Cryptologia
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ucry20

EVIDENCE OF LINGUISTIC STRUCTURE IN THE VOYNICH MANUSCRIPT USING SPECTRAL ANALYSIS

Gabriel Landini [a]

[a] Oral Pathology Unit, School of Dentistry, The University of Birmingham, St. Chad's Queensway, Birmingham B4 6NN ENGLAND. G.Landini@bham.ac.uk

Published online: 04 Jun 2010.

To cite this article: Gabriel Landini (2001) EVIDENCE OF LINGUISTIC STRUCTURE IN THE VOYNICH MANUSCRIPT USING SPECTRAL ANALYSIS, Cryptologia, 25:4, 275-295, DOI: 10.1080/0161-110191889932

To link to this article: http://dx.doi.org/10.1080/0161-110191889932

EVIDENCE OF LINGUISTIC STRUCTURE IN THE VOYNICH MANUSCRIPT USING SPECTRAL ANALYSIS*

Gabriel Landini

ADDRESS: Oral Pathology Unit, School of Dentistry, The University of Birmingham, St. Chad's Queensway, Birmingham B4 6NN ENGLAND. G.Landini@bham.ac.uk.

ABSTRACT: This paper reports several statistical characteristics of text of the Voynich manuscript which are common to natural languages. In particular, the spectral analysis of the text without spaces suggests that the manuscript shares similar properties with natural languages and is not a random collection of characters. It also shows an indication of the modal token length without relying on the accurate coding of spaces or word separators and gives further evidence of Currier's observation that the text line may be a functional entity. Those characteristics are compared with what is found in other texts including random and encoded versions.

KEYWORDS: Voynich, Kraus, linguistic structure, spectral analysis.

THE VOYNICH MANUSCRIPT

The Voynich manuscript is a medieval or early modern book (234 pages, on vellum) written in an unknown script and what appears to be an unknown language or code. The book is profusely illustrated, suggesting a medical or scientific treatise. According to the nature of drawings, the manuscript has been subdivided into a herbal section (mostly drawings of unidentified and bizarre plants), an astronomical section (with zodiac symbols), a biological section (with drawings resembling anatomical structures and naked human figures), a cosmological section (with circles, stars and celestial spheres), a pharmaceutical section (with

* This paper was one of several presented at the "History of Cryptography Conference" which took place at Cambridge University, 24 June 2000. The sponsor of the conference was the British Society for the History of Mathematics.

vases and parts of plants) and the so-called recipes section (with many short paragraphs, each preceded by a star).

A BRIEF HISTORY OF THE MANUSCRIPT

The origins of the manuscript (date and place) as well as the author remain unknown, but a letter, which was attached to the book, from J. M. Marci (1595-1667) (rector of Prague University) to Athanasius Kircher (1602-1680) (in Rome) locates the book in Prague in 1665/6. Marci related that he inherited the manuscript from a friend and that he was sending it to Kircher, as a present, for decryption. Kircher was a Jesuit priest interested in languages and cryptology, and according to the letter, Kircher had previous knowledge of this manuscript. Marci also reported claims by R. Mnishovsky (1580-1644) that: 1) the manuscript belonged to Holy Roman Emperor Rudolf II (1552-1612), 2) Rudolf paid 600 gold ducats for it, and 3) Roger Bacon (1214-c1292) had been named as one possible author. The manuscript was bought in 1912 by Wilfrid M. Voynich, an antiquarian book dealer, with other manuscripts from the Jesuit college at the Villa Mondragone in Frascati, Italy.

The book was studied by several world renowned cryptologists including William Friedman, John Tiltman and Prescott Currier, as well as by linguists and historians, but it remained unread, giving it the reputation of "the most mysterious manuscript in the world". Voynich died in 1930, but his wife Ethel continued to try to have the puzzle solved. She provided photostatic copies to several scholars. She bequethed the manuscript to Miss A. Nill, Wilfrid's secretary and her long-time companion. In his autobiography, the book dealer H. P. Kraus (1978) relates his buying the manuscript from Miss Nill for $24,500 and his subsequent unsuccessful efforts to sell it. Finally, he donated the manuscript to Yale University in 1969, where it remains at the Beinecke Rare Book Library with catalogue number MS 408.

Several "solutions" have been published [21, 8, 27, 3, 24, 14], all claiming different contents, authors and languages. None has been widely accepted. An account of the problem in the context of cryptology can be found in Kahn's book [10], while detailed information about most work on the manuscript up to the late 1970's is available in D'Imperio [7].

Recently, Zandbergen [32] found new evidence of ownership of the manuscript previous to 1665, ruling out a modern hoax as suggested by Barlow [1]. There still is, however, the possibility of a fabrication, perhaps as part of a plan to get

a large sum from Rudolf, who was interested in arts, science and the occult [3].

Consequently, one of the key problems in the analysis of this puzzling book is to be able to differentiate a real language from meaningless writing and from cryptographic methods used in the middle ages and early modern times. Questions have been raised about the possibility of the text being irrecoverable, for example as a lossy encoding or as the only surviving example of a lost language, natural or artificial. Such questioning about the recoverability of the text seems to have arisen, at least in part, as a consequence of the inability to find a cryptologically sound solution. This, of course, does not prove that a solution cannot be found and so the manuscript remains an interesting cryptological problem.

CHARACTERIZATION OF THE TEXT

Several electronic versions of the Voynich manuscript are available [22, 28]. These use arbitrary alphabets for transcription of the symbols. The latest alphabet, EVA, was created by R. Zandbergen [31] and allows 99.86% of the text to be represented with 27 basic characters including the space. An extended set of about 72 additional characters covers variations of the basic shapes or rare characters (most appearing only once in the whole book). A computer font representing the characters was produced by the author to render a look-alike of the Voynich manuscript script written in the transcription alphabet. In general, several characters resemble the Roman alphabet (a, o, e, ch, sh), some are like Arabic numerals (2, 4, 8, 9), while others are similar to symbols used as Latin abbreviations (cr, 9), characters used in Italian medieval codes (cr, g) or ornamental shapes (f, p) in the Middle Ages (see, for example [4]). The characters are written in groups separated with variably sized spaces forming "words," form lines of text arranged in what appear to be paragraphs. In a few instances, character sequences suggest some order or key. The paragraph endings and the intrusion of drawings in the text indicate that the writing direction is from left to right and top to bottom. Detached words and phrases associated with images, similar to "labels", are also common throughout the manuscript.

