Fingerprinting Gibberish: A Quantitative Comparison of the Voynich and Sloane MS 3188

Alexander Boxer

Abstract

Texts that resemble natural language but convey no semantic content — that is, gibberish — still exhibit statistical fingerprints allowing them to be identified from other, superficially similar texts. This paper summarizes the results of a quantitative comparison between the Voynich and the extended passages of "Enochian" transcribed by John Dee and Edward Kelley in Sloane MS 3188. Preliminary to this analysis, and by far its principal labor, is a new transcription of the Voynich text which I have prepared and made freely available on GitHub. Using this transcription, and a transcription of the relevant sections of Sloane MS 3188, several important similarities can be observed. Most notably, both the Voynich and Enochian exhibit an unusual degree of epizeuxis, or token repetition, including triplets. However, the running fraction of hapax legomena, a key statistical fingerprint which distinguishes Enochian from natural language texts, also distinguishes Enochian from the Voynich. Our analysis therefore suggests that even if the Voynich is gibberish, it does not belong the same family of gibberish as Enochian.

Keywords

voynich, enochian, gibberish, natural language, transcription, epizeuxis, hapax legomena

1. Introduction

This article stems from a larger, independent effort to explore the Voynich manuscript using modern computational methods. The core of this effort is a new, machine-readable transcription of the Voynich manuscript. A secondary component has been the collection and curation of machine-readable texts in various languages to facilitate comparisons with the Voynich. The initial tranche of comparison texts is weighted toward documents in Latin, German, and other European languages, primarily from the late medieval and early modern periods. Certain texts in other languages have been examined as well, especially if these are deemed to have some bearing on the Voynich mystery. This is the case with Sloane manuscript 3188, which was composed by John Dee and Edward Kelley between 1581 and 1583, and which contains extended passages written in an otherwise unknown language called Enochian.

A connection between Dee and the Voynich manuscript was first proposed by Wilfrid Voynich in 1921 [1]. Moreover, since Dee's associate, Kelley, is assumed to have invented the Enochian language, the suggestion that Kelley, Dee, or both, was somehow involved in creating the Voynich manuscript, perhaps as an elaborate hoax to sell to the emperor Rudolf II, has been explored by several authors [2, 3, 4]. It is not the intent of this paper to evaluate that hypothesis. Rather, this paper sets out to showcase the utility of our transcription, along with the computational

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

alex@idolsofthecave.com (A. Boxer)

https://alexboxer.com (A. Boxer)

tools developed to anlayze it, by providing a quantitative comparison of the Voynich and Sloane MS 3188. The results of this comparison indicate that the statistical properties of the Voynich manuscript differ substantially from those of the Enochian sections of Sloane MS 3188.

2. Voynich-Attack

voynich-attack is a public GitHub repository containing resources and tools developed to facilitate an exploration the Voynich manuscript [5]. The primary resource is a new, freely shareable transcription of the Voynich manuscript. The transcription is formatted as a comma-separated values (.csv) file and consists of 33,669 word-like tokens comprising 147,485 characters.

All pages of block-text are included (197 manuscript pages); pages dominated by illustrations or where the text cannot be reduced to an ordered sequence are omitted (36 manuscript pages). The transcription was hand-typed from the digitized edition of the Yale University Library Digital Collections [6]; in a few instances, where finer detail was required, the 2016 facsimile edition served as a guide [7].

The impetus to create a new Voynich transcription arose entirely from personal curiosity. The effort was not undertaken to supplant or impugn any of the existing transcriptions of the manuscript. All transcriptions of the Voynich manuscript must deal with the challenges of determining which characters are distinct and which are variants, and whether certain sequences of characters should be grouped as one token or several. Accordingly, all transcriptions of the Voynich, ours included, possess an unavoidable element of subjectivity. For this reason, the convenience of a single, sanctioned transcription of the manuscript may be illusory; instead, the community of Voynich researchers may be better served by multiple, independent transcriptions.

