Enciphered after all? Word-level text metrics are compatible with some types of encipherment

Claire L. Bowern [1] and Daniel E. Gaskell [1]

[1] *Yale University, PO Box 208236, New Haven, CT 06520, USA*

Abstract

Many Voynich manuscript analyses have relied on statistical properties of the text to distinguish enciphered natural language from non-language (or gibberish). Schinner (2007) and Rugg & Taylor (2016) have argued that Voynichese is unlikely to be natural language because of its extreme predictability. Conversely Bowern & Lindemann (2021), Sterneck et al. (2021), Layfield (2021) and others focus on topic modeling and larger textual units, showing that beyond the paragraph level, Voynichese has many properties in common with enciphered natural language. The question then becomes whether one can discover ciphers that produce the textual characteristics that make Voynichese unusual at the word level, while preserving topic structure across a larger sample. To this end, we investigate the statistical properties of 22 methods of textual manipulation on a sample of historical and contemporary texts. For consistency of comparison we use the same metrics as <Voynich submission>. While many historical encipherment methods (such as substitution ciphers) are phonological structure-preserving (and therefore not tested here), others, such as the Crema cipher, are not. Results show that there are multiple types of encipherment which reduce conditional entropy; the encoding of multiple phonemes (or orthographic characters) as bigraphs, for example, lowers character entropy to the levels seen in Voynichese, for Latin-encoded texts. Adding null characters (in some patterns) also increases predictability of word formation. While such results do not "prove" that the Voynich manuscript is enciphered, it indicates that the unusual word-level predictability highlighted in previous work is not conclusive evidence that the Voynich manuscript is gibberish.

Keywords[1]

Enciphered Hypothesis, Hoax Hypothesis, Linguistics, Natural Language Processing

1. Introduction

Voynich manuscript analyses such as Reddy & Knight [1], Montemurro & Zanette [2], and Amancio et al. [3] have relied on statistical properties of the text to distinguish enciphered natural language from non-language (or gibberish). Schinner [4] and Rugg & Taylor [5] have argued that Voynichese is unlikely to be natural language because of its extreme predictability. Conversely Bowern & Lindemann [6], [7], Sterneck et al. [8], Layfield [9] and others focus on topic modeling and larger textual units, showing that beyond the paragraph level, Voynichese has many properties in common with natural language. This implies that Voynichese is a cipher of a natural language. Properties of linguistic systems are defined by Hockett [10], among others. For the purposes of comparison with non-language in this task, the properties most crucial to a linguistic system are that there are words with arbitrary form-meaning correspondences, and they combine to form sentences in consistent ways that impart meaning. That is, linguistic systems have phonology, morphology, semantics, and syntax. They have

[1]International Conference on the Voynich Manuscript 2022, November 30--December 1, 2022, University of Malta.
EMAIL: daniel.gaskell@yale.edu (A. 1); claire.bowern@yale.edu (A. 2)
ORCID: 0000-0002-0306-7943 (A. 1); 0000-0002-9512-4393 (A. 2)

'words' that have predictable meanings in particular contexts; those words combine to form sentences, and someone who knows the system can recover meaningful information from the text [11]. Gibberish, in contrast, has no such underlying meaning and no formal information-carrying syntax; it has only the phonology which one can deduce from the forms of words in the sample. Voynichese phonology is very unusual, showing many features that set it apart from a wide array of natural language samples. Language can be enciphered in such a way as to preserve recoverability of the morphology, semantics, and syntax, while making the phonology appear distinct from natural languages. The question then becomes whether one can discover ciphers that produce the textual characteristics that make Voynichese unusual at the word level, while preserving topic structure across a larger sample.