The following sections deal with some characteristics shared between the Voynich manuscript and natural languages. The analyses were performed in a preliminary version of the Voynich manuscript produced as a part of a current transcription project [13]. In the sections below, a "token" refers to any string of characters separated by spaces, while a "word" is a type of token, regardless of its frequency. In some instances, two types of randomisation were applied

to the corpora: token-scrambled texts were produced by randomising the order of the tokens, while character-scrambled texts were produced by randomising the position of the characters. The first type preserves the words but destroys any grammatical structure beyond words, the second type destroys both word and grammar structure. Both types of randomisation preserve the character distribution of the original text.

ZIPF'S LAW OF WORD FREQUENCIES

A first step to the characterisation of linguistic corpora is by means of the word frequency distribution. Words in natural languages are not uniformly distributed but approach the so-called "Zipf's law of word frequencies." In this, word frequency is a power law of the rank with exponent $\sim$-1 [35]. Figure 1a shows that the Voynich manuscript approximately follows this law. Despite the fact that Zipf's law is followed in virtually all languages, it has been suggested to be "linguistically, very shallow" [16] since similar distributions arise in randomly generated sequences [15]. However, the word and token length distributions in random texts are markedly different from those in natural languages and consequently easy to identify (Figure 1b).

Unfortunately, token-scrambled texts maintain Zipf's law unaffected, so obviously Zipf's law is not evidence of "meaningfulness", but the vast majority of languages exhibit it, so it seems justified to note also its presence in the Voynich manuscript.

Departures from Zipf's law can be observed in polyalphabetic substitution ciphers because the same word can be enciphered in several ways, depending on its position in the text, thus producing a flattening of the Zipf's plot. Other texts deviating from Zipf's law include item lists and dictionary-like books.

ZIPF'S LAW OF WORD LENGTHS

Zipf also introduced an inverse relationship between the frequency of usage and the length (in phonemes) of words due to "abbreviatory acts of truncation". This relationship also seems to hold in the Voynich manuscript, although not in terms of phonemes, but in number of characters (since it is not known what constitutes a phoneme in the Voynich "language"). Figure 2 shows the word length as a function of frequency rank in the Voynich manuscript.

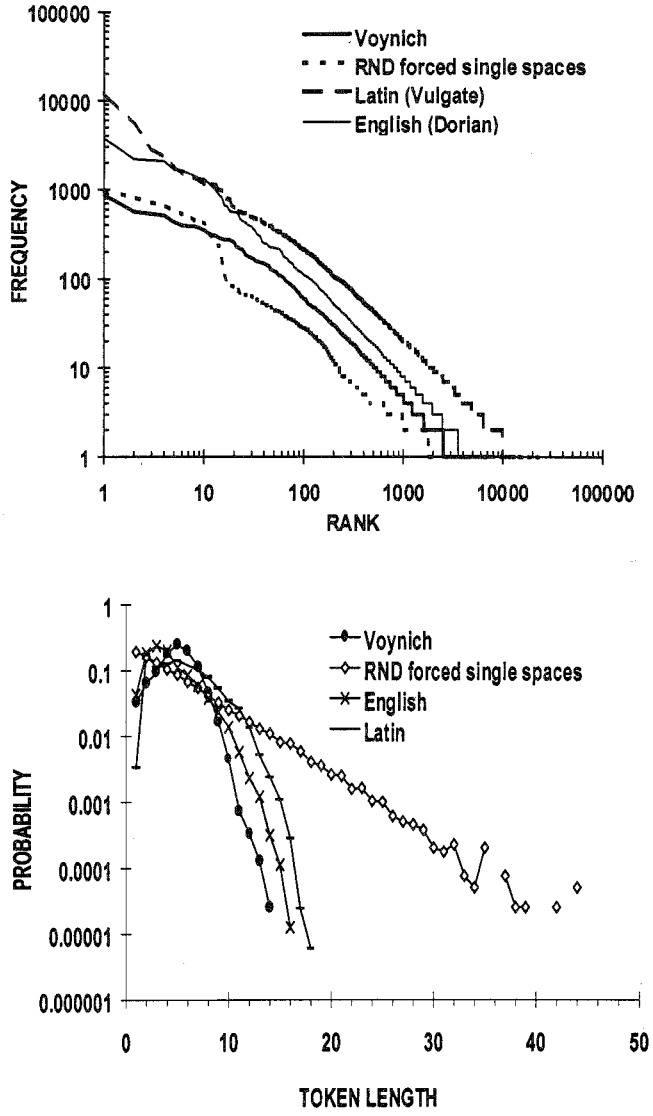


Figure 1a (top). Zipf's rank-frequency law in the Voynich manuscript, a character randomised version, English (The Picture of Dorian Grey, ~80,000 tokens) and Latin (first 10 books of the Vulgate Bible, ~168,000 tokens). The character scrambled version of the Voynich manuscript was forced to single spaces to preserve the character distribution of the original. The dotted line shows a slope of -1 (the theoretical value of the slope in Zipf's law). Deviations from the power law for the most common words is a well known feature of Zipf's plots. Figure 1b (bottom). Despite that random texts roughly follow the rank-frequency law (Figure 1), the token length distributions are quite different. Note that for the Voynich manuscript, the modal token length is 5 characters.

