Crux of the MATTR: Voynichese Morphological Complexity

Luke Lindemann[1]

[1] *The George Washington University, 2121 I St NW, Washington, DC, USA*

Abstract

This study examines Voynichese at the word level, using a corpus of language samples and word-level statistics in order to identify the most plausible language families for Voynichese, and to exclude statistically improbable language families. The results narrow down the possibilities for a source language and emphasize the likelihood that Voynichese does in fact encode meaningful language. Comparison texts include samples from 160 modern languages in ten major language families, as well as historical manuscripts written in Hebrew, Italian, Old Church Slavonic, and Welsh. The methodology employs two particular word distribution statistics, the Moving Average Type-Token Ratio (MATTR) and the Most Common Words percentage (MCW). A graph of the corpus language samples shows a clustering by language family and correlates with the presence of agglutinating or isolating morphology in a given language family. The methodology ranks plausible language families and excludes statistically improbable language families based on their distribution, and indicates that Voynichese possesses a middle range of morphological complexity that most closely resembles that of medieval Germanic languages. By contrast, Semitic languages like Hebrew and Arabic, and languages from the Slavic and Celtic families are possible but less likely candidates. Families that feature heavy agglutinating morphology like Turkic and Uralic can be effectively excluded.

Keywords

Voynich, Morphology, Natural Language Processing, Corpus Linguistics, Enciphered Hypothesis, Hoax Hypothesis

1. Introduction

Any grounded theory about the underlying meaning or construction of the Voynich manuscript must contend with the unusual distribution of characters in the text. Bennett [1], Stallings [2], and Bowern & Lindemann [3], [4] have noted the extreme predictability of characters in Voynichese, with character combinations only occurring in predictable positions of the word. Bennett compared the conditional character entropy of Voynichese to written Hawaiian, while Reddy & Knight 2011 [5] argue from statistical properties of the text that the writing system most closely resembles an abjad. Bowern and Lindemann argue that the character predictability in Voynichese has no parallels with any known text or writing system, but may be the result of encipherment of a natural language. Timm & Schinner [6] and Rugg & Taylor [7] argue that this predictability indicates that the Voynich text has no underlying meaning, and suggest methods for its construction.

However, an intriguing aspect of the Voynich text is that while the distribution of characters in the text is extremely unusual, in many ways the distribution of words is not. Landini [8], among others, notes that the Voynich word distribution follows Zipf's law, a power law that relates word rank with

International Conference on the Voynich Manuscript 2022, November 30--December 1, 2022, University of Malta.
EMAIL: luke.lindemann@gmail.com (A. 1)
ORCID: 0000-0003-2675-5244

frequency. Additionally, the most frequent words in the Voynich A and Voynich B Currier languages [9] occur at similar rates to the most common words in natural language texts.

Examining Voynichese at the word level allows us to largely set aside unsettled questions of script encoding and encipherment, and to develop a morphological profile of Voynichese among plausible candidates for language families. Note that this relies upon the assumption that the Voynich script uses spaces in a conventional way, i.e., to separate distinct words.

This paper proposes two measures, Moving Average Type-Token Ratio (MATTR) and the Most Common Word measure (MCW), which correlate with the morphological complexity of a text without reference to internal properties of the words themselves (e.g. spelling, word length, or the presence of common suffixes). Morphological complexity, as used here, roughly equates with the number of morphemes per word. Individual languages and language families can be placed along a spectrum from more to less morphologically complex based upon the abundance of morphological structures of inflection and affixation. On the more complex side are languages with agglutinating morphology like Turkish, in which a word[2] typically contains multiple morphemes:

(1)	Gör-üş-tür-ül-e-me-ye	de	bil-iyor	mu-ydu-nuz?
	see-REC-CAUS-PASS-PSB-NEG-PSB	also	PSB-IMPF	INT-P.COP-2.PL

‘Did it also sometimes happen that you were not allowed to see each other?’

(Göksel and Kerslake [10])

On the less morphologically complex side are languages with isolating morphology like Hawaiian, in which each word typically contains a single morpheme. There are fewer agreement features, and information like tense and aspect may be conveyed with separate words:

(2)	Ke	kali	nei	au
	PRES	wait	now	I

‘I’m waiting now.’