To this end, we investigate the statistical properties of 22 methods of textual manipulation (including encipherment) on a sample of historical, contemporary, and constructed texts, along with the 'gibberish' construction method of Timm and Schinner [4]. For consistency of comparison, we use the same metrics as <other submitted paper>. 'Encipherment' methods here range from text-destroying to text preserving. Lindemann and Bowern [6], [7] show that character entropy measures are roughly equivalent across languages with different numbers of phonemes and writing systems; we therefore focus on methods that increase character predictability across words, such as removing phonemic contrasts. While many historical encipherment methods (such as substitution ciphers) are structure-preserving (and therefore not tested here), others, such as the Crema cipher described by Gabriele de Lavinde, are not.[2]

Results show that there are multiple types of encipherment which reduce conditional entropy; the encoding of multiple phonemes (or orthographic characters) as bigraphs, for example, lowers character entropy to the levels seen in Voynichese, for Latin-encoded texts. Adding null characters (in some patterns) also increases predictability of word formation. The full paper surveys a wider range of metrics. While such results do not "prove" that the Voynich manuscript is enciphered, it indicates that the unusual word-level predictability highlighted in previous work is not conclusive evidence that the Voynich manuscript is gibberish.[3]

The main questions examined in this paper are as follows:

1) What effects do textual manipulations (including encipherment) have on statistical measures of text?
2) What types of manipulations (if any) produce outcomes that are similar to Voynichese?
3) Is there an interaction between language and manipulation?
4) Are there manipulations which produce a set of markers which is compatible with Voynich text?

2. Methods

2.1. Underlying dataset

We begin with a range of documents in natural and constructed languages. These samples comprise a subset of the Latin alphabet Wikipedia and historical document corpora used in [6], along with the constructed language (conlang) and gibberish samples from <Voynich submission>.

[2] The proceedings of HistoCrypt (e.g. [12], [13]) only became available after the work for this paper was already substantially complete. We plan future work that enciphers a broader range of historical texts through the methods discussed in the DECODE project.

[3] As a reviewer points out, there is a risk of circularity if we select text manipulation metrics in order to find statistics that resemble Voynich metrics, and then argue that such encipherment techniques were used *because* they mimic the characteristics of Voynichese. We emphasize that in this exploratory work, we are not directly attempting to decode the Voynich manuscript. Rather, we are asking *how* certain text manipulations affect text, in comparison to plain text and gibberish. The strongest claim we make is that some text manipulations do produce some results that are consistent with Voynich text.

Because the appropriate transcription and interpretation of the VMS glyphs is uncertain [6], [14], we included five different transcriptions or subsets in our VMS corpus: Glen Claston's minimally-decomposed v101 transcription [15]; Takeshi Takahashi's maximally-decomposed EVA Full and EVA Basic transcriptions [16]; and the Currier A and B subsets in EVA Basic, as identified in Jorge Stolfi's interlinear file v16e6 [17]. This allows us to treat the properties of the VMS as a distribution of possibilities rather than assuming which method is correct, as well as to compare the Currier hands.[4]

2.2. Text manipulation

In order to gauge the effects of encipherment on underlying text in different languages (point 3 above), we employ a range of text manipulations. Because the texts are orthographic (rather than phonemic), different text manipulations do not have an impact on the text in the same way. For example, consider a manipulation that removes vowels. For an abjad orthography (such as is used for Hebrew, Arabic, or Aramaic), such a manipulation would have no effect, as the vowels are already not represented[5] in the writing system. Conversely, other manipulations should affect all languages equally. For example, adding an identical sequence of letters to the end of each word will increase predictability in all texts.

Text manipulations can be text-preserving or text-destroying. A manipulation that replaces all vowels with V and all consonants with C, for example, is text-destroying, as it is impossible to recover the message from the resulting "encipherment". Conversely, a simple substitution cipher (where a=1, b=2, etc.) is text-preserving; in fact, it is sufficiently "text-preserving" that it performs identically to unenciphered text on the metrics typically used to examine Voynichese.[6]

Table 1
Description of text manipulations

Variable	Definition
2v6c	Collapse of alphabet to 2 vowels and 6 consonants (preserving vowel frontness/backness, and consonant place/manner)
addB	Turn all letters to bigraphs by adding B after every letter
affix	Neutralises common prefixes to qo[7]
alf	Add "c9" or "B" to the end of every word (depending on the last letter)
alfb	Add "c9" to the end of the word if it ends in a vowel; else B9 or D9
bigraph1	Make letters into bigraphs based on shape (a > ci, b > lo, d > cl, g > cj, l > lc, etc)
bigraph2	Turn the most frequent[8] letters into bigraphs (a > ci, e > cc, o > ic, t> ch, s > sh, n > uc)
bigraph3	distributes the Latin alphabet across 20 2-letter sequences.
conflc	Conflate most common consonants to single letter
Crema	Implements the Crema cipher: reverse alphabetic cipher with multiple nulls