ENTROPY

Bennett (1976) showed that the second-order character entropy of the Voynich manuscript text is lower than in many other European languages (Figure 3, see also Stallings 1998). Given that, on the whole, the Voynich manuscript text (in marked contrast to Latin and English, Figure 1b) contains relatively few words longer than 10 characters, and that many of its symbols resemble medieval Latin abbreviation signs, it has been conjectured that long words were systematically abbreviated.

It seems therefore paradoxical that entropy could be so low in an abbreviated text since abbreviation, in general, reduces word length selectively at the expense of removing redundancy. Thus this has an effect of increasing the entropy of the text rather than reducing it.

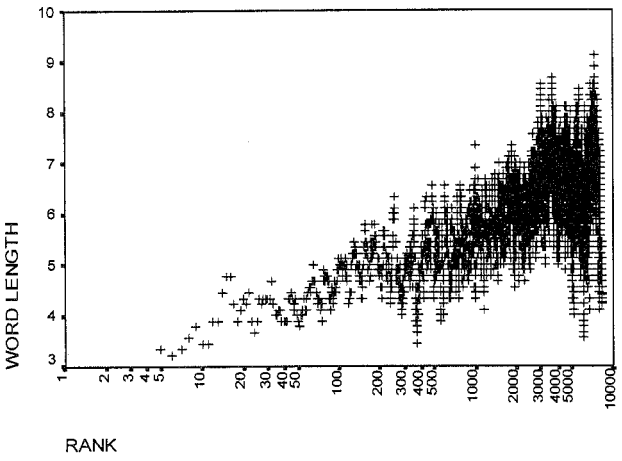


Figure 2. An alternative representation of Zipf's word length-frequency law showing that word length is reduced with use. As phonemes have not been identified in the Voynich script, here the length in characters was used. The dots (+) represent a running average of length 9.

Bennett [2] pointed out that in very redundant ciphers a considerable amount of information lies in the key. He indicated that this gave some support to Brumbaugh's solution based on an ambiguous code [3]. One way to reduce the entropy of a text is to replace some characters with multi-character strings. While this procedure increases the mean token length, it can be compensated for by including spaces in the replacement strings. If the replacement strings

are properly chosen, a reader can learn to read the modified text directly, with little effort. It is not difficult to design such a map so that the transformed text still obeys Zipf's law of word frequencies [12]. Although such a redundant substitution is probably not used in the Voynich manuscript, schemes which reduce second order entropy, shorten word length and still allow a real time decoding are possible. Therefore the low entropy in the Voynich manuscript, although remarkable, may not be as puzzling as Bennett once suggested.

Zandbergen [33], investigating the entropy contributed by each word in the manuscript, found that the vocabulary was as diverse as in Latin texts of similar length. While the first and second character of Voynich words show a lower entropy than Latin, the Voynich words contain more "information" from the third character onwards. These results highlight the problem of working with cryptanalysis in which there is no certainty that each character is in fact a single character.

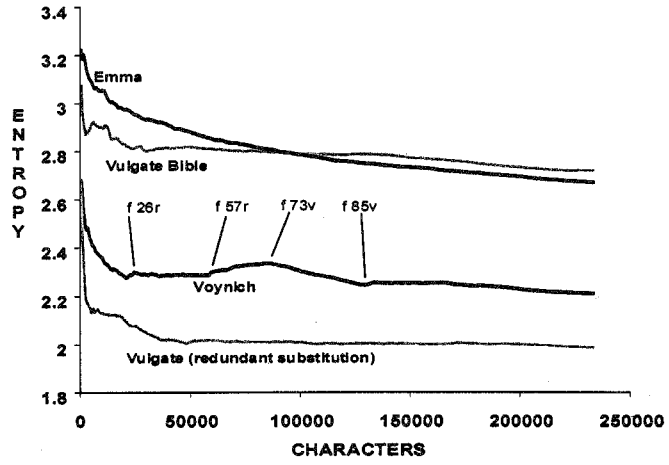


Figure 3. The asymptotic entropy calculated with a universal algorithm for sequential data compression [36] in various texts. In the Voynich text, entropy is lower than that of English (Jane Austen's *Emma*) or Latin (Vulgate Bible). Several changes in the plot of the Voynich manuscript indicate the changes between Currier's "languages" A and B (Currier 1976). Up to folio 26r all the text is Herbal A; from 26r to 57r there is a 40% of folios in B language. The fragment from 57r to 73v corresponds with the Cosmological, Astronomical and Zodiac sections. From folio 73 onwards 73% of the pages are written in language B. Note the drastic reduction of the entropy in the Vulgate Bible when encoded with a redundant substitution as described in the Entropy section.

CRYPTOLOGIA October 2001 Volume XXV Number 4

FURTHER MEASURES OF LANGUAGE STRUCTURE

There is obvious organisation in written communication: a piece of text may reveal, for instance, characters, syllables, words, phrases, sentences, paragraphs and sections or chapters. For a string of characters, the space character provides the method for isolating words (in spoken language, word segmentation is more difficult), so once the word set is retrieved, then concordance lists can be used to reveal syntactical structures or grammatical meaning. However, these relationships rely heavily on the accurate coding of the text. Unfortunately, in the Voynich manuscript, such accurate coding remains problematic since no punctuation exists and a proportion of the characters have alternative readings because it is written by hand. For example, the spacing between words varies throughout the manuscript and, where the spacing is tight, relatively common words may be run together and misrepresent the text. Instead, a more relaxed measure of structure in the corpus would be appropriate. This was investigated using symbol correlation analysis. Before going into the details of symbol correlations in the Voynich manuscript, a very brief introduction to correlation and spectral analyses is given below.

Correlation analysis

For random processes in time, the variation of a quantity x between times t and $t+\tau$ can be estimated by the autocorrelation function C,

$$C_x(\tau) = \langle x(t)x(t+\tau)\rangle$$

where the $\langle\rangle$ brackets indicate ensemble averages. Alternatively the variation can be estimated in the frequency (f) domain by the spectral density S:

$$S_x(f) \propto \frac{\left|\int_{-\infty}^{\infty} x(t)e^{-2\pi ift}\,dt\right|^2}{\Delta f}$$

where $\propto$ indicates proportional. Furthermore, both quantities are related since the power spectrum $S(f)$ is the Fourier transform of a C_x.