(Elbert and Pukui [11])

The goal of this analysis is to place Voynichese along this spectrum of morphological complexity by comparing it against a corpus of language samples in several different language families with different types of morphology.

2. Text Corpora

Voynichese is compared against a multilingual corpus derived from different language versions of Wikipedia articles. Additional texts used for this analysis include transcriptions of the Voynich Manuscript and a collection of historical manuscripts. In order to ensure uniformity of length, each text has been divided into equal partitions of 10,000 words. A uniform sample size controls against the possibility of length effects. The window of 10,000 words was chosen because it is the approximately maximum size for comparison with Voynichese. It is roughly the number of words in Voynich A (and approximately half the number of words in Voynich B). Cleaning, text manipulations, and analyses were done with R statistical computing software [12].

[2] Here *‘word’* is defined typographically as any sequence of characters separated by spaces. This definition is in opposition to, for example, definitions of the prosodic or phonological word. This is relevant because of differences in orthographic convention. Some texts had to be excluded from our sample because they were written in scripts which do not separate words within a sentence. In others language families, some of the languages are conventionally written with spaces separating affixes as if they were separate words, while in other languages of the same family affixes are written as part of the word. This was a reason for excluding, for example, the Bantu family.

The samples of Voynichese consist of the running text in paragraphs (excluding labels on pictures and diagrams). The Voynich A sample consists of the first 10,000 words of the Voynich manuscript written in Currier Language A. The Voynich B samples consists of the Voynich B text partitioned into two samples of 10,000 words each. The Full Voynich sample consists of three 10,000-word samples of running text in the Voynich manuscript undivided by Currier Language. The samples are rendered in the Extensible Voynich Alphabet (EVA) transcription, and were derived from the Landini-Stolfi Interlinear Gloss files [13]. The Takahashi transcription, which is the most complete, was the sole transcription consulted.[3]

The analysis compares the Voynich text against a multilingual corpus derived from different language versions of Wikipedia articles. The original corpus was obtained from supplementary materials provided for [3]. It contains samples of 311 languages from 38 families. For the purpose of this analysis, I further restricted the corpus to languages that (a) are written in scripts which conventionally separate words using spaces or punctuation (excluding, e.g., Chinese, Burmese, and Thai), (b) contain more than ten thousand total words, and (c) are members of families with at least one hundred representative samples (where each sample is a partitioned text of 10,000 words). See Table (4) in the appendix for a full list of languages by language family.

The 160 language texts were partitioned into equal samples and divided into ten separate language families: Celtic, Germanic, Indic, Iranian, Phillipine, Romance, Semitic, Slavic, Turkic, and Uralic. For each family, 100 samples were chosen at random so that each family is represented by an equal number of samples.

The comparison historical texts include the the Italian (Romance) *La Rettorica* [14], the Hebrew (Semitic) *Bereshit* of the *Tanach* [15], the Welsh (Celtic) *Mabinogian* [16], and the Old Church Slavonic (Slavic) *Codex Suprasliensis* [17]. These were cleaned and partitioned into equal samples of 10,000 words each.

2.1. MATTR

The first statistic we will examine is the moving average type-token ratio (MATTR). The type-token ratio of a text is the number of unique words divided by the total number of words. TTR is useful but highly dependent upon the length of the given text. To compensate for this, MATTR is the average type-token ratio across equally partitioned sections of text. The application of MATTR in the identification of the Voynich manuscript was suggested by Gheuens 2019 [18] and further explored in Bowern and Lindemann 2020 [4].

MATTR correlates with morphological complexity because the morphological expression of categories like gender and case allow for the proliferation of textually unique words. For example, the Latin word *'librorum'* is the plural genitive of the masculine noun *'liber'* ('*book'*). A translation into English, a more isolating language, might render this single Latin word with the phrase *'of the books'*. We would expect *'librorum'* to occur very infrequently in a Latin text, while each of the words *'of,'* *'the,'* and *'books'* will occur with comparatively greater frequency in an English text.

