

Polygraphia III: The cipher that pretends to be an artificial language

Jürgen Hermes¹

¹*Department of Digital Humanities, University of Cologne, 50923 Cologne*

Abstract

This paper demonstrates the existence of a cipher method in the early modern period (Polygraphia III by Johannes Trithemius), which – applied as a random procedure – is able to produce a text that can mimic the oblique properties of the so-called Voynich Manuscript (VMS). This result is quite exciting since it brings back into play highly-debated approaches claiming the existence of hidden comprehensible information within the text of the VMS (which is often referred to as Voynichese). The paper briefly outlines some of the most salient and difficult-to-explain statistical properties of Voynichese, shows how Trithemius stepwise developed a cipher system whose application looks like an artificial language, points out how an application of this cipher generates a text that comes very close to the statistical properties of Voynichese and finally discusses possible starting points of cryptanalytic attacks on a cipher that operates similar to the Polygraphia III encryption.

Keywords

Artificial Languages, Statistical Properties, Cryptanalysis, Johannes Trithemius

1. Introduction: The odd statistical properties of Voynichese

It doesn't matter what aspect of the VMS is dealt with, abysmal mysteries open up everywhere. One of the many unsolved riddles that have haunted us with this manuscript is undoubtedly the extremely strange statistical behaviour of its text – a feature yet to be observed in any other historical writing. The odd properties of the VMS have often been discussed. Some of them suggest a natural language origin such as the Zipf distribution [1], the word entropy [2] and the prefix-midfix-suffix structure of words [3]. Other features do not even come close to those of known natural languages. Due to the limited space, only three particularly striking features can be taken up here: The low joint entropy, the binomial word length distributions and the astonishing repetitiveness of the text.

1.1. Joint entropy

At first, it was thought that the peculiarity of Voynichese was a very low joint entropy h_2 (1), otherwise only found in some Polynesian languages [4]. This means that one can predict a

International Conference on the Voynich Manuscript 2022, November 30–December 1, 2022, University of Malta.


✉ hermesj@uni-koeln.de (J. Hermes)

🌐 tinyurl.com/hermesatidh/ (J. Hermes)

🆔 0000-0002-8367-8073 (J. Hermes)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

letter of the VMS – based on knowledge of its predecessor – much better than it is the case with European languages.

$$h_2 = \sum_{I=1}^n p(I, J) * b(I, J) = - \sum_{I=1}^n \sum_{J=1}^n p(x_I) * p(y_J|x_I) * \log_2(p(y_J|x_I)) \quad (1)$$

What was not taken into account was that natural languages with a low h_2 (like Hawaiian and Japanese) have only very few different letters and thus also a very low character entropy h_1 (2). It is Stallings' merit to work out that the difference between the two measurement numbers ($h_1 - h_2$) is much more meaningful here and that the low value that the Voynich manuscript shows does not occur in any known letter script and that at best syllabic scripts come close to it [5]. Even at the character level, Voynichese is a very strange outlier, "and we might be tempted to conclude from this that the text is meaningless." [6] So let's take a look at the word level.

$$h_1 = - \sum_{I=1}^n p(I) * \log_2(p(I)) \quad (2)$$

1.2. Word lengths distribution

When looking at word lengths the text of the VMS is astonishingly uniform (hardly any words have less than 3 or more than 10 characters). Even more surprising is the similar behaviour of type lengths and token lengths. Following Zipf's law of abbreviations, natural languages should all have a rather left-skewed distribution in the frequency of token lengths [7][8]. Although Voynichese tokens are also slightly shorter on average than types, the word length distributions of both, types and tokens, is almost binomial [9] [10].

1.3. Distribution of similar words

The words of the VMS text are not only relatively similar in terms of lengths, they are also alike in terms of combinations of characters used. In [11] it is shown, that "for each common word, there is at least another one differing from it by only a single quill stroke". This makes it possible to build an extremely dense similarity network out of the types of the manuscript. This feature alone makes the text of the VMS seem very repetitive. This impression is heightened by the fact that similar, sometimes even identical words can be found in the direct vicinity. A well-known example from in the first paragraph of f78r (Takahashi transcription) is *qokeedy qokedy shedy tchedy [...] qokal otedy qokedy qokedy dal qokedy qokedy rgam*. These sequences may be reminiscent of children's writing exercises or the enumeration of inflection paradigms, but not of flowing texts of natural languages.