2.1. The RefText class

In addition to a transcription of the manuscript, the `voynich-attack` repository contains corpora of curated texts for comparison against the Voynich, as well as computational tools to facilitate general exploration. For investigations in the Python programming language, a `RefText` class is defined which allows all of the curated texts within the `voynich-attack` repo, including the Voynich transcription, to be analyzed as convenient `RefText` objects. (Note: the use of the `RefText` class is entirely optional; the Voynich transcription is a simple .csv file which can be examined using whatever tools the analyst sees fit.)

Accessing `RefText` objects is achieved simply by importing them from the `corpora.py` module in any Python script. For example, an examination of the Voynich manuscript and Julius Caesar's *Commentarii de Bello Gallico* would include the following import statement:

```
from corpora import vms, caesar
```

Here, `vms` and `caesar` are the `RefText` objects of their respective texts. A `RefText` object contains just a handful of attributes. The `charlist` attribute is an ordered list of all the characters of the text, and the `tklist` attribute is an ordered list of all the tokens. Thus, `caesar.charlist` returns `['g', 'a', 'l', 'l', 'i', 'a', 'e', 's', 't', ...]` and `caesar.tklist` returns `['gallia', 'est', 'omnis', 'divisa', ...]`

Wherever possible, the source text of a `RefText` object is stored as a `Pandas` DataFrame. `Pandas` DataFrames are the standard way spreadsheet-like data is represented in Python, and they support many powerful and fast query operations. The DataFrame representation of a `RefText` source text is accessed via the `df` attribute. Thus, `caesar.df` returns a DataFrame (spreadsheet) of Caesar's *Commentarii* with each row of text indexed by its book, chapter, and line number. Similarly, `vms.df` returns a DataFrame of the Voynich transcription, with each row of text indexed by its page, paragraph, and line number, and with individual tokens occupying individual columns within their row.

A structured data type such as a DataFrame has many advantages over a simple text string. DataFrames allow the results of a query — say, the pattern matches of a regular expression `findall` operation — to be easily referenced back to their locations within the text. DataFrames also facilitate more directed queries, for instance, compiling statistics on tokens appearing at the beginnings or ends of lines. In order to format a text as a DataFrame, it is helpful to start with a structured representation. The `xml`-formatted texts of the Deutsches Textarchiv (for German-language texts) and the Perseus Digital Library (for Latin and Greek texts) have been particularly useful in this regard [8, 9].

`RefText` objects also have two built-in methods, `chardf` and `tkdf`, which generate character or token n-grams of any rank. These methods take an optional rank argument (for ranks greater than one) and return a `Pandas` DataFrame of n-gram statistics. For example, the following line of code returns the top 5 tokens of the Voynich transcription, their counts and percentages:

```
In [1]: vms.tkdf().head(5)
Out[1]:
```

	gram	n	pct
0	8,a,m	794	2.36
1	cc,c,8,9	487	1.45
2	o,x	485	1.44
3	c^c,c,8,9	417	1.24
4	a,m	379	1.13

Similarly, to obtain the top 5 bigrams of Caesar's *Commentarii*:

```
In [2]: caesar.chardf(2).head(5)
Out[2]:
```

	gram	n	pct
0	e-r	7416	2.34
1	u-m	5752	1.82
2	q-u	5525	1.75
3	t-i	5487	1.73
4	i-s	5479	1.73

An especially useful feature of the `RefText` class is the ability to quickly instantiate `RefText` objects on the fly. For instance, to examine the statistics of just the botanical sections of the

Voynich, one can easily create a subset of the Voynich DataFrame by selecting just the relevant manuscript pages. This subset DataFrame can be converted into a `RefText` object using the appropriate function from the `reftext.py` module. In the following example, a `RefText` object name `plants1` is instantiated from a DataFrame named `plants1_df`.

```
import reftext

plants1 = reftext.from_dataframe(
        plants1_df,
        language = 'voynich',
        read_from_col = 3,
        comma_split_tokens = True
        )
```

From here, `plants1` can be analyzed using the attributes and methods of the `RefText` class. Notably, several interesting observations result from comparing the token n-grams of the Voynich subsections against each other, although a fuller discussion of this remains outside the scope of this paper.