The window length for MATTR must be calibrated to the comparison corpus. We want to choose a window length such that the language samples cluster closely within their own language family, while keeping family clusters as separate as possible. In Figure (1), we examine every window length from

[3] A reviewer suggested the deletion of ambiguous/unresolved words that are rendered differently by different transcribers. This would require deleting 45% of the Full Voynich running text (41% of the Voynich A running text and 46% of the Voynich B running text), which speaks to the depth of unresolved issues in the Voynich script. I elected to use Takahashi's transcription unaltered, because spelling errors in a consistent transcription should have a relatively small effect on word-level statistics. In any case, the result of deleting ambiguous words does not substantially impact our conclusions: an 8.4% difference for the MATTR statistic and -6.5% for the MCW statistic, in a direction suggestive of lower morphological complexity. The Voynich A and Voynich B samples change roughly the same amount, and remain similarly distinct from each other.

1-20,000 words. For each window, we calculate the overall statistical variance of MATTR values in the entire corpus. We then calculate the mean statistical variance within each language family, and subtract this from the overall statistical variance. The peak is at 1000 words, so this is our optimal word window for calculating MATTR.

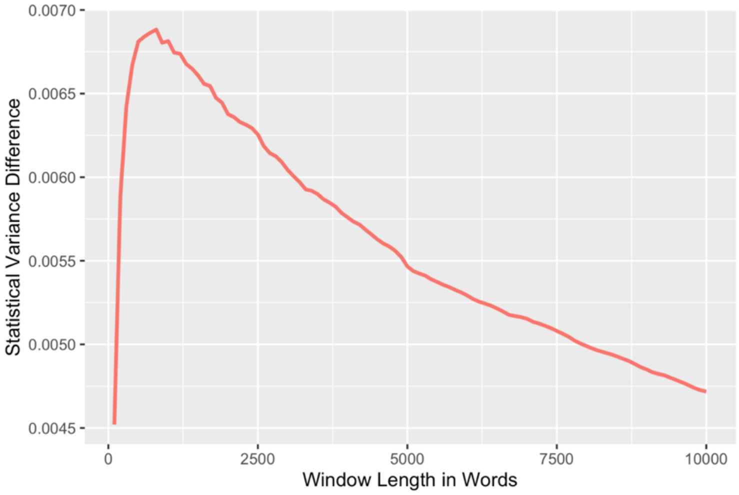

Figure 1: Difference in Language Family Variance for MATTR

2.2. MCW

The second statistic is the most common words measure (MCW). This is the proportion of overall tokens in a text that consist of the most frequent word types. The most common words in a language are typically determiners, conjunctions, copulas, or prepositions. The MCW takes the aggregate frequencies of the highest ranked words (e.g., the five/ten/twenty/hundred most common words).

MCW correlates with morphological complexity because the addition of morphological structure decreases the frequency of the most common words. Consider the most common word in English: the definite determiner '*the*'. This word accounts for 7% of the words in our English sample. The most common word in German, a related Germanic language that is more morphologically complex, is also a definite determiner (the masculine singular nominative '*der*'). This word accounts for only 4% of the overall text. It is one of several other forms of '*the*' based on case, gender, and number. The role of the single English word '*the*' is spread out over multiple German forms ('*der*', '*die*', '*das*', '*den*', '*dem*', and '*des*').

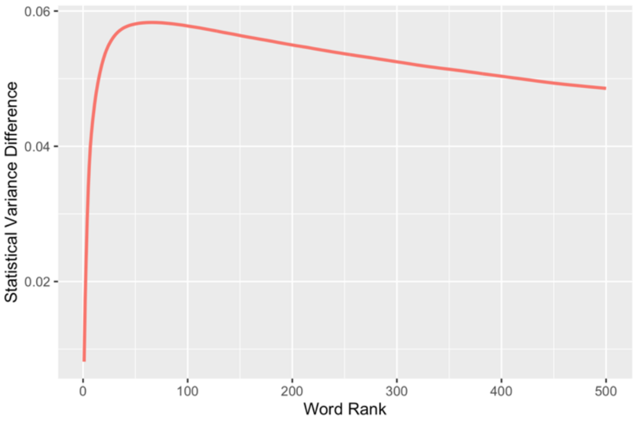

Figure 2: Difference in Language Family Variance for MCW Word Ranks

As with MATTR, we choose an optimal selection of ranked words by examining the MCW values in our corpus for each word rank. In Figure (2), we calculate the overall statistical variance of MCW values and subtract from this the average statistical variance of the language families. The optimal rank is at 70 words, so our MCW measure will be the proportional frequency of the most common seventy words in the sample.