1.4. Interim conclusion: What Voynichese cannot be

Consequently, all approaches assuming an unknown transcription of a natural language or a simple substitution cipher as the origin of Voynichese must fail in providing a comprehensible explanation of these statistical features of the text – which explains why Elizebeth Friedman described statistical analyses of the VMS as "doomed to utter frustration". [12] Elizebeth's

husband, William F. Friedman, widely regarded as one of the greatest cryptanalysts of the 20th century left for posterity his hypothesis that the manuscript might be “an early attempt to construct an artificial or universal language of the a-priori type”. [13] However, this approach obviously also has some limitations: The earliest universal language designs of Wilkins [14] and Dalgano [15] are basically far too systematic to produce a rather wild text like the VMS. Moreover, they were drafted in the advanced 17th century, while – as far as we know – the VMS is definitely older. [16]

2. Method: Trithemius’ stepwise mimification of an artificial language

To resolve this contradiction, we would like to introduce a much older (but still about a 100 years younger than the VMS) draft of a cipher developed for the purpose of encryption, whose application leads to a text that has the character of an artificial language. This cipher is found in the first printed book on cryptology, namely the Polygraphia written by Johannes Trithemius. [17]

Trithemius (1462-1516) was a German Benedictine abbot and a polymath. The Polygraphia from 1506 was the first ever printed book on the subject of cryptology, although it was already Trithemius’ second book on this subject, for he wrote the Steganographia as early as 1499/1500. [18] However, the book’s arcane characteristics were so incomprehensible to the public, including representatives of the Catholic Church, that it was placed it on the Index of Forbidden Books for more than 250 years. [19]. Following the example of the Jewish Kabbala, Trithemius develops a series of procedures in the Steganographia that hide, (and - in some cases additionally encrypt) secret messages in non-suspicious looking texts. [20] He himself provides instructions for the procedures in the first two books, but remains silent on the third book (S.III). Not until almost 500 years later, in one of the greatest cryptanalytic achievements of the last 50 years, it could be shown that S.III also contained an encryption method. [21][22]

The procedures described in the Steganographia are based on the fact that the letters of the plaintext occur at specified positions of the ciphertext and the rest is filled with nulls. This comes with a great deal of effort for the creator of the message, because he has to self-invent an extremely large amount of inconspicuous text. With the procedures introduced in the Polygraphia, Trithemius attempts to limit this effort. [23] In the first two parts of his book Trithemius shows how to camouflage a plaintext within a Latin prayer - without speaking one word of Latin - by replacing individual letters of the plaintext with Latin words. The words are arranged in substitution tables in a way that the cipher text appears syntactically and semantically coherent.

Table 1 illustrates the principle of the so-called Ave Maria (Hail Mary) cipher: To produce the ciphertext, one must look for the corresponding word in the correct row for each letter of the plaintext (leftmost column). For the next letter, one has to use the same method with the following column. Consequently, the encoding of the English word *secret* would be *Conſervator magnuſ conſervanſ cuncta iuſtiſ ſuiſ in felicitatibuſ amen*. For the first two books of the Polygraphia (P.I and P.II), Trithemius creates a total of almost 700 columns (383 in P.I and 308 in P.II) with which a plaintext containing a corresponding number of letters (and even more, when one gets back

to the start when reaching the end) can be camouflaged by an inconspicuous Latin prayer.