3. Sloane MS 3188 and the Enochian Language

During the later decades of the 16th century, John Dee, English mathematician and sometime advisor to Queen Elizabeth I, made a concerted attempt to communicate with angels. To this end, Dee employed a crystal-ball reader, or scryer, named Edward Kelley. Dee's records of the spiritual visions and messages received by Kelley — but never by Dee — are preserved in several manuscripts currently housed in the British Library. Sloane MS 3188 is the earliest of these, covering the period from December 1581 to May 1583.

Starting in March 1583, Dee's notes contain dictations from a book said to be written in the angels' own language, a "Caelestiall speche" previously revealed only to Enoch and a few other biblical figures [10]. Although not named so by Dee or Kelley, subsequent literature refers to the language of these dictations as Enochian. The linguistic characteristics of Enochian were examined by Laycock [11] who observes:

> Statistical studies in linguistics show that patterning of this nature is rare in normal language — though it is found in poetry and magical charms. It is also characteristically found in certain types of meaningless language (such as glossolalia), which is often produced under conditions similar to trance . . . All the facts seem compatible with Kelley pouring out a string of gibberish while in a trance state.

3.1. Relevance to the Voynich Manuscript

There are two ways an analysis of Enochian can contribute to an understanding of the Voynich manuscript. The first concerns the theory, currently out of favor, which asserts that Kelley and

or Dee is the author of the Voynich. It is not the intent of this paper to address this hypothesis in any detail, except to report that our analysis did not discern any compelling statistical patterns of the sort that would indicate the Voynich and Enochian are directly linked.

The second way an examination of Enochian can offer insight into the Voynich is to recognize Enochian as an important, extended sample of gibberish, one whose composition is roughly contemporaneous with that of the Voynich. Consequently, Enochian provides a highly relevant point of reference in assessing whether the Voynich encodes a natural language text or not.

3.2. Source Text Description

The sample of Enochian examined in this article is that contained in Sloane MS 3188. This comprises the first two pages of the angelic book supposedly dictated to Kelley by the archangel Raphael. The book, in its entirety, was to consist of 49 leaves (98 pages), each containing a 49×49 grid of text for a grand total of 235,298 textual units.

Initially these units were word-like and dictated letter-by-letter using a 21-letter angelic alphabet. This scheme, however, was abandoned for being too cumbersome after only the second row of text. Subsequent rows were not spelled-out but spoken aloud by Kelley and transcribed by Dee using Roman letter equivalents. Yet even this abbreviated approach did not remain consistent. By the conclusion of the book's second page, the 49 units of each row had become simply 49 individual letters.

The remaining 48 leaves of the book were transcribed by Kelley without Dee present and are preserved as Sloane MS 3189. A collection of additional Enochian dictations is preserved in Sloane MS 3191. Each of these texts is probably best regarded as a distinct variant of Enochian. Indeed, Laycock refers to the exemplar from Sloane MS 3191 as the "true Enochian language" which he contrasts with the earlier "angelic" language of Sloane MSS 3188 and 3189 [12].

There exists a digitization of Sloane MS 3188 available online from the British Library [13]. For the present paper, however, the annotated edition of Whitby [10] served as the reference for a hand-typed transcription. This transcription has been uploaded to the `voynich-attack` repository. It can be downloaded as a .csv file, or imported into a Python script as a `RefText` object using: `from corpora import enoch`. Altogether, the transcription consists of 4,079 words comprising 19,087 letters.

4. Analysis of Sloane MS 3188 and Comparison with the Voynich

4.1. Qualitative Observations

A closer examination of Enochian was motivated by the observation that some of the unusual linguistic features found in the Voynich can also be seen in Enochian. Noteworthy among these are: epizeuxis, alliteration, and compounding.