For non correlated random time series, the power spectrum is flat, while for so-called scaling processes, the spectrum is characterised by a power law:

$$S_x(f) \propto \frac{1}{f^\beta}.$$

The generally accepted models for series with long range correlations are the fractional Gaussian noises $(-1 < \beta < 1)$, and fractional Brownian motions $(1 < \beta < 3)$ [17]. A large number of natural phenomena exhibit fluctuations with $\beta = 1$, but most surprisingly $1/f$ fluctuations are present in many human-driven activities such as in vehicle traffic in roads, exchange market prices and pitch and loudness fluctuations in music and speech (see for example [18]). The origin of such fluctuation has, in many cases, remained elusive.

Correlations in symbolic sequences. For symbolic sequences with K symbols the assignment of numeric values to symbols introduces bias. This can be avoided by coding the position of a symbol in the sequence using the equal-symbol multiplication introduced elsewhere [29]:

$$x_n x_m = 1 \quad \text{if} \quad x_n = x_m, \quad 0 \quad \text{otherwise.}$$

This consists in decomposing the sequence into K binary sequences indicating position of symbol k. The Fourier transform of the k sequences gives the frequency components $U_k(f)$, from which the power spectrum $S_k(f)$ can be calculated:

$$S_k(f) \propto |U_k(f)|^2$$

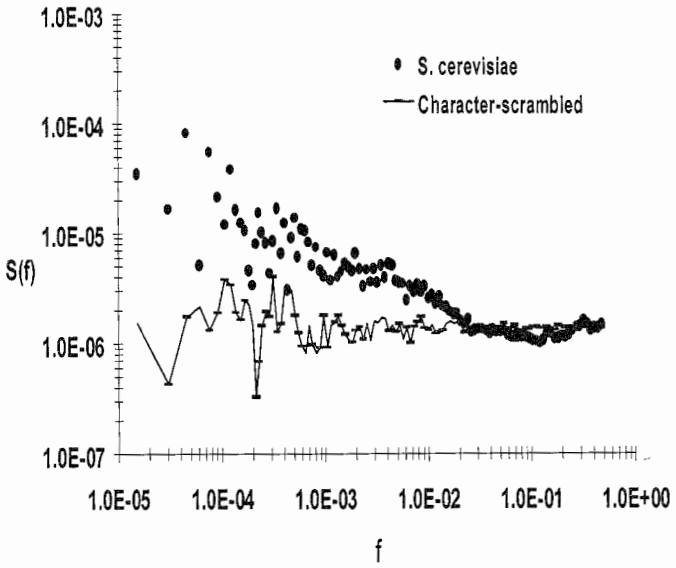
and for the ensemble of symbols:

$$S(f) = \sum_{k=1}^{K} S_k(f).$$

To reveal the underlying trends, spectra can be smoothed by averaging with filters of effective bandwidth Δf proportional to f, although other alternatives are possible [5]. Again, long range correlations in the sequence appear as a power law relationship in the power spectra when plotted on log-log scales. The power spectrum is relevant to the present problem because a periodic trend in the sequence x_i with period τ shows up as a peak in the power spectrum, at frequency n/τ.

DO LINGUISTIC STRUCTURES GENERATE CORRELATION PATTERNS?

Spectral analysis has not been applied to written works before, but it has been used to characterise global properties of DNA sequences [29, 30]. Briefly, the analysis of the entire holdings of the Genbank release 73 revealed two main findings. First, the spectral power of genetic sequences has power law-like long range



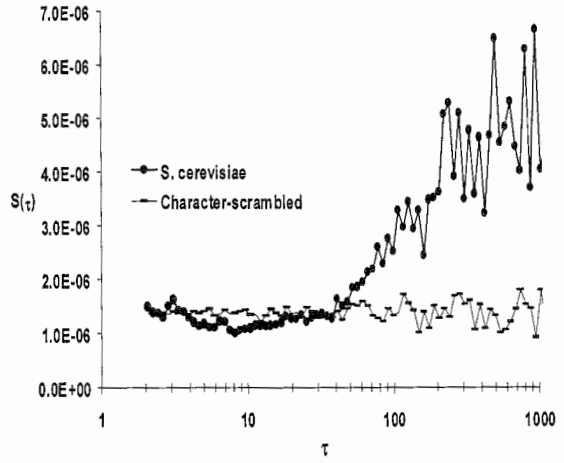


Figure 4a (top). Spectral analysis of the Saccaromyces cerevisiae mitochondrial complete genome (GenBank Accession NC_001224) and in the base-scrambled sequence. The long range correlation behaviour appears as a deviation of the plot from the horizontal. In Figure 4b (bottom), a semi-log plot in function of the period shows the codon peak at $\tau = 3$ which is absent in the randomised sequence.

correlations at very large scales (Figure 4a). The meaning of these correlations remains obscure; it has been argued that they may provide a means of information transfer with reduced sensitivity to noise [29], but they could also be due to the variable local density of bases in the genome. Moreover, the power spectrum plots show a sharp peak at period $\tau = 3$ (Figure 4b). Interestingly, this peak coincides with the "word" length of the genetic code (DNA stores information coding for amino acids in "words" 3 bases long called "codons") despite the fact that there are no word-delimiter symbols or codons in DNA. The scrambled version of the same sequence (RND in Figure 3) lacks the peak at period=3 and the long range correlation behaviour. Instead, it approaches a flat spectrum characteristic of a white noise (random) source. This prompted the idea that spectral analysis may reveal structural differences between original and randomised versions of written texts. This was investigated in texts with various periodic features (Table 1).