Column	P.I-1	P.I-2	P.I-3	P.I-4	P.I-5	P.I-6
a	Deus	clemens	creans	celos	sanctis	celis
b	Creator	clementissimus	regens	celestia	electis	celestibus
c	Conditor	pius	conservans	supercelestia	predilectis	supercelestibus
d	Optifex	piissimus	moderans	mundum	sanctissimis	eternum
e	Dominus	magnus	gubernans	mundana	iustis	perpetuum
f	Dominator	excelsus	ordinans	homines	iustificatis	sempiternum
g	Consolator	maximus	ornans	humana	predestinatis	fecola feculorum
h	Arbiter	optimus	egornans	angelos	angelis	euum sanctum
i	Judex	sapientissimus	constituens	angelica	archanges	feculum
k	Illuminator	inuisibilis	dirigens	terram	amatoribus	regno celorum
l	Illustrator	immortalis	producens	terrana	cultoribus	altissimis
m	Rector	eternus	decorans	tempus	amicis	excelsis
n	Rex	sempiternus	stabiliens	temporalia	apostolis	paradiso
o	Imperator	gloriosus	illustrans	euum	prophetis	olympo
p	Gubernator	fortissimus	intuens	euiterna	discipulis	paradisiacis
q	Factor	sanctissimus	monens	omnia	martyribus	olympicis
r	Fabricator	incoprehensibilis	confirmans	cuncta	sanctificatis	fulgoribus
s	Conservator	omnipotens	custodiens	uniuersa	dominationibus	felicitate
t	Redemptor	pacificus	cernens	orbem	dilectis	felicitatibus
u	Auctor	misericos	discernens	astra	ciuibus	gloriosis
z	Princeps	miserordissimus	illuminans	solem	seruis	honore
y	Pastor	cunctipotens	fabricans	stellas	famulis	magnificentia
z	Moderator	magnificus	saluificans	vitam	ministris	luce perpetua
w	Saluator	excellentissimus	faciens	diuentia	confessoribus	patriacelesti

Table 1

The first six columns of the Ave Maria cipher from Polygraphia I. Trithemius gives the additional instruction to insert an *uis* in before the sixth column and an *amen* after it

Within the next two parts of the Polygraphia, Trithemius maintains the principle of replacing individual letters with whole words. He uses no longer Latin words as substitution ciphers, but words that he invented especially for this purpose. In the third book (P.III), he makes use of features of natural languages, which appear as if the same stem had been inflected in different ways. In P.IV he goes the opposite way, since same word endings are combined with different word stems (see table 2). Attentive observers will notice that in P.IV the plaintext letter is always in the second position of the replacement word. This part is therefore based on a very simple steganographic procedure. Although the principle in P.III looks similar, it works completely differently, because one cannot determine the plaintext letter from any component of the substitution cipher and only from its position in the substitution table. Trithemius intended that one creates the ciphertext as in the Ave Maria cipher, that is, using each column in the order given. However, it is also possible to use only one column to produce the ciphertext. For example, it is possible to encipher *secret* with *pasil pasu pasi pasel pasu pasol* by only using column P.III-5. The result is a very repetitive text, where words of a certain similarity consequently encode different letters, what was hardly known from natural languages. In the next section we will demonstrate that an application of the cipher generates a text that also comes very close to

other odd statistical properties of Voynichese.

For the sake of completeness: In P.V, Trithemius publishes the now famous transposition table, with which he goes down in history as one of the three inventors of polyalphabetic substitution. In P.VI he treats different alphabets, besides Latin the Greek and various Frankish ones, some of which he may have invented himself.

Column	P.III-5	P.III-15	P.IV-8	P.IV-15
a	pafa	mafra	baron	famelech
b	pafe	mafte	abaron	ebramelech
c	pafi	maftri	ocarion	achalech
d	pafv	maftrv	adelon	adelmech
e	pafu	maftru	meron	nemelech
f	pafan	maftran	ofilon	afemelech
g	pafen	mafren	agion	agefelech
h	pafin	maftrin	chorion	thomelech
i	pafon	maftron	libion	diralech
l	pafun	maftrun	afyron	afafelech
l	pafas̄	maftral	elychon	alanech
m	pafes̄	mafrel	amaron	amalech
n	pafis̄	maftril	enorion	onamech
o	pafos̄	mafrol	morifon	fomelech
p	pafus̄	maftrul	aporion	apomelech
q	pafal	maftras̄	aquilon	aquifalech
r	pafel	mafres̄	armaon	tramelech
s̄	pafil	maftris̄	ofarion	afomelech
t	pafol	mafros̄	atharon	ftomelech
u	paful	maftrus̄	cuburon	tumelech
ξ	pafar	maftraff	agion	axomelech
η	pafar	maftrereff	tymeon	pymelech
δ	paftr	maftriff	azaron	ozzifelech
iv	pafur	maftruff	puualon	iuuemelech

Table 2

Selected columns from Polygraphia III and IV: Replacement ciphers using artificial words.