Epizeuxis: One the most vexing questions for any attack on the Voynich is: Do the word-like tokens of the Voynich actually encode words? A major obstacle in identifying the Voynich tokens with words is that several hundred of these, or nearly 1% of all token pairs, are direct repeats, e.g., cc,o,x - cc,o,x. The technical term for this is *epizeuxis*. The rate of epizeuxis in a natural language such as Latin or German is typically several orders of magnitude smaller. (In

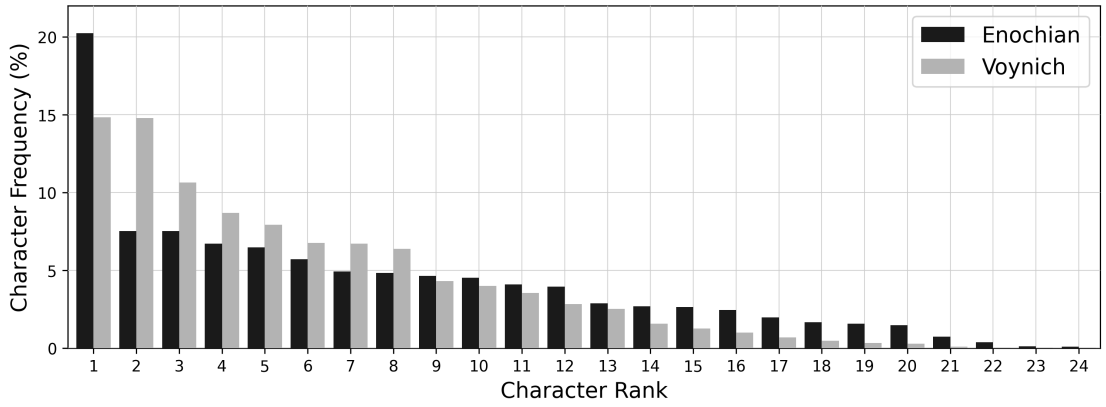

Figure 1: Frequency Distributions of Enochian and Voynich characters. The Enochian text from Sloane MS 3188 exhibits a frequency distribution of characters which differs markedly from that of the Voynich.

Caesar's *Commentarii* there are no instances at all.) The Enochian of Sloane MS 3188 represents an intermediate case, with an epizeuxis rate of 0.15%.

Even more striking is the presence of epizeuxis triples. The Voynich contains ten such instances, for example: o,N,a,m - o,N,a,m - o,N,a,m. Similarly, there are Enochian triples in Sloane MS 3188, e.g., "Ádgmach ádgmach ádgmach." Note that in Enochian, the tokens are explicitly words; here, "ádgmach" is glossed as "much glory" [14].

Alliteration: As noted by Laycock, many highly alliterative passages in Sloane MS 3188 are more indicative of glossolalia, or speaking in tongues, than a natural language. For instance: "gasmat gasque gasla gasna gasmaphes gasmagel gasnunabe" [11]. Similarly, the Voynich contains multiple sequences of six or even seven tokens in a row that each begin with the prefix-like characters "4,o".

Compounding: As mentioned previously, it can be difficult to determine whether a sequence of characters in the Voynich manuscript should be grouped into one or split into several tokens. One reason for this is that many sequences — "o,x" being a prominent example — are just as likely to be found compounded with other characters, e.g., "cc,o,x" as they are standing on their own. A very similar phenomenon is seen in Sloane MS 3188. Indeed, Dee himself on several occasions expresses confusion on how to group the syllables he has transcribed, knowing that the total number of words per line needed to sum to 49 [15].

4.2. Character Analysis

Given the qualitative similarities observed between the Voynich and Sloane MS 3188, the question arises as to whether these texts exhibit deeper similarities at the statistical level. Our analysis indicates that they do not. Most immediately, Figure 1 shows a marked divergence in the frequency distribution of the characters appearing in the Voynich and in the Enochian sections of Sloane MS 3188.