3. Language Family Comparison

Figure (3) is a graph of the centroid values of each language family for MATTR (with a partition window of 1000 words) against MCW (with a rank of 70 words). Language families that exhibit the highest morphological complexity appear in the upper left portion of the graph, with a low MCW value and a high MATTR value. Families with more isolating languages appear in the bottom right portion of the graph, with a high MCW value and a low MATTR value. The highest tier of morphological complexity consists of the Turkic and Uralic languages, followed by the Semitic and Slavic languages, the Indic and Iranian, the Germanic and Romance, and finally the Phillipine languages.

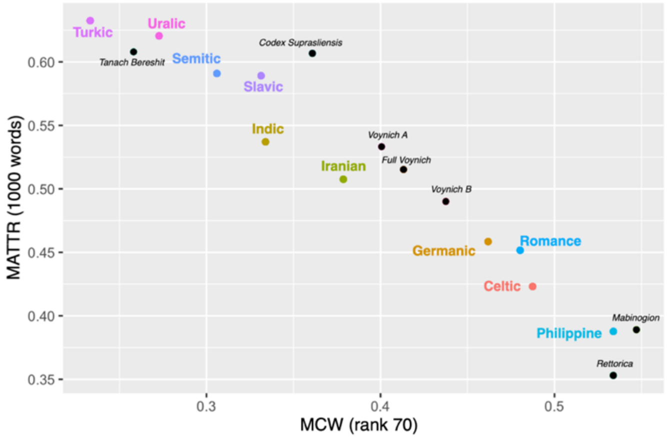

Figure 3: Average MATTR and MCW Values by Language Family and Manuscript

The average values for the samples from each of the Voynich and other historical texts are also given in Figure (3). The values for the Voynich samples are in the middle, suggesting a level of morphological complexity that is higher than the average for Germanic and Romance languages and lower than Semitic and Slavic. Voynich A appears to be more morphologically complex than Voynich B, and the closest proximal centroid is the Iranian family.

Note, however, that some of the historical manuscripts are removed from the centroids of their modern languages, particularly for the *Mabinogian* (Celtic) and *Rettorica* (Romance). This is partly attributable to variation within each family due to the properties of specific languages and variation within each text. This is illustrated for the Romance family in Figure (5), which shows the individual values for every Romance language sample in the Wikipedia corpus along with those of the Italian historical text and the Voynich manuscript.

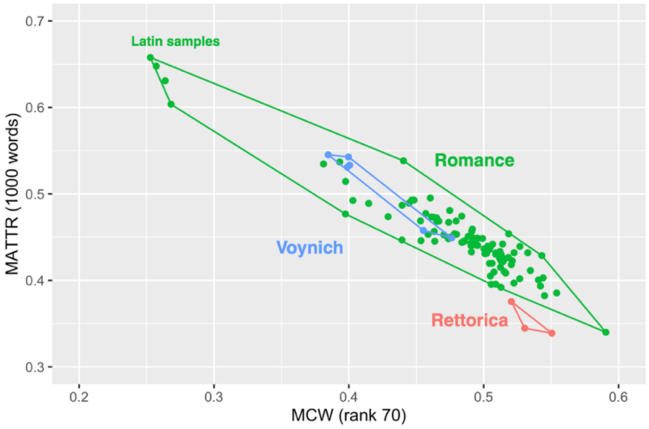

Figure 4: Individual MATTR and MCW Values in the Romance Family

While the historical Italian samples (and the Voynich samples) are somewhat removed from the Romance centroid, they are not unexpected given the overall Romance distribution. The Romance family also includes samples of Latin. Latin is not technically a Romance language, but it was included because it is the parent language from which the Romance languages evolved. The Latin samples are clear outliers among the Romance languages, which reflects the much higher degree of morphological complexity in Latin compared with the modern languages.

