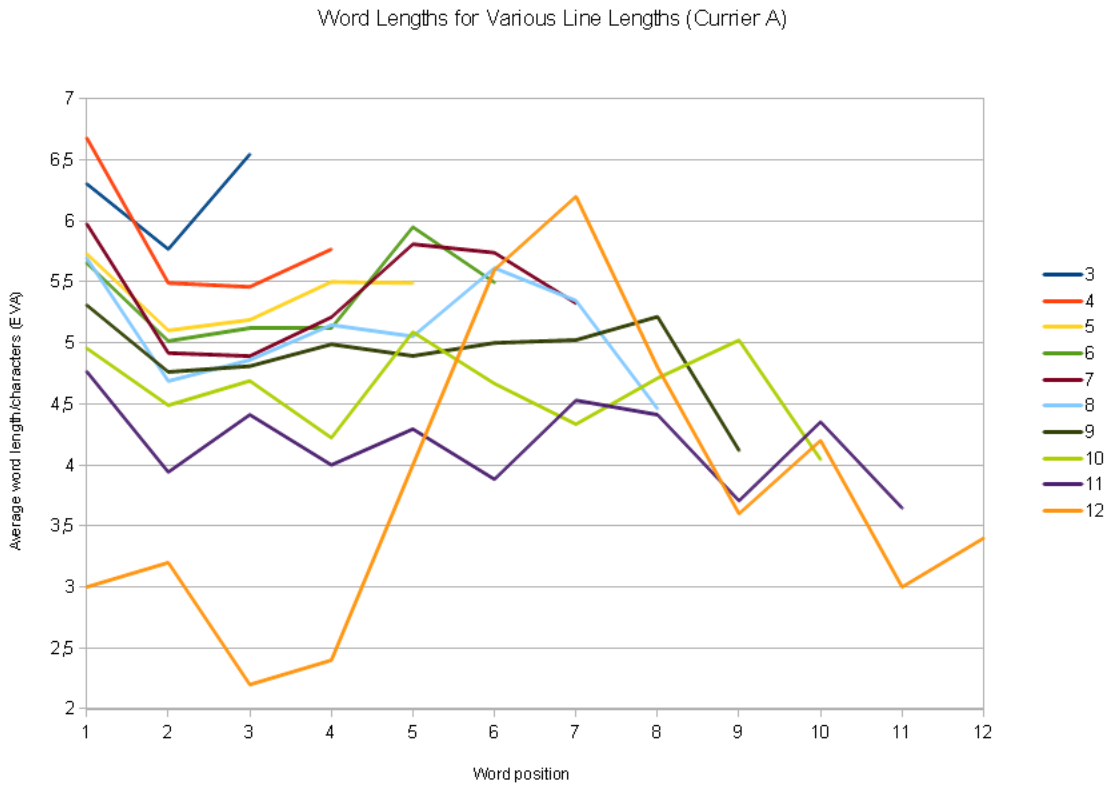

Figure 3: Average word length $\bar{l}_i$ plotted over the word position $i$on the line, for various total line lengths n (Currier transcription, hand »A«)

on it – otherwise the scribe would have had to word-wrap the line long before. The consequence of this for our naïve approach is that a fair mix of all words enters into the calculation for small values of i, while for large i a bias towards short words appears

This can be put to the test by splitting our data according to total line lengths n, and in figures 3 and 4, exactly this was done.

Every plotted line represents the course of the average word length $\bar{l}_i$ along the line for one particular line length of n words. For example, the light blue line labelled »8« represents the average word length distribution for all lines with a length of *exactly* $n = 8$. (This differs from the previous figures where all lines with $n \geq i$ were considered for each i.)

It is immediately obvious that with growing n, all average word lengths shrink, regardless of their position on the line. This can be explained by the fact discussed above, namely that short words will tend to increase the length n of the line they're found on.[7] Thus, it's not the position on the line which reduces the word length, but the small word length which only makes high line positions possible in the first place.