[4] An alternative option would be to divide by Davis' identification of hands [18].
[5] Unless added with diacritics.
[6] Simple substitution ciphers are therefore not further discussed here.
[7] These are based on common Latin prefixes
[8] Frequency is based on English, but the most frequent letters identified here are also among the most frequent in all the writing systems of major European languages.

mCClass	Group consonants by manner: replace all stops with p, all fricatives with k, all approximants with t, and turn glides to vowels (and delete geminates)
mvCClass	as MCClass, but also reduce the number of vowels to two.
No_Voicing	conflate voicing distinctions
No_Vowels	removed vowel letters (a, e, i, o, u + accented)
pCClass	Group consonants by place: replace all labials with p, all velars with k, all apicals with t, delete doubled letters, remove h
Sort_Alph	sorts the letters of a word into alphabetical order
spl	one symbol for the first half of the alphabet, another for the second
Two_Vowels	Replace i and u with "i" and all other vowels with "a"
u_vowels	replace all vowels with "u"
VC1	Replace all vowels with one character, all consonants with another
VC2	Replace vowels with one of two characters, consonants with a different two characters
wdb	add text at word boundaries (a variant of the "alf" and "alfb" methods)

The text manipulations tested here are given in Table 1. They were chosen to illustrate a variety of encipherment techniques [13] as well as structure preserving and destroying manipulations. Text manipulation was achieved through bespoke functions which replaced text by regular expression in R [19]. Note that the choices made here are not meant to include all potential encipherment methods; they are aimed to test the relationship between types of encipherment methods, natural language, and the statistical measures typically used to investigate Voynich text.[9] We leave exhaustive reviews of 15th century ciphers to future work but note for now that such ciphers can be monotonic (single associations between plaintext and ciphered characters), homophonic (where multiple cipher characters are associated with a single plaintext item), or polyphonic (where multiple plaintext characters are associated with a single cipher character); cf. Lasry et al 2020. Encipherment methods also include encoding sequences larger than a single character (bigrams or syllables, for example), or where multiple plaintext characters are encoded by a single cipher character. This last type of cipher is of particular interest here because it increases predictability of character transitions (thus decreasing H2 conditional entropy measures).

All texts in this dataset use the Latin alphabet, but some use diacritics, which were removed before processing. As a reviewer notes, there is considerable variety in transcription and spelling of medieval texts. We would add that in addition, some orthographies are straightforwardly phonemic (that is, that differences in characters represent differences in sounds of the spoken language), while others have more distance between the written representation of the language and the phonemic distinctions in speech. The results here are manipulations of *orthography*, not of phonology. This distinction is unlikely to affect the results, especially at this exploratory stage.

2.3. Statistical tests

42 statistical parameters were calculated for each document. These are summarized in Gaskell and Bowern (2022); they include metrics for character skew across words; entropy,

[9] We welcome discussion on other plausible types of encipherment to investigate.

proportions of repeated characters, positional word biases, and unique words. We use the same metrics in order to compare enciphered materials with constructed examples and samples of natural language.

To neutralize the effects of sample length, variables were calculated on randomized 200-word excerpts from each document, taking the mean value over 100 iterations. Variables yielding two-tailed distributions (e.g., word lengths) were described using three parameters: mean, standard deviation, and skew. Variables yielding one-tailed distributions (e.g., rank-ordered character frequencies) were described using two parameters: the maximum observed value and a shape parameter β obtained by sorting values in rank order (highest values first) and fitting them to the exponential function

$$f(x) = \frac{e^{\frac{-x}{\beta}}}{\beta}, \tag{1}$$

where x is rank (as an integer).

In order to gauge whether a transformation is a plausible match for Voynich character metrics, we measure the Euclidean distance between Voynich (EVA) samples and other samples in the text. Euclidean distance is calculated as

$$\sqrt{\Sigma(a_i - b_i)^2} \tag{2}$$

Where $_i$ is each point of comparison (in our case, each text metric). Metrics were Z-scored (each point is subtracted from the mean and divided by the standard deviation) in order to normalize scores across comparisons.