The Enochian distribution also differs radically from typical natural language distributions (not shown). The disproportionate dominance of the most common Enochian letter "a" accords

with the theory that Enochian is constructed from compounds of glossolalia-like monosyllables meant to be spoken aloud. While the characters of the Voynich also do not align with known natural language distributions, neither do they exhibit this strong statistical fingerprint of Enochian.

4.3. Token Analysis

A more significant divergence between the Voynich and the Enochian of Sloane MS 3188 can be seen at the token level. The technical term for a token that appears only once within a text is *hapax legomenon*. Figure 2 shows the running fraction of hapax legomena for the Voynich, the Enochian sections of Sloane MS 3188, and several natural language texts. For a given text, each curve shows, token-by-token, the fraction of tokens which have appeared only once.

Among these curves, Enochian is the clear outlier, exhibiting a running fraction of hapax legomena that remains well over 50% for the entirety of the text. Moreover, of the 2,716 unique Enochian tokens in Sloane MS 3188, a remarkable 86% are hapax legomena. These numbers suggest strongly that there is no semantic content contained in the underlying text, and buttress the theory that the Enochian of Sloane MS 3188 is indeed gibberish.

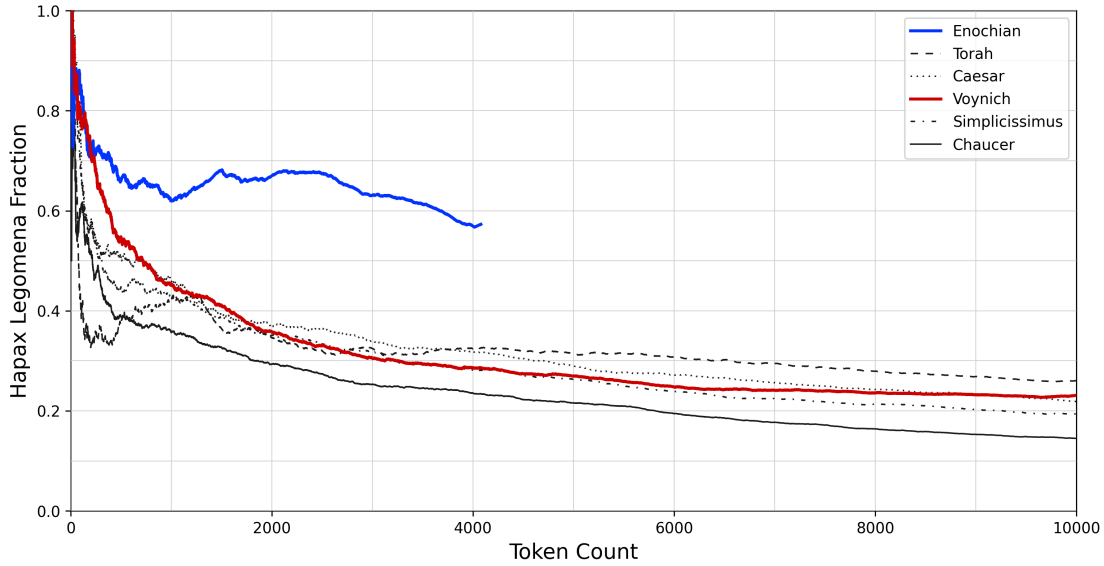

Figure 2: Running Fraction of Hapax Legomena for Selected Texts: the Enochian sections of Sloane MS 3188 (Enochian); Torah (Hebrew); Caesar's *Commentarii de Bello Gallico* (Latin); Voynich Manuscript; Grimmelshausen's *Der abenteuerliche Simplicissimus Teutsch* (German); Chaucer's *Canterbury Tales* (Middle English). The running fraction of hapax legomena is effective in discriminating Enochian from natural language texts. By this metric, the Voynich appears more similar to natural language than to Enochian.