[7]Bear in mind that we use the word »line length« to refer to the total number of words on one line, not to its physical width in centimetres or such.

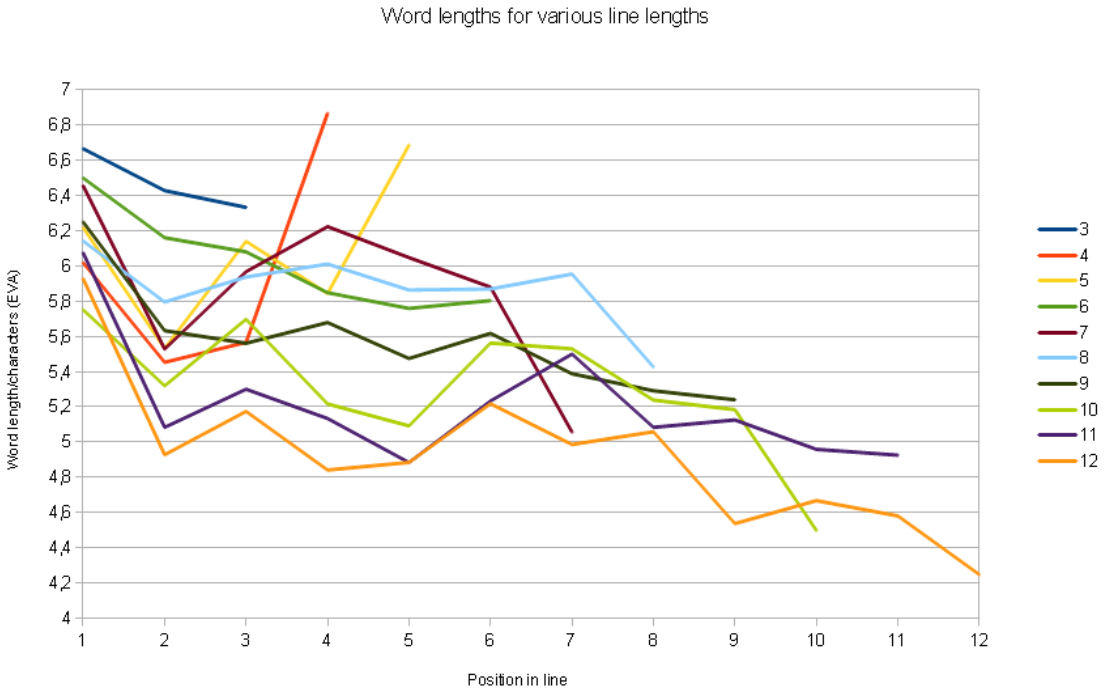

Figure 4: Same as 3, but for Currier transcription, hand »B«

Secondly, once the odd behaviour of the first two words on the line is disregarded, the rest of the data shows an almost constant $\bar{l}_i$ for each of the n'sconsidered. Note that for the last plot representing a line count of $n = 12$ words, there exist only 5 lines in hand »A«, which means that there are considerable staistical uncertainties, which is an excuse we can't put forth for the oddities of the very small line length: For lines with $n = 3 \ldots 5$, the last word l_n of each line seems to be significantly longer than would be expected, but there are 99 hand »A« lines with $n = 3$, and 196 for $n = 5$, so statistical fluctuations should cancel out. For longer lines ($n \geq 6$), this trend is reversed and the last word is actually *shorter* than average.

The second effect will be discussed in section 5. For the »upturn« of the first effect, we have no concrete explanation. It may be worth to point out though that the majority of these short lines occurs at paragraph ends – ie the line termination here is not »necessitated« by space running out, but results from properties of the text, for example a topic being exhausted.[8]

[8] One mechanism resulting in an increased length of the last word on such a paragraph-terminal line could be the inclusion of punctuation in the enciphered text. Assume for example that each ciphertext word represents one plaintext word, and that the plaintext word includes also punctuation like sentence-terminals periods, much like we are used to attach puntucation to the preceding word. Assuming paragraph endings are also sentence endings (which should not be unreasonable) this would mean that every paragraph-terminal word would on average be one character longer than »regular« words, namely the period.