3. Results

Figure 1 gives a heatmap which shows the similarity of documents and encipherment methods. Items were z-scored to normalize across metrics. Lighter, more yellow colors indicate items that are closer to the Voynich Manuscript combined EVA sample, whereas darker (and bluer) squares are those that are further away. The normalization methods means that all statistics contribute roughly equally (though some are correlated); this should be considered a rough metric of similarity at this stage.

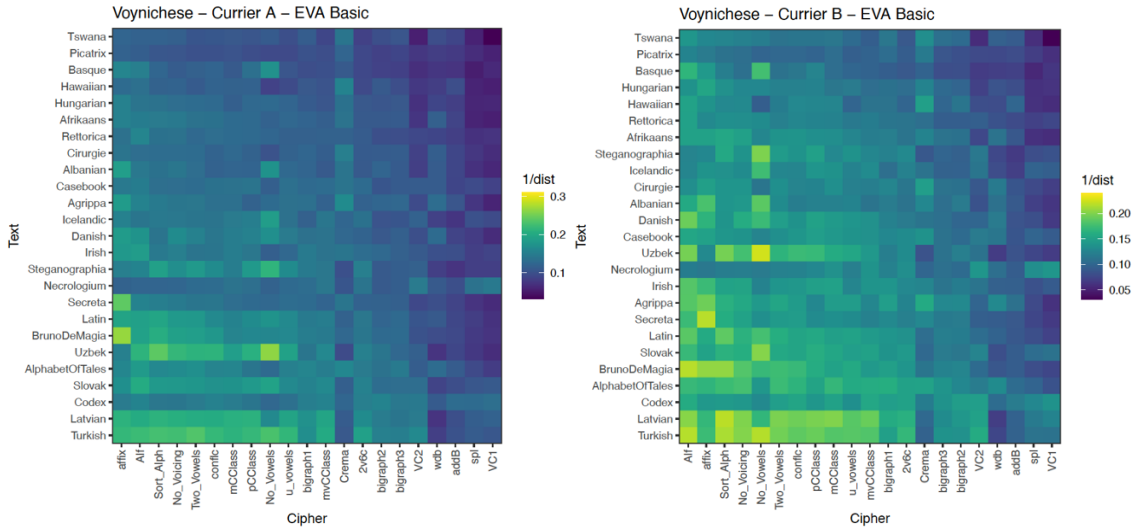

Figure 1: Heatmap of the Euclidean distances from Currier A (left) and Currier B (right).

Currier B produces closer matches than Currier A. The closest matches are with the text manipulations that remove phonemic contrasts (such as voicing, place, or manner), that replace

prefixes with a single sequence, that remove vowels, or that sort letters in words into alphabetical order. There is also an interaction between language and encipherment method. Though the vowelless examples produce closer distances than replacing all vowels and consonants with a single character (VC1), Uzbek and Turkish are much closer on this method (to Currier B) than Tswana or Hawaiian. Note also that the text also appears to make a difference; the vowelless *Steganographia* is closer than the vowelless *De Magia* or the *Secreta Secretorum*, even though all are in Latin; but the affix-transformation of the *Secreta* is a closer match than the *De Magia*.

Figure 2 gives the Euclidian distance from Voynich scores. In the interests of readability, we show the closest items from the combined Voynich EVA Basic. The number of comparisons totaled 893. The closest comparisons are other subsets or renderings of the Voynich Manuscript.