Text	Author	Language	Date	Nature
GenBank NC_001224	—	genetic code	—	*S. cerevisiae mitochondrial genome*
Metamorphoses	Ovid	English translation (Garth)	1AD (c1717)	Verse
Canterbury Tales	Chaucer	English	c 1390	Verse
The Picture of Dorian Grey	Wilde	English	1890	Prose
Emma	Austen	English	1815	Prose
Vulgate Bible	unknown	Latin	c405	Prose
Voynich Manuscript	unknown	unknown	unknown	cipher?

Table 1. Text sources used in the symbol correlation analysis.

All the texts (except the DNA sequence) were analysed in 3 versions: raw, token-scrambled and character-scrambled. "Raw" texts were the original corpora converted to lower case, with all punctuation, numerals and spaces removed and truncated at 217 characters. "Token-scrambled" texts had, in addition, the token order randomised while in "character-scrambled" texts had the letter order randomised. Spaces were also removed after randomisation.

METHODS AND RESULTS

For all texts, including the Voynich manuscript, the spectral analysis was performed on text strings converted to lowercase, with all spaces, punctuation and numerals removed. The spaces were removed because in the case of the Voynich

manuscript, they are the least reliable character to transcribe. The strings were $n = 2^{17}$ characters long (the nearest power of 2 to the length of the Voynich manuscript). After the power spectra for the individual symbols were calculated, the data were smoothed by averaging them in 112 logarithmically sized bins spanning the full frequency range to visualise just the trends. The averages of the spectra were calculated and displayed as functions of the frequency f or the period τ.

Figure 5 shows the spectral patterns in Ovid's *Metamorphoses* and in the token- and character-scrambled versions. The original and the word-scrambled versions showed a peak at $\tau = 3.2$, but the peaks disappeared in the character-scrambled sequence. The main difference between the token-scrambled sequence and the original was a peak at $\tau = 32.7$ These two peaks occurred close to the modal token length and modal verse length of 3 and 34 characters respectively as shown in Figure 5b. Another peak was observed at $\tau = 16.4$ which is half the value of the modal verse length. Note that the original text tended to have, despite the peaks just mentioned, a lower value of $S(\tau)$ than the corresponding word-scrambled text in the range from $\tau = 10$ to 200.

In general these differences between the plots described above were also found in the other texts investigated: the character-scrambled plots were essentially flat from 2 to 1000 characters, while the original and token-scrambled plots showed a local maximum somewhere in the 2 to 10 characters range and a local minimum roughly in the 50 to 100 characters range. This local minimum was always smaller in the original versions than in the token-scrambled texts.

Similar results were obtained for the *Canterbury Tales* (Figure 6a); the modal verse length was 31 characters; this showed on the correlation graph as a peak at $\tau = 32.7$. The modal token length was 3 characters and the peak in the $S(\tau)$ graph appeared at $\tau = 2.9$. These peaks gradually disappeared when the text was coded using polyalphabetic substitutions with increased number of alphabets (Figure 6b).

For the prose writings investigated, the sentence length was considerably more variable than the verse length of the pieces analysed, and the effects of sentence length on the symbol correlation spectrum became less obvious. For "The Picture of Dorian Grey," the modal token length was visible in the graphs of the original as well as the token-scrambled version, but the sentence length was not obvious. Despite this, the values of $S(\tau)$ tended to be smaller than those of the token-scrambled version as in the other examples. Similar results were observed in Jane Austen's *Emma* and the Vulgate Bible (graphs not shown). Evidently the

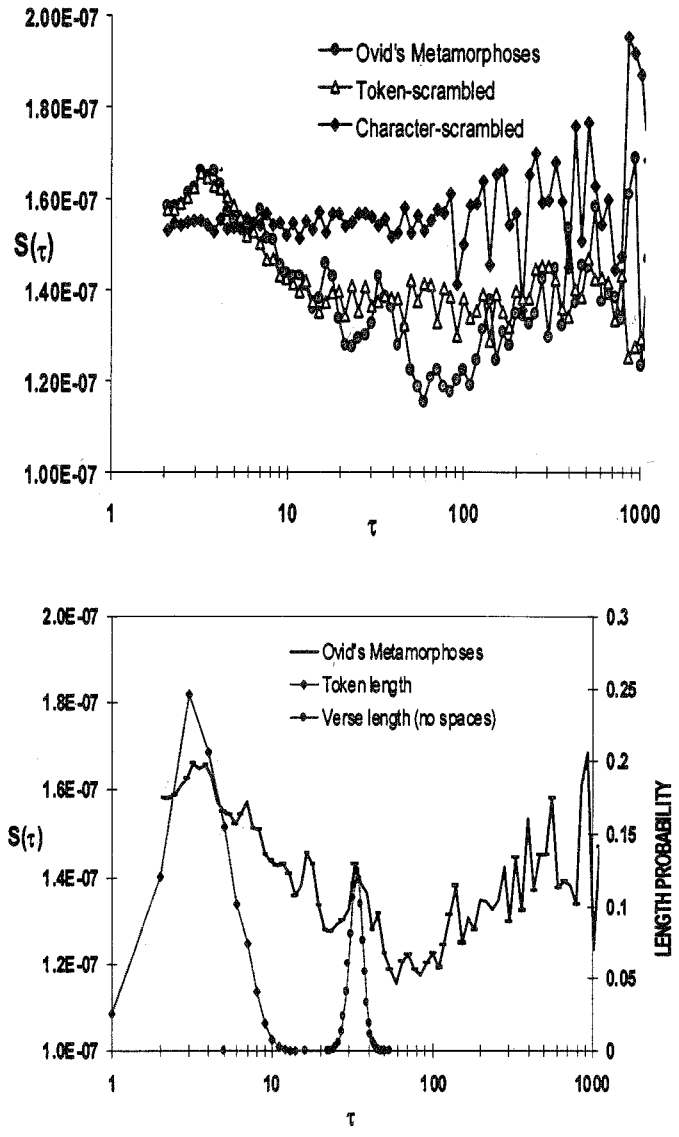


Figure 5a (top). Ovid's Metamorphoses (spacing removed) and the token- and character-scrambled sequences. Note the peaks at $\tau \approx 3.6$ and 32. Both peaks are absent after character-scrambling, but the first remains after token-scrambling. Figure 5b (bottom) shows that those peaks correspond with the token length and verse lengths.