5 The Last shall be the First: Examining Claim 1

The previous section has discussed how the effect of a drop in the average word length for large word counts i can be explained. There exists a similar explanation which ties two of the other effects together, namely the *increase* in the the length or line-initial words, and the *decrease* of line-terminal word lengths.

Consider for example a text being written with hitherto i words on line k, while the line leaves space for x more characters.[9]. Now, while the length of the ith word, l_i^k clearly will on average be equal to $\bar{l}$, what happens to the $i+1$st word? Obviously, there are two alternatives:

- For $l_{i+1}^k \leq x$, the word will fit on the remainder of the line, but
- For $l_{i+1}^k > x$, the word will have to wrap around to the next line, and it will end up as l_1^{k+1}.

But this amounts simply to the rule that *short words have a higher chance to remain the last words on a line, while long words have a higher probability to become the first word of the subsequent line*. The net effect of this would be just what is observed: That the last word of a line on average is shorter, but the first word is longer.

6 Putting it to the test

Now of course it's all good practice to go and put our results to the test. Our basic hypothesis is that the overall drop of the word length along the line is the result of the fact that only short words result in long lines, plus that wrapping words around the end of lines results in shorter line-terminal words, and longer line-initial words.

This can easily be put to the test, since all our arguments are equally valid for »natural« text, and we made no particular assumptions about the VM. Thus, the effects should be observed in the same way when applying the statistics to regular text.[10]

To this end, we used manuscripts in various languages,[11] namely

- Mark Twain: *Tom Sawyer* (english)
- Mark Twain: *Tom Sawyer* (german translation)
- Jules Verne: *20000 Lieues sous les mers* (french)

[9] suggesting without a loss of generality that all characters have the same width and require the same amount of space

[10] Here and subsequently, when there is mention of »natural« languages, then unenciphered plaintext in human language is meant. The word is used in contrast to the »processed« or »engineered« ciphertext of the VM. It is *not* to suggest that the VM was written in a constructed language, or is made up mostly of non-verbal information. We actually do believe that the plaintext for the VM was written in one of the contemporary languages of central Europe.