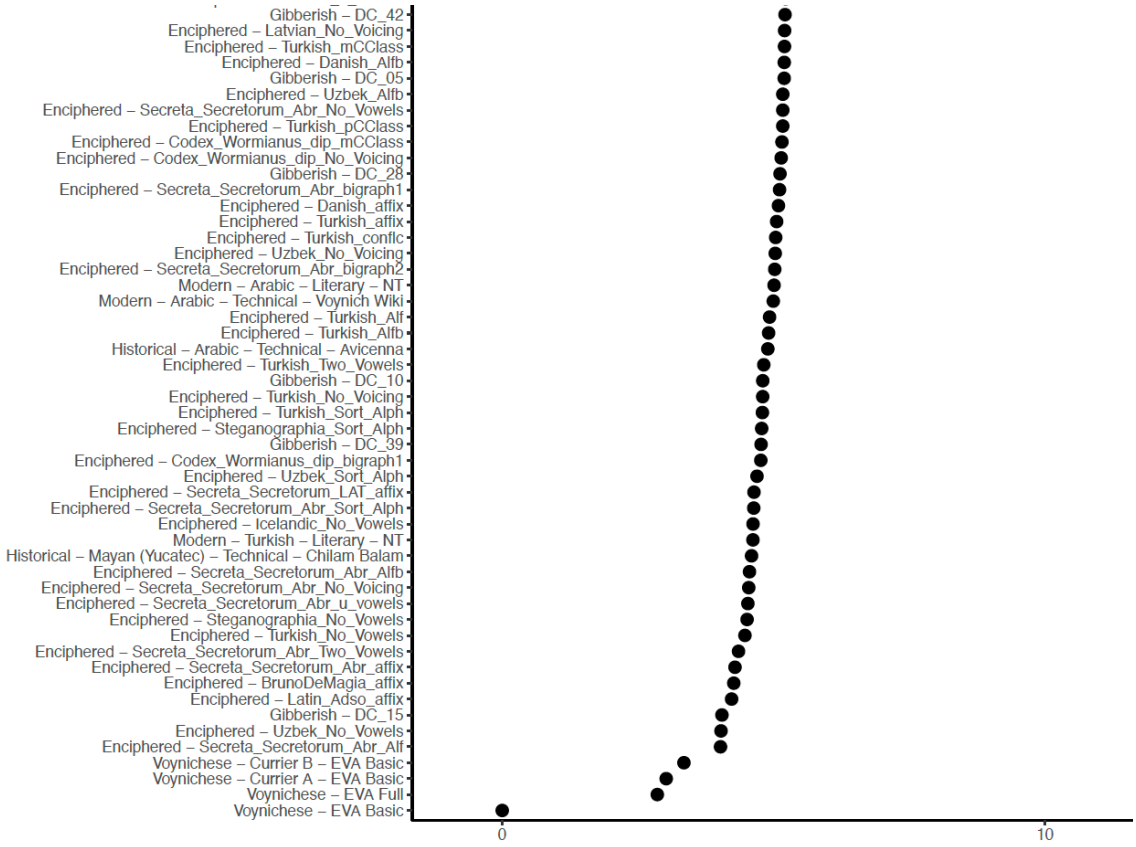

Figure 2: Samples that are closest to the Voynich EVA basic sample (truncated)

There are, as can be seen from Figure 2, many items that are approximately equidistant from the Voynich samples. They include enciphered Latin (e.g. *De Magia* and *Adso*, as well as the version of the *Secreta Secretorum* with scribal abbreviations). The "No_vowel" condition for manuscripts in Uzbek, Turkish, and Latin also results in close matches. But note that similar items also include several gibberish samples from <Voynich submission>, and while these manuscripts are the closest, no manuscript is as close as the other samples of Voynichese, implying that they are not particularly good matches.

In order to examine the contribution of particular text measures to the distance from Voynich samples, we plotted each measure separately. Figure 3 illustrates three of the closest overall matches to the Voynich metrics, along with the automatic composition (gibberish) generated by Timm and Schinner [4] for comparison.