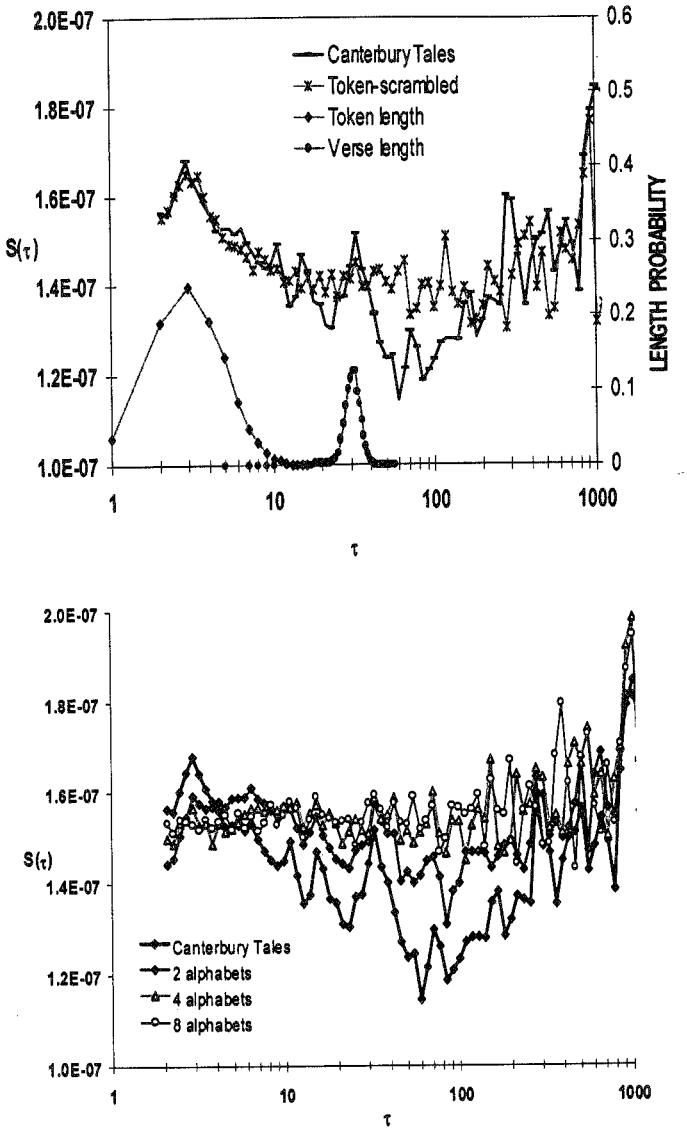


Figure 6a. The Canterbury Tales by Chaucer (spacing removed). Again the token and verse length distributions coincide with peaks in the symbol correlation graphs as in Figure 5. In Figure 6b is shown the effect of polyalphabetic substitutions: the graphs approach a flat spectrum when the number of alphabets increases.

Text	Modal token length	Token peak	Token peak after token-scrambling	Token peak after character-scrambling	Modal sentence/line length	Sentence peak
GenBank NC_001224	3	3.1	—	no	—	—
Metamorphoses	3	3.2	3.2	no	34	32.7
Canterbury Tales	3	2.9	2.9	no	31	32.7
The Picture of Dorian Grey	3	3.2	3.2	no	broad peak (local max=30)	not obvious
Emma	3	3.2	3.2	no	broad peak (local max=43)	not obvious
Vulgate Bible	2	2.7	2.5	no	not estimated	not obvious
Voynich Manuscript	5* or 6+	5.9	5.4	no	44	46.1

Table 2. Periodic features in the texts samples.

* Calculated from the version coded with spaces in EVA alphabet (Figure 1b); because of the nature of the script and the variable spacing, this value may not be fully reliable.

+ Calculated from the line-length distribution in EVA alphabet (Figure 8b).

power spectrum curves in grammatically meaningful texts, even without spaces, give an indication of the modal token length and some indication of structure (modal verse length), if present, and help to differentiate between character- and token-scrambled texts.

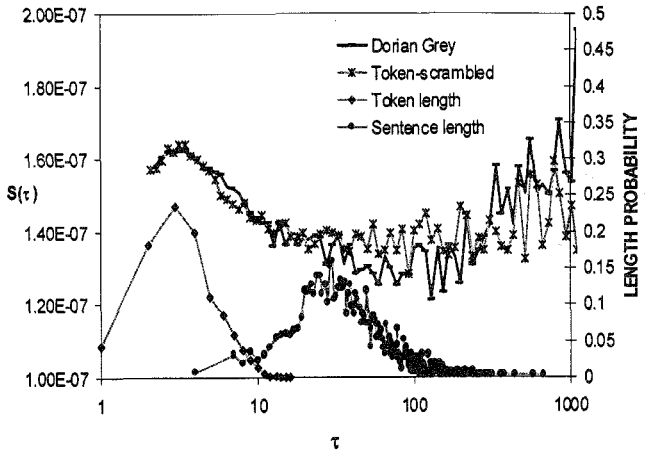


Figure 7. The analysis of The Picture of Dorian Grey (in prose) shows that the sentence length effect on the correlation patterns is less clear than in the previous examples in verse. There is no clear peak corresponding with the maxima of the sentence length distribution.

Would the correlation curve give any indications of modal token length and sentence/verse length in the Voynich manuscript where the spacing is unreliable? In the plots of the Voynich manuscript without spaces there was a marked long range correlation slope when compared with the scrambled versions (Figure 8a). As observed for all the other samples, the $S(\tau)$ curve showed smaller values in the range from $\tau = 10$ to 200 than the other two scrambled versions. Two peaks in the power spectrum were present at $\tau = 5.9$ and $\tau = 46.1$ (Figure 8b). Both peaks disappeared, as in previous cases, after character-scrambling and the peak at $\tau = 46.1$ disappeared after token-scrambling. It seems reasonable to assume that the first peak relates to the rules of word construction while the second is a consequence of the token arrangement in the text. This prompted the question of what do these peaks corresponded with in the Voynich manuscript.