3. Results: More similar than almost anything else

The hypothesis that should be tested is: Can a cipher whose mode of operation corresponds to P.III produce a text with similar statistical properties to the VMS? To do this, we would have to simulate a setting in which an encryptor has the P.III substitution tables available and uses them to generate a ciphertext to an arbitrary plaintext. What is difficult here is the simulation of human intuition: How exactly are the many available substitution ciphers selected? We have opted for random selection as a first step and generated the cipher text from all 132 substitution columns (P.III all), resp. from only 10 different columns (P.III small). Future work could simulate this closer to an early modern reality, for example, by involving test subjects as encryptors.

We will then compare the P.III-encoded text with texts of Central European languages and

Voynichese. The early modern comparative texts chosen were "The Mathematicall Praeface to Elements of Geometrie of Euclid of Megara" (English, <https://www.gutenberg.org/files/22062>) by John Dee, "La Divina Commedia" (Italian, <https://www.gutenberg.org/files/1012/>) by Dante and "Das Buch Paragranum" (German, <http://www.zeno.org/Philosophie/M/Paracelsus/Das+Buch+Paragranum>) by Paracelsus. Ancient latin is represented by Caesar's "De Bello Gallico" (<https://www.gutenberg.org/ebooks/218>). For Voynichese the complete EVA-transcription by Takahashi was used, separate values are shown for Currier A and B.[24] All texts as well as the necessary code for the generation of the PIII cipher and the calculation of the statistical properties can be found on GitHub: https://github.com/hermesj/R_Voynich_Stats.

3.1. Joint entropy

Table 3.2 shows that that the difference between h_1 and h_2 is larger in the generated P.III cipher text than it is in the early modern comparative texts. Nonetheless, it only comes close to the Voynich manuscript if one restricts the number of substitution columns (from which, however, the word entropy also suffers). This means either that Voynichese is still much more repetitive than the P.III cipher at the character level or that the selection process in the cipher needs to be worked out even better.

	Dee	Dante	Paracelsus	VMS A	VMS B	P.III big	P.III small
h_{words}	9.27	9.66	8.93	9.88	9.89	10.74	6.52
h_0	5.36	4.59	4.95	4.46	4.46	4.52	4.00
h_1	4.10	3.93	4.04	3.85	3.88	4.00	3.57
h_2	3.29	3.11	3.14	2.17	2.01	2.90	1.90
$h_1 - h_2$	0.81	0.82	0.90	1.68	1.87	1.10	1.67

Table 3

Word entropy and the three character-based entropies: h_0 is h_{max} , h_1 is the entropy of single, h_2 of joint characters. The last column gives the difference between h_1 and h_2 (see [5])

3.2. Word lengths distribution

Figure 1 illustrates the distribution of word lengths for types and tokens of the texts to be analysed. Very striking is the pronounced binomial distribution of the Voynichese, which is not matched at all by the P.III cipher and the comparison texts in the distribution of type word lengths. In the case of the P.III, this is a consequence of the word length distribution within the substitution columns. However, this could easily be replaced by a binomial distribution without affecting the semantics. Even more striking, however, is that in natural languages the curves for types and tokens are far apart, since they become very left-skewed with the tokens. This is definitely not the case for Voynichese and P.III.

3.3. Distribution of similar words

If we look at the measurement of minimal pairs (Levensthein distance 1) and directly successive repetitions of words (cf. Table 4), the high values for Voynichese emphasize the special nature