Figure (3) gives an impressionistic picture of Voynichese along the spectrum of morphological complexity. We can use a more quantitative method to exclude unlikely candidates for language families. Figure (5) is a density plot which shows the MCW values for the Celtic family, with a fitted normal distribution in red. The solid red lines indicate cut-offs at three standard deviations from the mean. The dotted red line is the average value for the Welsh *Mabinogian* sample, and the dotted blue line represents the average value for the Old Church Slavic *Codex Suprasliensis*. Because the latter value is outside the cut-off for the Celtic normal distribution, we can conclude that Celtic is unlikely to be a candidate language family for the *Codex Suprasliensis,* but is a candidate language family for the *Mabinogian* (both correct results).

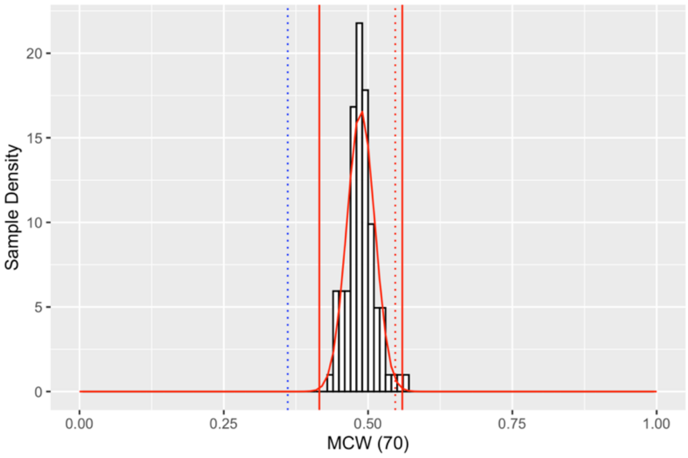

Figure 6: Sample Density and Normal Distribution for the Celtic Family (solid red); average values from Welsh *Mabinogian* (dotted red) and Old Church Slavonic *Codex Suprasliensis* (dotted blue)

Accordingly, the method is to calculate the average values for each historical text and rank according to the two measures and rank the families by simple Euclidean distance. Then, for each measure and language family, we calculate a probability distribution and determine which texts are likely to be in the distribution ($0.003 < p < 0.997$). We consider a language family to be an unlikely candidate if it is statistically excluded by both measures.

4. Results

Table (2) summarizes the results for the historical manuscripts. Each family is ranked by Euclidean distance from the average (MCW, MATTR) pair of the manuscript to the average for the family. The shaded cells represent language families that are statistically excluded: light gray indicates that they are excluded by one statistic, either MATTR or MCW, and dark gray indicates that they are excluded by both. Note that the correct language family is among the top three ranked language families in each case. About a third of the language families are statistically excluded according to both measures, and in no case is the correct language family excluded. Table (3) summarizes the results for the full Voynich text and to Voynich A and Voynich B separately.

Table 2
Historical Manuscripts

Mabinogian (Celtic)	*Codex Suprasliensis* (Slavic)	*Tanach Bereshit* (Semitic)	*Rettorica* (Romance)
Phillipine	Slavic	Uralic	Phillipine
Celtic	Semitic	Turkic	Celtic
Romance	Indic	Semitic	Romance
Germanic	Uralic	Slavic	Germanic
Iranian	Iranian	Indic	Iranian
Indic	Turkic	Iranian	Indic
Slavic	Germanic	Germanic	Slavic
Semitic	Romance	Romance	Semitic
Uralic	Celtic	Celtic	Uralic
Turkic	Phillipine	Phillipine	Turkic

Table 3
Voynich Text

Full Voynich	Voynich A	Voynich B
Iranian	Iranian	Iranian
Germanic	Indic	Germanic
Indic	Slavic	Indic
Romance	Germanic	Romance
Slavic	Semitic	Celtic
Celtic	Romance	Slavic
Semitic	Celtic	Semitic
Phillipine	Uralic	Uralic
Uralic	Turkic	Phillipine
Turkic	Phillipine	Turkic

5. Conclusions

This paper has demonstrated a method for developing a morphological profile of an unknown language text along a spectrum of morphological complexity. Applying this method to the Voynich texts, we find that Voynichese possesses a middle range of morphological complexity that is somewhat higher than the average for Romance and Germanic but lower than Semitic and Slavic. On both the high and low side of the spectrum, we can effectively eliminate the Phillipine and Turkic languages, and most likely the Uralic and Celtic languages as well. Romance remains a possibility, although

Voynich appears somewhat below Latin but higher than what we find from the majority of the modern Romance languages. The profile of Voynich A suggests that it is more morphologically complex than Voynich B, which may indicate that it encodes a separate language or dialect.