As the manuscript lacks any recognisable punctuation, the only reliable measure that could be extracted to be compared with was the line length. Figure 8b shows that the line length distribution was bimodal with local maxima at 6 and at 44 characters. The first peak corresponded with isolated words (or "labels") across the manuscript while the second was close to the line length in the folios

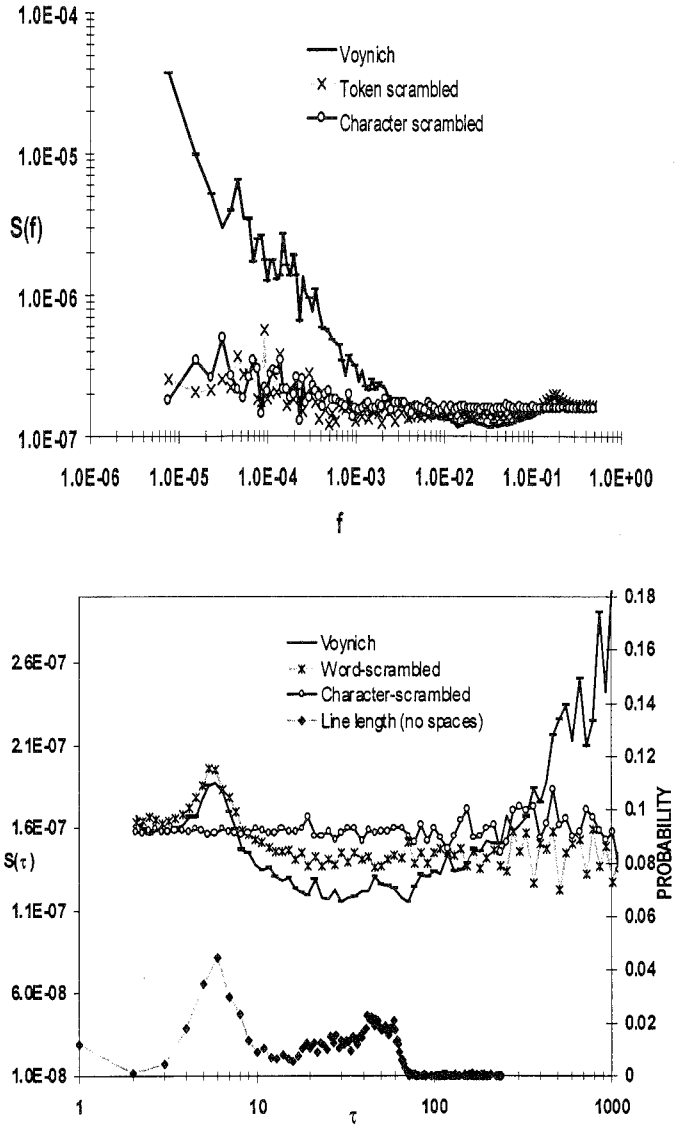


Figure 8a. Spectral analysis of the Voynich manuscript. Note the differences in long range correlations compared with the scrambled, versions of the same text. Figure 8b shows that the line length distribution maxima has a corresponding peak in the power spectrum. The peak at $\tau = 5.9$ for the original is very close to the local maxima of 6 (single words or "labels" through the manuscript). It is also close to the modal token length (5 characters) estimated from the text coded with spaces in Figure 1b.

rich in text paragraphs. These two peaks appeared at τ values very close to those in the $S(\tau)$ plot. Table 2 shows the periodic features found in the texts analysed.

CONCLUSIONS

Spectral analysis was used to characterise symbol correlation in various writings and revealed:

a) long range correlation at large scales,

b) a peak indicating the modal token length,

c) a decay in the power spectrum compared with the scrambled versions of the text, and

d) a peak that corresponded with periodic structures in the text (verse length in poetry).

Regarding these features in the Voynich manuscript, the long range correlation at large scales ($\tau > 400$ characters) may be, as speculated for DNA sequences, the result of the uneven character density along the text. There are several reasons why this could be so: changes in the "language" as observed by [6], change of subject as depicted by the drawings in the different sections of the book, or the insertion of the labels in the text. Further evidence in the unevenness of the local character density along the text is shown in Figure 3.

The decay in the power spectrum relative to the scrambled versions suggests that the text in the Voynich manuscript is not a random collection of characters. A similar conclusion was reached by [20] using an alternative metric they introduced, called letter serial correlation analysis.

The peak at length $\tau = 46.1$ reinforces an original observation by Currier [6] regarding the structure of lines of text in the Voynich manuscript. Currier noticed that certain characters and words tended to occur in specific locations (i.e. line-initial, line-terminal) while some others never occurred in those positions. Furthermore, certain characters ( ,  ,  and ) were very common at the beginning of paragraphs which raised the question of whether they were ornamental or nulls since some other characters appeared to behave as in natural languages [9].

Friedman believed that the Voynich manuscript was written in some kind of artificial or synthetic language (see, for example, [34]). Currier thought that the

"words" were not strictly words and appears to have contemplated the synthetic language theory too. Stolfi [25] suggested that the manuscript could be in some tonal language like Chinese, but written in an alphabetic script, which would explain the constrained structure of the Voynich words [26].

The findings shown here favour the natural language theory. But given the various oddities seen in the Voynich manuscript, other alternatives should not be discarded without solid evidence. Whether Currier's line structure observation, confirmed here by the correlation analysis, is due to metric constrains (like prayers or hymns), the presence of null or ornamental characters or even an artificial language deserves further investigation.

ACKNOWLEDGEMENTS

The author thanks Drs. James Reed, Jorge Stolfi, and René Zandbergen for many useful comments during the preparation of this paper and Dr. Judith V. Field for the kind invitation to the Meeting on the History of Cryptology, Cambridge, 2000.