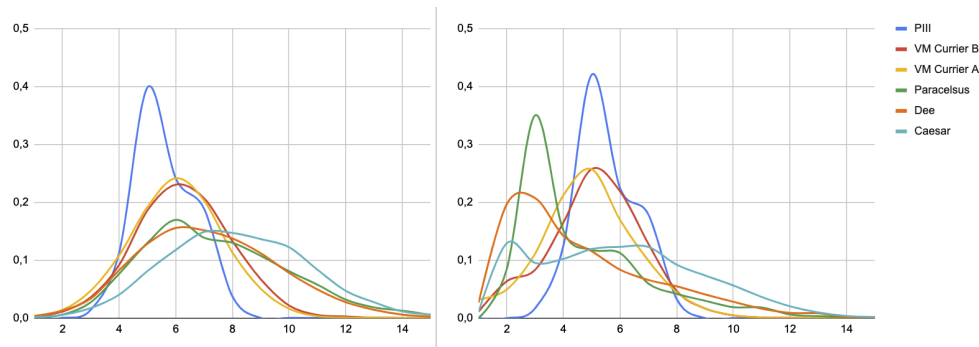


Figure 1: Distribution of word lengths (relative frequencies) for types and tokens of Voynichese (Currier A and B), P.III cipher and comparative texts of the early modern age and ancient Latin.

of this text. It can also be seen that the P.III ciphers have increased values, which can be raised almost arbitrarily if fewer replacement columns are used. Here, too, future research should demonstrate to what extent other selection procedures in ciphering produce more realistic results.

	Dee	Dante	Paracelsus	VMS A	VMS B	P.III big	P.III small
Minimal pairs	5.53	2.61	6.40	35.32	30.25	11.12	222.23
Minimal triplets	0.08	0.10	0	2.26	1.70	0.06	47.21
Word repeated	0.98	0	2.56	9.59	8.46	1.36	25.65

Table 4

Counting of adjacent minimal pairs and minimal triplets as well as directly consecutive word repetitions.

3.4. Summary of results

The experiments demonstrate that a cipher that replaces individual letters with invented words that are very similar to each other has the potential - in contrast to natural languages - to approximate the statistical properties of Voynichese. What this study still lacks is the simulation of a comprehensible selection process that can mimic a medieval encryptor. What is also missing from the study is an exploration of the possibility of inferring the underlying replacement tables from a text. This would be the foundation for a cryptanalytic attack and will be discussed in the next section.

4. Discussion: Solutions and Problems of the P.III hypothesis

To be clear, this contribution does not assume that Trithemius had anything to do with the creation of the VMS since the time of origin determined via the radiocarbon method speaks against his authorship. Additionally, the perception of the Polygraphia as somehow a blueprint-codebook for the VMS is fraught with several difficulties, of which the selection process mentioned above is only one. The effort to create something like the VMS with the help of replacement tables would have been enormous. Moreover, the hypothesis has a lot of prerequisites, such

as the assumption that a codebook has existed, even if no reference to it has never been found. Therefore, it should not be concealed here that another method has already been published that is capable of generating a text with the characteristics of the Voynich manuscript.[25] On the one hand, this method (known as the autocopist theory) has the advantage that no codebook is needed, because it claims that the text is created by adding gradual variations to an initial state. On the other side, it has the disadvantage that no information can be transported with it. If the VMS was produced as an autocopy, its glyphs only have form, no meaning.

What could be demonstrated here, however, is the existence of a cipher method in the early modern period, which – applied as a random procedure – is able to produce a text that can mimic the oblique properties of the VMS. This result is quite exciting since it brings back into play highly-debated approaches claiming the existence of hidden comprehensible information in the text of the VMS.

The question remains whether one could reconstruct the information that was encrypted with such a method. To answer it, we should look at all the ciphers from the first four parts of the Polygraphia. For the Hail Mary cipher in P.I and P.II, a reconstruction of the plaintext should be impossible without the underlying codebooks. Although functionally identical and meaning-related words were grouped in columns, these do not contain any information about the hidden plaintext letter.

The simple steganographic encoding in P.IV is considered by Trithemius himself to be insecure – but convenient to use – since one does not need the codebook for decoding.

The P.III cipher is more closely related to P.I than to P.IV, but the regular formations of the words [10] make it possible to draw conclusions about their position in the substitution tables. Above all, two rules can be assumed: Same words encode same letters, similar words encode different ones. How far one can get with this simple assumptions in a text like the one from the VMS is up for discussion.