This provides compelling evidence that Voynich does, in fact, encode meaningful text. While character-level measures mark Voynichese as a complete outlier among historical European manuscripts, the word-level measures for Voynich place it comfortably between the Old Slavonic and Hebrew historical texts on one side and the Italian and Welsh texts on the other. A meaningless text in which the characters of each word are chosen at random would have a flat distribution, with each word having an approximately equal frequency. This would result in MATTR values approaching zero and MCW values approaching one (far off to the bottom right of the chart). On the other hand, a method of encipherment which produces the anomalous character-level statistics we see may have a minimal effect on the overall distribution of words.

These results do not exclude the hypothesis that Voynich is meaningless, because the writers could have employed a method to artificially imitate the word distribution of medieval European texts. However, any proposed method for the creation of Voynich-as-gibberish should be able to produce reasonable word-level statistics like those that we find here. In any case, improving upon this methodology may help to geographically situate Voynichese, either by identifying an underlying natural language or by suggesting the kind of language that a medieval hoaxer could have been imitating.

The focus of future research will be on increasing the quantity of samples from language families to develop a more accurate picture of the range of results, and on researching other word-level measures that may correlate more exactly with morphological complexity. Voynichese is similarly middle range among other word-level statistics that we have examined, including word-level entropy measures, the proportion of tokens which consist of hapax legomena, and average word length. These statistics also correlate with morphological complexity, but produce a wider range of variation in individual languages that may suggest that they are more sensitive to the preferences and styles of individual authors.

6. Appendix

Table 4

Wikipedia Samples by Language Family

Family	Language Samples
Celtic	Breton, Cornish, Irish, Manx, Scottish Gaelic, Welsh
Germanic	Afrikaans, Alemannic, Anglo Saxon, Bavarian, Danish, Dutch, Dutch Low Saxon, English, Faroese, German, Icelandic, Limburgish, Low Saxon, Luxembourgisch, North Frisian, Norwegian (Bokmål), Norwegian (Nynorsk), Palatinate German, Pennsylvania German, Ripuarian, Saterland Frisian, Scots, Simple English, Swedish, West Flemish, West Frisian, Yiddish, Zeelandic
Indic	Assamese, Awadhi, Bengali, Bihari, Bishnupriya Manipuri, Dvehi, Doteli, Fiji Hindi, Goan Konkani, Gujarati, Hindi, Maithili, Marathi, Nepali, Oriya, Punjabi, Romani, Saaraiki, Sanskrit, Sindhi, Sinhalese, Urdu, Western Punjabi
Iranian	Gilaki, Kurdish, Mazandarani, Ossetian, Pashto, Persian, Sorani, Taji, Zazaki
Phillipine	Cebuano, Central Bicolano, Gorontalo, Ilokano, Kapampangan, Pangasinan, Tagalog, Waray-Waray
Romance	Aragonese, Aromanian, Asturian, Catalan, Corsican, Emilian-Romagnol, Extremaduran, Franco-Provençal, French, Friulian, Galician, Italian, Ladin, Ladino, Latin, Ligurian, Lombard, Mirandese, Neapolitan, Norman, Occitan, Picard, Piedmontese, Portuguese, Romanian, Romansh, Sardinian, Sicilian, Spanish, Tarantino, Venetian, Walloon
Semitic	Amharic, Arabic, Egyptian Arabic, Hebrew, Maltese, Moroccan Arabic, Tigrinya
Slavic	Belarusian, Belarusian (Taraškievica), Bosnian, Bulgarian, Croatian, Czech, Kashubian, Lower Sorbian, Macedonian, Old Church Slavonic, Polish, Russian, Rusyn, Serbian, Serbo-Croatian, Silesian, Slovak, Slovenian, Ukrainian, Upper Sorbian

Turkic	Azerbaijani, Bashkir, Chuvash, Crimean Tatar, Gagauz, Karachay-Balkar, Karakalpak, Kazakh, Kirghiz, Sakha, South Azerbaijani, Tatar (Cyrillic), Tatar (Latin), Turkish, Turkmen, Tuvan, Uyghur, Uzbek
Uralic	Erzya, Estonian, Finnish, Hill Mari, Hungarian, Inari Sami, Komi, Komi-Permyak, Livvi-Karelian, Meadow Mari, Northern Sami, Udmurt, Vepsian, Võro

7. References

[1] W. R. Bennett. and engineering problem-solving with the computer. Prentice Hall Series in Automatic Computation. Prentice Hall, Hoboken, New Jersey, 1976. ISBN: 0137958072.