REFERENCES

1. Barlow, M. 1986. The Voynich Manuscript – By Voynich? *Cryptologia*. 10(4): 210-216.

2. Bennett, W. R. 1976. *Scientific and engineering problem solving with the computer*. Englewood Cliffs: Prentice-Hall.

3. Brumbaugh, R. S. 1977. *The world's most mysterious manuscript*. Carbondale: Southern Illinois University Press, 1978. London: Weidenfeld and Nicholson.

4. Cappelli, A. 1929. *Lexicon Abbreviaturarum. Dizionario di abbreviature latine ed italiane*. Milan: Editore Ulrico Hoepli (reprint of sixth edition, 1967).

5. Chechetkin, V. R. and A. Yu Turygin. 1994. On the spectral criteria of disorder in non-periodic sequences: Application to inflation models, symbolic dynamics and DNA sequences. *Journal of Physics A: Mat Gen.* 27: 4875-4898.

6. Currier, P. 1976. Some important new statistical findings, and The Voynich Manuscript, some notes and observations. In D'Imperio, M. E., editor. *New research on the Voynich manuscript: proceedings of a seminar*. Washington DC: Privately printed pamphlet, 30 November 1976.

7. D'Imperio, M. E. 1978. *The Voynich Manuscript - An elegant enigma.* Laguna Hills CA: Aegean Park Press.

8. Feely, J. M. 1943. *Roger Bacon's cipher: the right key found.* Rochester NY. (No publisher, mentioned by [7]).

9. Guy, J. B. M. 1997. The distribution of letters ⟨c⟩ and ⟨o⟩ in the Voynich Manuscript: Evidence for a real language? *Cryptologia.* 21(1): 51-54.

10. Kahn, D. 1967. *The Codebreakers.* New York: Macmillan pp. 863-872, 1120-1121.

11. Kraus, H. P. 1978. *A rare book saga. The autobiography of H. P. Kraus.* New York: Putnam's.

12. Landini, G. 1998. The "dain daiin" hypothesis.
http://web.bham.ac.uk/G.Landini/evmt/daindaiin.htm.

13. Landini, G. and R. Zandbergen. 1996.
http://web.bham.ac.uk/G.Landini/evmt/rules.htm.

14. Levitov, L. 1987. *Solution of the Voynich Manuscript: A liturgical manual for the endura rite of the cathari heresy, the cult of Isis.* Laguna Hills CA: Aegean Park Press.

15. Li, W. 1992. Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE Transactions on Information Theory.* 38(6): 1842-1845.

16. Mandelbrot, B. B. 1965. Information theory and psycholinguistics. In Wolman B. B. and E. N. Nagel. *Scientific Psychology: Principles and Approaches.* New York: Basic Books. pp. 550-562.

17. Mandelbrot, B. B and J. W. Van Ness. 1968. Fractional Brownian motions, fractional noises and applications. *SIAM Review.* 10(4): 422-437.

18. Mandelbrot, B. B. 1982. *The Fractal Geometry of Nature.* San Francisco: Freeman.

20. McKay B., and M. Perakh. 1999.
http://www.nctimes.net/~mark/Texts/.

21. Newbold, W. 1921. The Cipher of Roger Bacon. *Proceedings of the College of Physicians and Surgeons of Philadelphia.* pp. 431-74. Philadephia, read on April 20.

22. Reeds, J. 1995. William F. Friedman's transcription of the Voynich Manuscript. *Cryptologia.* 19(1): 1-23.

23. Stallings, D. J. 1998. Understanding the second-order entropies of Voynich text. http://www2.micro-net.com/~ixohoxi/voy/mbpaper.htm.

24. Stojko, J. 1978. *Letters to God's Eye: The Voynich Manuscript for the first time deciphered and translated into English.* New York: Vantage Press.

25. Stolfi J. 1997. The generalized Chinese theory.
http://www.dcc.unicamp.br/~stolfi/voynich/97-11-23-tonal/.
26. Stolfi J. 2000. A grammar for Voynichese words.
http://www.dcc.unicamp.br/~stolfi/voynich/00-06-07-word-grammar/.
27. Strong, L. C. 1945. Anthony Askham, the author of the Voynich Manuscript. *Science*. 101(15 June): 608-9.
28. Takahashi T. 1998. Transcription of the Voynich Manuscript.
http://www.voynich.com/pages/index.htm.
29. Voss, R. F. 1992. Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. *Physical Review Letters*. 68(25): 3805-3808.
30. Voss, R. F. 1994. Long-range fractal correlations in DNA introns and exons. *Fractals*. 2(1): 1-6.
31. Zandbergen, R. 1996. Transcription of the Voynich manuscript.
http://www.voynich.nu/transcr.html.
32. Zandbergen, R. 1999. Voynich manuscript history after 1600.
http:// www.voynich.nu/history.html.
33. Zandbergen, R. 2000. From digraph entropy to word entropy in the Voynich manuscript.
http://www.voynich.nu/wordent.html.
34. Zimansky, C. A. 1970. William F. Friedman and the Voynich Manuscript. *Philological Quarterly*. 49(4): 433-42
35. Zipf, G. K. 1935. *The Psycho-biology of Language*. Boston: Hough- ton Mifflin Co.
36. Ziv J. and A. Lempel. 1977. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*. IT-23: 337-343.

BIOGRAPHICAL SKETCH

Dr. Gabriel Landini is currently a Senior Lecturer in Analytical Pathology at The University of Birmingham, UK. His main research interests are medical imaging and fractal geometry applied to the analysis and quantification of invasive patterns of cancer. Dr. Landini obtained his Doctor in Odontology degree from the Republic University (Uruguay) in 1983 and a PhD in Oral Pathology from Kagoshima University (Japan) in 1991. His interest in Cryptology and the Voynich manuscript started while researching on the information contents and statistical properties of symbolic sequences and large-scale patterns in DNA.