[11] as provided by http://www.gutenberg.org

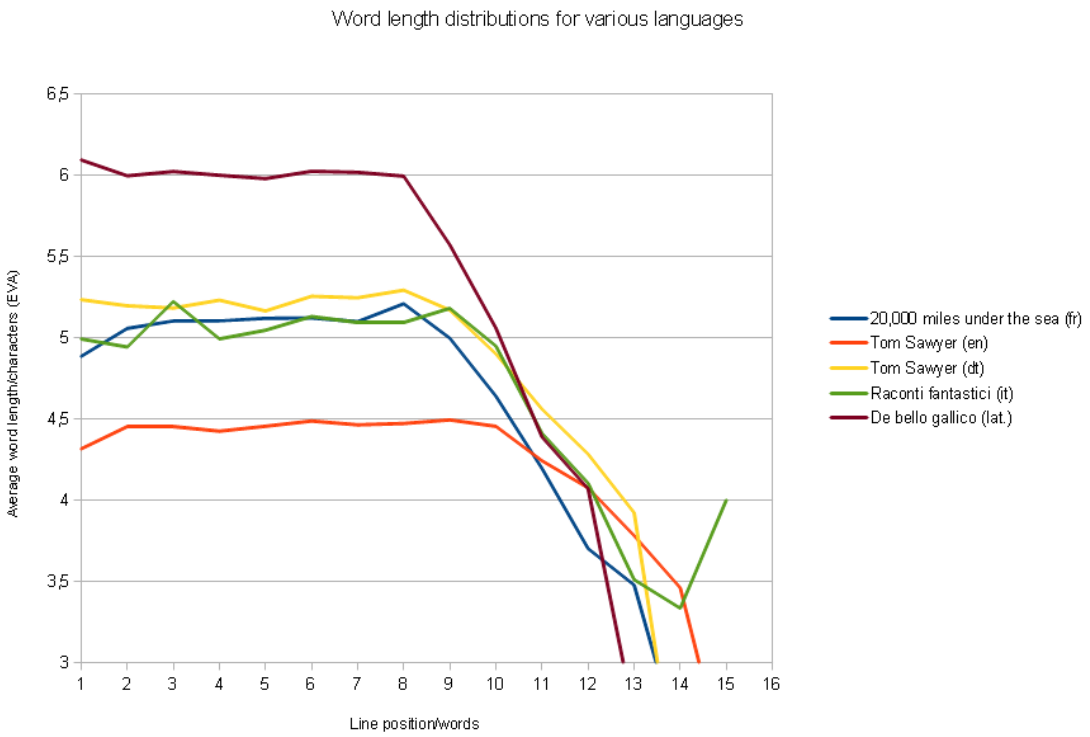

Figure 5: Word length statistics for various natural languages, calculated with the same method as before

- Iginio Ugo Tarchetti: *Raconti fantastici* (italian)
- Julius Caesar: *De bello gallico* (latin)

These had between 3500 (*De bello gallico*) and some 20,000 lines (*20,000 miles ...*), so the corpus is of roughly the same order of magnitude as the Voynich manuscript (Currier, hand »B«: 1500 lines). The ASCII files were reformatted to a (constant) line width of 62 characters, which is slightly lower than the line width of the VM, which fits up to 80 characters in a line.[12] The overall average word length of the various languages ranged between $\bar{l} = 6.0$ for latin and $\bar{l} = 4.5$ for english, while the VM words average around $\bar{l} = 5.8$.

6.1 Word length drop-off along the line

The results for the »constant drop off« are shown in figures 5 for an internal comparison of the languages, and 6 for a comparison of latin (which was still the best match) with Currier »A« and »B«.

[12] Though this is not a hard limit, due to the handwriting nature of the VM.