References

- [1] G. Landini, Evidence of Linguistic Structure in the Voynich Manuscript Using Spectral Analysis, *Cryptologia* 25 (2001) 275–295.
- [2] R. Zandbergen, From Digraph Entropy to Word Entropy in the Voynich MS, 2022. URL: <http://www.voynich.nu/extra/wordent.htm>, accessed 25.08.2022.
- [3] J. Stolfi, A Prefix-Midfix-Suffix Decomposition of Voynichese Words, 1997. URL: <http://www.ic.unicamp.br/stolfi/voynich/97-11-12-pms/>, accessed 25.08.2022.
- [4] W. R. Bennett, *Scientific and Engineering Problem-Solving with the Computer*, Prentice Hall PTR, Upper Saddle River, NJ, 1976.
- [5] D. Stallings, Understanding the Second-Order Entropies of Voynich Text, 1998. URL: <http://ixoloxi.com/voynich/mbpaper.htm>, accessed 25.08.2022.
- [6] L. Lindemann, C. Bowern, Character Entropy in Modern and Historical Texts: Comparison Metrics for an Undeciphered Manuscript, 2021. URL: <http://arxiv.org/abs/2010.14697>, accessed 25.08.2022.
- [7] C. Bentz, R. Ferrer-i Cancho, Zipf’s law of abbreviation as a language universal, *Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics*. University

- of Tübingen (2016). URL: <https://publikationen.uni-tuebingen.de/xmlui/handle/10900/68639>, accessed 25.08.2022.
- [8] P. Grzybek, History and Methodology of Word Length Studies. The State of the Art, in: P. Grzybek (Ed.), Contributions to the Science of Text and Language. Word Length Studies and Related Issues, Springer, Dordrecht, NL, 2006, pp. 15–90.
 - [9] J. Stolfi, On the VMS Word Length Distribution, 2000. URL: <http://www.ic.unicamp.br/stolfi/voynich/00-12-21-word-length-distr/>, accessed 25.08.2022.
 - [10] J. Hermes, Textprozessierung - Design und Applikation, PHD thesis, Universität zu Köln, 2012.
 - [11] T. Timm, A. Schinner, A possible generating algorithm of the Voynich manuscript, Cryptologia 44 (2020) 1–19.
 - [12] M. D’Imperio, The Voynich Manuscript—An Elegant Enigma, Aegean Park Press, Langley, 1978.
 - [13] W. F. Friedman, E. Friedman, Acrostics, Anagrams, and Chaucer, Philological Quarterly 38 (1959).
 - [14] J. Wilkins, An Essay Towards a Real Character, Gellibrand, London, 1668. Reprint Menston: The Scholar Press 1968.
 - [15] G. Dalgano, Ars Signorum, Self print, London, 1661. Reprint Menston: The Scholar Press 1968.
 - [16] R. Zandbergen, Radio-carbon dating of the Voynich MS, 2022. URL: <http://www.voynich.nu/extra/carbon.html>, accessed 25.08.2022.
 - [17] J. Trithemius, Polygraphia libri sex, Haselberg, Basel, 1518.
 - [18] J. Trithemius, Steganographia, Berner, Frankfurt, 1606.
 - [19] K. Arnold, Johannes Trithemius (1462-1516), Schöningh, Würzburg, 1971.
 - [20] G. Strasser, Lingua Universalis: Kryptologie und Theorie der Universalsprachen im 16. und 17. Jahrhundert, number 38 in Wolfenbütteler Forschungen, Harrassowitz, Wiesbaden, 1988.
 - [21] T. Ernst, Schwarzweiße Magie. Der Schlüssel zum dritten Buch der Steganographia des Trithemius, Daphnis (1996) 1–205.
 - [22] J. Reeds, Breakthrough in Renaissance Cryptography. A Book Review, Cryptologia 23 (1999) 59–62.
 - [23] M. Gamer (Ed.), Die Polygraphia des Johannes Trithemius nach der handschriftlichen Fassung (Band 1): Edition, Übersetzung und Kommentar, Brill, 2022.
 - [24] P. Currier, Papers on the Voynich Manuscript, 1976. URL: http://www.voynich.nu/extra/curr_main.html, accessed 25.08.2022.
 - [25] T. Timm, How the Voynich Manuscript was created, arXiv:1407.6639 [cs] (2015). URL: <http://arxiv.org/abs/1407.6639>, arXiv: 1407.6639.