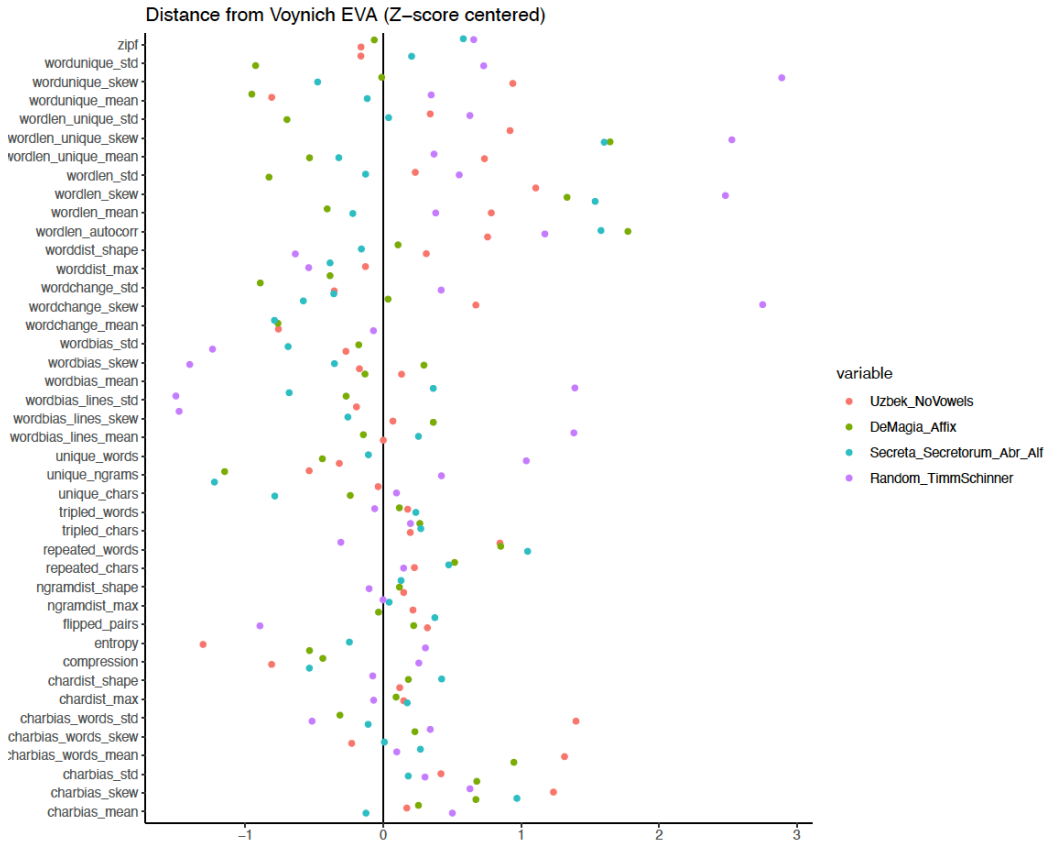

Figure 3: Z-scored metrics, centered on Voynich EVA (line at 0). Points show deviation from value of EVA basic (combined Currier A and B).

As can be seen from Figure 3, although these texts are overall among the closest to the Voynich data, individual metrics vary substantially. The *De Magia* is closest on the wordunique_skew, but among the furthest for the word-length metrics. Devowelled Uzbek is close on word biases on lines, but far from Voynich for entropy measures.

4. Discussion and Conclusions

The main questions examined in this paper are as follows (repeated from Section 1):

1) What effects do textual manipulations (including encipherment) have on statistical measures of text?
2) What types of manipulations (if any) produce outcomes that are similar to Voynichese?
3) Is there an interaction between language and manipulation?
4) Are there manipulations which produce a set of markers which is compatible with Voynich text?

The textual manipulations produce a range of effects on the statistical measures studied here. Many manipulations produce outcomes that are similar to Voynich text on at least some metrics, while showing differences on others. Encipherment approaches that merged phonemic contrasts yielded the closest overall results, suggesting that ciphers and writing systems with this property may be a useful target for future research. There is an interaction with language; removing the vowels from an Uzbek text produces closer samples than doing so with Hawaiian, for example. Likewise, there is also an interaction with the specific text in question (as witnessed by the different degrees of similarity of various Latin texts). These text manipulations did not produce a single set of changes that mimic Voynich text. They do

show, however, that aberrant character-level text metrics are not necessarily an indication that Voynich text is gibberish.

5. References

[1] S. Reddy and K. Knight, "What We Know About The Voynich Manuscript," in *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, Portland, OR, USA, Jun. 2011, pp. 78–86. Accessed: Jul. 26, 2022. [Online]. Available: https://aclanthology.org/W11-1511

[2] M. A. Montemurro and D. H. Zanette, "Keywords and Co-Occurrence Patterns in the Voynich Manuscript: An Information-Theoretic Analysis," *PLOS ONE*, vol. 8, no. 6, p. e66344, Jun. 2013, doi: 10.1371/journal.pone.0066344.