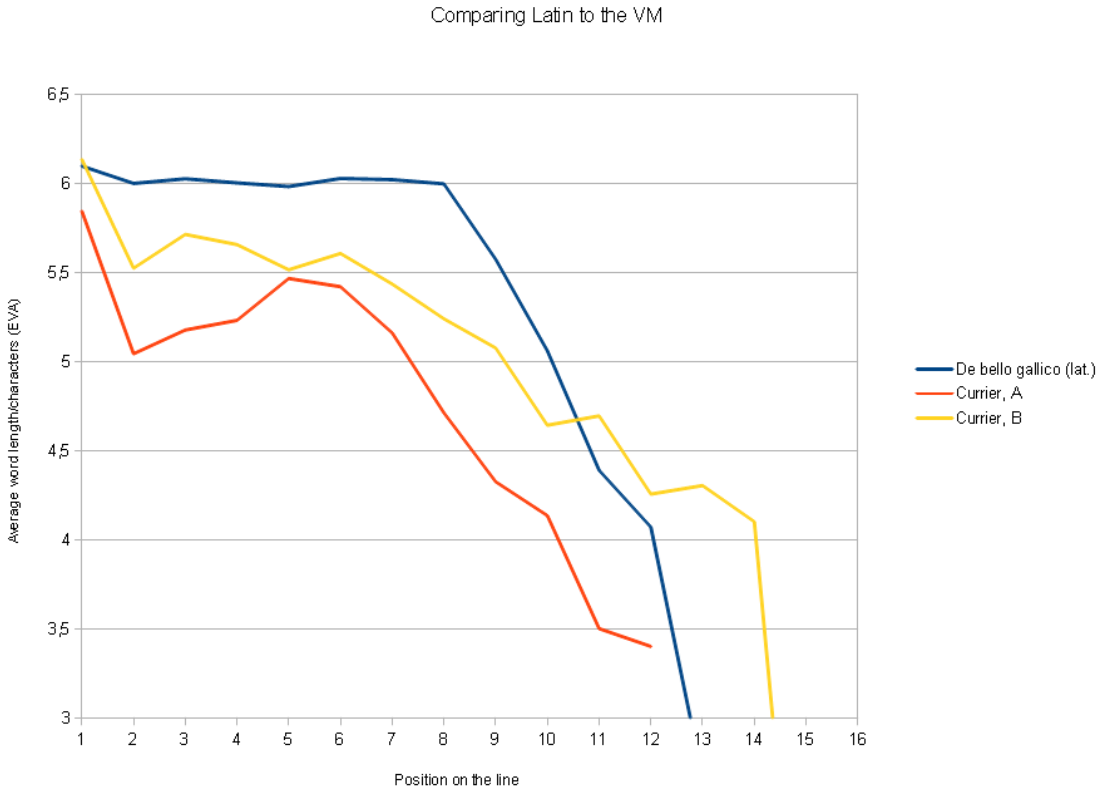

Figure 6: Comparison of exemplary latin word characteristics with the Currier transcription of hands »A« and »B«

If you compare figure 5 to 2, it is immediately obvious that, while in the VM the word length is steadily dropping along the line, in all natural languages the word length is almost constant up to a certain point, and then drops off sharply at a position around $n \approx 8 \ldots 10$. Figure 6 gives you a direct comparison between the behaviour of latin text, and that of the hand »B« in the Currier transcription.

This effect can be attributed to the (physical) line width of the natural language texts being constant throughout the MSs, while in the VM there had been many instances of the available space being narrower than the actual page width, cutting lines short on various positions. Figure 7 illustrates this. Both the latin text (*Bello gallico*) and the german text (the translation of *Tom Sawyer*) exhibit pretty narrow peaks in the distribution of line lengths around a maximum of $n = 8$ and $n = 9$ words, resp.[13], and drop off sharply on both sides. In contrast to this, the VM line length distribution is much wider, owing to the different page layouts which were avilable to the scribe. Taking this into account, it seems reasonable that the drop off of the word length $\bar{l}$ in the VM is really due to the »short words make long lines« effect mentioned earlier.

[13]Which is equivalent to the total line width divided by the overall average word length