[2] D. Stallings, Understanding the second-order entropies of Voynich text, 1998. URL: http://ixoloxi.com/voynich/mbpaper.htm.

[3] L. Lindemann and C. Bowern. "Character entropy in modern and historical texts: comparison metrics for an undeciphered manuscript." Yale Voynich Working Group, Yale University, New Haven, Connecticut. 2022. arXiv:2010.14697v2.

[4] C. Bowern and L. Lindemann. "The Linguistics of the Voynich Manuscript." Annual Review of Linguistics 7 (2021): 285-308.

[5] S. Reddy and K. Knight. What we know about the Voynich manuscript. Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities 2011: 78–86. URL: http://dl.acm.org/citation.cfm?id=2107647.

[6] T. Timm and A. Schinner, "A possible generating algorithm of the Voynich manuscript," Cryptologia, 44(1), 1–19, Jan. 2020, doi: 10.1080/01611194.2019.1596999.

[7] G. Rugg and G. Taylor, "Hoaxing statistical features of the Voynich Manuscript," Cryptologia, 41(3). 247–268, May 2017, doi: 10.1080/01611194.2016.1206753.

[8] G. Landini, "Evidence of linguistic structure in the Voynich manuscript using spectral analysis," Cryptologia 25(4). 275–295. 2001. doi:10/dhtvfp.

[9] P. H. Currier. Papers on the Voynich manuscript. 1976. URL: http://www.voynich.nu/extra/img/curr/main.pdf.

[10] A. Göksel and C. Kerslake. Turkish: A comprehensive grammar, p. 48. 2004. Routledge. ISBN: 9780415114943

[11] S.H. Elbert and M. K Pukui. Hawaiian Grammar, p. 60. 2004. University of Hawaii Press.

[12] R Core Team, R: A language and environment for statistical computing, 2019. URL: https://www.R-project.org/.

[13] R. Zandbergen, Text Analysis – Transliteration of the Text, Voynich MS, 2022. URL: http://www.voynich.nu/transcr.html.

[14] B. Latini. La Rettorica, 1261. F. Maggini (ed.), Florence, 1915. Electronically prepared by G. Ferraresi, E. Rinke, and M. Goldbach, Hamburg 2005. TITUS (Thesaurus Indogermanischer Text- und Sprachmaterialien), J. Gippert, Johann Wolfgang Goethe University in Frankfurt am Main. URL: http://titus.uni-frankfurt.de/texte/etcs/ital/aital/latrett/latre.htm

[15] Tanach Bereshit. Westminster Leningrad Codex. URL: https://www.sacred-texts.com/bib/tan/index.htm

[16] I. Williams (ed.) The Mabinogian: Pedeir Keinc y Mabinogi allan o Lyfr Gwyn Rhydderch. 1930. Electronically prepared by E. Parina, 2005. TITUS (Thesaurus Indogermanischer Text- und Sprachmaterialien), J. Gippert, Johann Wolfgang Goethe University in Frankfurt am Main. URL: http://titus.uni-frankfurt.de/texte/etcs/celt/mcymr/pkm/pkm.htm

[17] S. Sever'janov (ed.) Codex Suprasliensis, 1904. Electronically prepared by J. Lindestedt, Helsinki 1994-2005. TITUS (Thesaurus Indogermanischer Text- und Sprachmaterialien), J. Gippert, Johann Wolfgang Goethe University in Frankfurt am Main. URL: http://titus.uni-frankfurt.de/texte/etcs/slav/aksl/suprasl/supra.htm

[18] K. Gheuens. Examining lexical diversity in the Voynich manuscript, 2019. URL: https://herculeaf.wordpress.com/2019/05/04/type-token-ratio/.