[3] D. R. Amancio, E. G. Altmann, D. Rybski, O. N. Oliveira, and L. F. da Costa, "Probing the Statistical Properties of Unknown Texts: Application to the Voynich Manuscript," *PLoS ONE*, vol. 8, no. 7, pp. e67310–e67310, Jul. 2013, doi: 10/f48qmr.

[4] T. Timm and A. Schinner, "A possible generating algorithm of the Voynich manuscript," *Cryptologia*, vol. 44, no. 1, pp. 1–19, Jan. 2020, doi: 10.1080/01611194.2019.1596999.

[5] G. Rugg and G. Taylor, "Hoaxing statistical features of the Voynich Manuscript," *Cryptologia*, vol. 41, no. 3, pp. 247–268, May 2017, doi: 10.1080/01611194.2016.1206753.

[6] L. Lindemann and C. Bowern, "Character Entropy in Modern and Historical Texts: Comparison Metrics for an Undeciphered Manuscript." arXiv, May 18, 2021. doi: 10.48550/arXiv.2010.14697.

[7] C. L. Bowern and L. Lindemann, "The Linguistics of the Voynich Manuscript," *Annu. Rev. Linguist.*, vol. 7, no. 1, pp. 285–308, Jan. 2021, doi: 10/gmczhg.

[8] R. Sterneck, A. Polish, and C. Bowern, "Topic Modeling in the Voynich Manuscript." arXiv, Jul. 06, 2021. doi: 10.48550/arXiv.2107.02858.

[9] C. Layfield, L. van der Plas, M. Rosner, and J. Abela, "Word Probability Findings in the Voynich Manuscript," in *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, Marseille, France, May 2020, pp. 74–78. Accessed: Aug. 17, 2020. [Online]. Available: https://www.aclweb.org/anthology/2020.lt4hala-1.11

[10] C. F. Hockett, "Animal" languages" and human language," *Human Biology*, vol. 31, no. 1, pp. 32–39, 1959.

[11] S. Romaine, "Language in Society: An Introduction to Sociolinguistics," 1994.

[12] B. Megyesi, C. Tudor, B. Láng, A. Lehofer, N. Kopal, and M. Waldispühl, "What Was Encoded in Historical Cipher Keys in the Early Modern Era?," *International Conference on Historical Cryptology*, pp. 159–167, Jun. 2022, doi: 10.3384/ecp188404.

[13] M. Héder and B. Megyesi, "The DECODE Database of Historical Ciphers and Keys: Version 2," *International Conference on Historical Cryptology*, pp. 111–114, Jun. 2022, doi: 10.3384/ecp188397.

[14] M. D'Imperio, *The Voynich Manuscript: An Elegant Enigma*. Fort George G. Meade, Md.: National Security Agency/Central Security Service, 1978.

[15] R. Zandbergen, "Voynich MS - Text Analysis - Transliteration of the Text," *Voynich.nu*, 2022. http://www.voynich.nu/transcr.html (accessed Aug. 06, 2022).

[16] T. Takahashi, "Voynich Manuscript - transcription," 2000. http://www.voynich.com/pages/index.htm (accessed Oct. 11, 2016).

[17] J. Stolfi, "Reeds/Landini's interlinear file converted to EVA," *Jorge Stolfi homepage*, Dec. 29, 1998. https://www.ic.unicamp.br/~stolfi/voynich/98-12-28-interln16e6/ (accessed Jul. 27, 2022).

[18] L. F. Davis, "How Many Glyphs and How Many Scribes? Digital Paleography and the Voynich Manuscript," *Manuscript Studies: A Journal of the Schoenberg Institute for Manuscript Studies*, vol. 5, no. 1, pp. 164–180, 2020, doi: 10.1353/mns.2020.0011.

[19] R Core Team, "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria, 2017. [Online]. Available: www.R-project.org

[20] C. M. Urzúa, "A simple and efficient test for Zipf's law," *Economics Letters*, vol. 66, no. 3, pp. 257–260, Mar. 2000, doi: 10.1016/S0165-1765(99)00215-3.