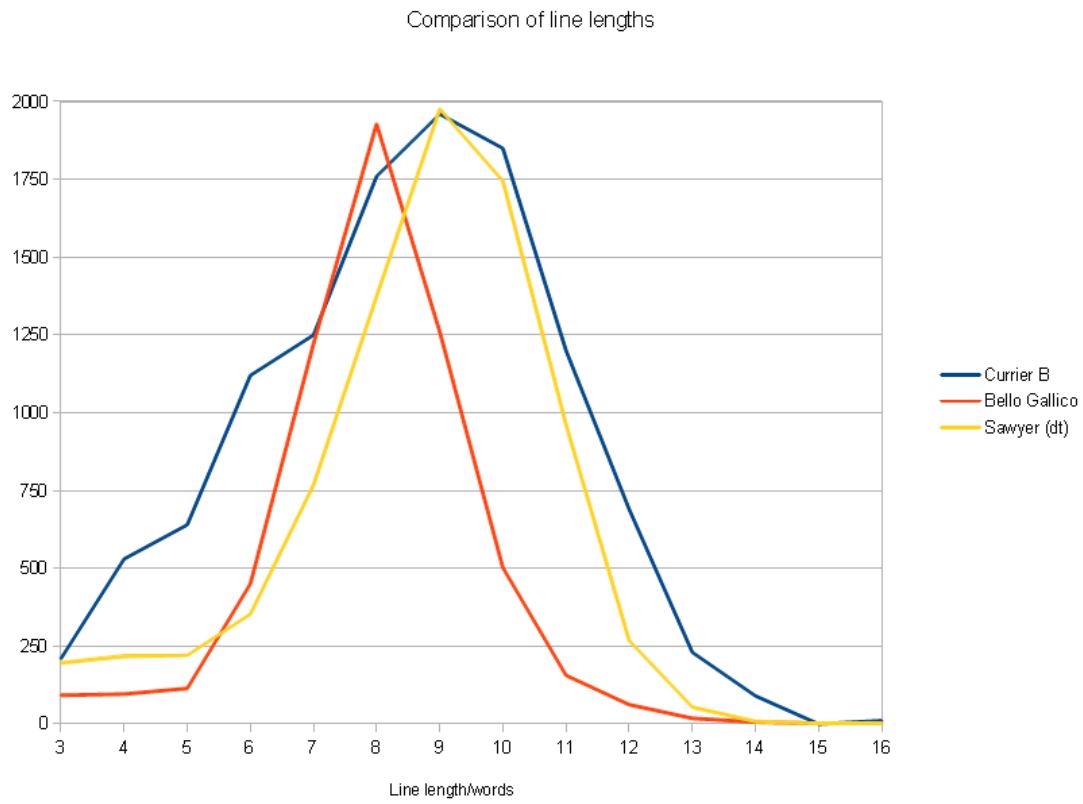

Figure 7: Comparing the number of lines with a certain amount of words on them between latin and german texts with constant page width, and with the VM. Note the much broader peak of the VM. (All curves have been normalized to have the maximum value around 2000 lines to make for a better comparison.)

6.2 Line-initial Word Length Peak, Dip at the Second Word

Fairly frustrating is the fact that the predicted »word wrap« effect simply refuses to take place: Latin, german and italian only exhibit a very small above-average word length for the inital word $\bar{l}_1$, while french and english actually show a *decrease* below average. Now, if one was in a really tolerant mood, it might be said that for example latin, *does* show a line-initial word length some 0.1 characters above average, and that if this was used as the source language for the VM, its encipherment may have (somehow ... magically ...) amplified this to the observed increase of the length of the first word to $\bar{l}_1 \approx \bar{l} + 0.6 \ldots 0.8$ above the general average. If one views it more realistically, then with two out of five sampled languages showing the actual *reverse* of the expected behaviour, and the other three only mildly fluctuating, the word wrap effect most probably simply isn't there.[14]

[14]Though it stumps me – pardon: »us« – as to why there is none.

The dip of the second word *below* the average word length $\bar{l}_2 < \bar{l}$, for which we had no statistical explanation in the first place, is only found in the VM and, very faintly, in the italian language sample: We suggest that this »italian dip« is only the consequence of random fluctuations, and that what ever caused the »second word dip«, doesn't occur in natural languages.

7 Summary

I guess what I'm experiencing is the most familiar feeling for VM researchers of them all, »I really don't know what to make of it.«

Regarding the three statistical effects (or »claims«) we have presented in section 1.1, we have come to the following conclusions:

- Although we could suggest a statistical mechanism for the »first word effect« (section 1.1, claim 1), we were mostly unable to corroborate this hypothesis with a test on natural language.
- The »second word effect« (claim 2) with a dip in the word length of the second word is undeniably there in the VM. We could neither provide an idea how this effect may have originated in the VM, nor could we find similar effects occuring in natural language texts.
- We think we have demonstrated how effect 3, the continuous drop of the average word length towards the end of a text line, occurs naturally as the result of text composition along lines, namely that short words will result in lines with more words, and thus higher word counts.

The scripts I wrote for the evaluation of the transcriptions, and the resulting data are available on request.

Elmar Vogt
Ludwigstr. 57
90763 Fürth
`elvogt@gmx.net`
Tel.: (++49) 173/591 29 93