The most intriguing observation from Figure 2, however, is that the Voynich does not exhibit this particular statistical fingerprint for gibberish. On the contrary, the Voynich, at least as measured by its running fraction of hapax legmomena, appears more similar to natural language

than it does to Enochian. This observation does not, by itself, provide evidence that the Voynich is or is not gibberish. Nevertheless, it does establish that the Voynich cannot be gibberish in the same way that the Enochian of Sloane MS 3188 is gibberish. Note that this assessment runs counter to that of Daruka [4], who, however, based his analysis primarily on the Enochian of Sloane MS 3189.

5. Conclusion

This paper examined the Enochian text of Sloane MS 3188 in an effort to better understand the Voynich manuscript. Enochian is a historically significant corpus of gibberish text which provides an important point of reference in assessing whether or not the Voynich is also gibberish.

To analyze the Voynich and Sloane MS 3188, resources and tools from the public GitHub repository voynich-attack were utilized, including a new, freely-shareable transcription of the Voynich manuscript.

Although the Voynich and Enochian exhibit certain qualitative similarities — notably a propensity for epizeuxis, alliteration, and compounding — our analysis revealed substantive, statistical differences between the two. The running fraction of hapax legomena was identified as an effective statistical fingerprint for distinguishing Enochian from natural language texts. The Voynich does not exhibit this statistical fingerprint, and by this metric appears more similar to natural language than to Enochian. This observation eliminates one type of gibberish to which the Voynich can belong. Specifically, it argues against they hypothesis that the Voynich encodes glossolalia-like monosyllables as does the Enochian of Sloane MS 3188; it argues even more strongly against the hypothesis that the Voynich is Enochian expressed in non-Roman characters. Our observation does not eliminate the possibility that the Voynich may belong to another class of gibberish which is substantively different from the Enochian of Sloane MS 3188.

References

[1] W. Voynich, A preliminary sketch of the history of the roger bacon cipher manuscript, Transactions of the College of Physicians of Philadelphia 43 (1921) 415–30. URL: https://archive.org/details/transactionsofco3431coll/page/414/mode/2up.

[2] R. S. Brumbaugh, Science in cipher, in: R. S. Brumbaugh (Ed.), The Most Mysterious Manuscript: The Voynich "Roger Bacon" Cipher Manuscript, Southern Illinois University Press, Carbondale, IL, 1978, pp. 128–30.

[3] G. Rugg, An elegant hoax? a possible solution to the voynich manuscript, Cryptologia 28 (2004) 31–46.

[4] I. Daruka, On the voynich manuscript, Cryptologia 45 (2021) 44–80. doi:10.1080/01611194.2019.1706063.

[5] A. Boxer, voynich-attack, GitHub, 2022. URL: https://github.com/alexanderboxer/voynich-attack.

[6] Yale University Library, Cipher manuscript, 2022. URL: https://collections.library.yale.edu/catalog/2002046.

[7] R. Clemens (Ed.), The Voynich Manuscript, Yale University Press, New Haven, CT, 2016.
[8] Berlin-Brandenburgische Akademie der Wissenschaften, Deutsches Textarchiv, 2022. URL: https://www.deutschestextarchiv.de/.
[9] G. Crane (Ed.), Perseus digital library, 2022. URL: http://www.perseus.tufts.edu/hopper/.
[10] C. Whitby, John Dee's Actions With Spirits: 22 December 1581 to 23 May 1583, Volumes 1 and 2, volume 2, Garland Publishing, New York, NY, 1988, p. 232.
[11] D. C. Laycock, The Complete Enochian Dictionary, Weiser Books, San Francisco, CA, 2001, pp. 33–5.
[12] D. C. Laycock, 2001, p. 39.
[13] British Library, Sloane MS 3188, 2022. URL: https://www.bl.uk/manuscripts/FullDisplay.aspx?ref=Sloane_MS_3188.
[14] C. Whitby, 1988, pp. 283-4.
[15] C. Whitby, 1988, pp. 266, 278.