A Quantitative Study of the Voynich Manuscript through the Kolmogorov-Smirnov Test

Anoushka Shrivastava, Robert Downing

Abstract

The Voynich Manuscript [MS408, VM] is a collection of drawings and text written in an unknown source language, using an unknown character set from the 15th century, and written by an unknown author, so virtually nothing is known about the significance of what it contains. From booksellers to cryptographers, no one has been able to decipher the contents of the VM, but they have come up with various theories, from alien influences to the manuscript being a hoax. Though there is much research on the qualitative aspects of the VM, there is a lack of quantitative analyses of the text. The goal of this paper is to provide a quantitative comparison between the possible source language used in the VM and modern-day languages through an application of the Kolmogorov-Smirnov [K-S] test, a robust statistical test that compares the distribution of two datasets. Results of this test indicate a strong similarity between the VM, Latin, and Italian, supporting the widely-regarded belief that the language of the VM is a Vulgar Latin derivative.

Introduction

Background on the Manuscript

The creator of the VM is unknown[1]. The first known owner of the VM is Emperor Rudolph II, who purchased the text[1]. After him, at least 3 individuals acquired the text until it came into the possession of Wilfred M. Voynich, a bookseller who acquired the manuscript from Jesuits and failed at deciphering its contents[1]. After he and his wife passed away, his secretary inherited the text, which she sold to a bookseller, who donated it to the Beinecke Library at Yale where it has been since 1969[1]. Throughout its course, many have tried to decipher the text, including Jesuit scholars, owners of the text, and even the US's top cryptographers, but all have failed in their attempts[1].

Results from carbon dating indicate that the parchment is from the 1400s, and another team of scientists determined that up to 15 calf skins were used to create the parchment[1]. Analyses on the ink consistently indicate that it is the type medieval scribes used[1]. After further analysis, no modern-day signs (pigments, etc.) were found, and details down to the style of the manuscript's binding are consistent with medieval times[1].

The manuscript itself is 10 inches by 7 inches, and contains pages, some with foldouts, with writings and numerous drawings[1]. The manuscript has many distinct sections: Botanical, with illustrations of various plants, some of which have not been identified; Anatomical, with human bodies; and Astrological, with pictures of stars and planets, to give some examples[1].

Fig. 1. Sample pages from the Botanical section of the Voynich Manuscript[2].

Fig. 2. Sample pages from the Astrological section Voynich Manuscript[2].

Various Transcription Alphabets

Transcribing the VM refers to creating a separate alphabet consisting of symbols a computer may recognize and using this alphabet to replace each handwritten character; in other words, transcription converts the text's handwritten symbols to a machine-readable alphabet[3]. This process should, in theory, preserve the entire manuscript and its contents, because we are simply replacing a symbol with another, but there are possible human errors such as reading the handwriting incorrectly, copying the wrong character, etc. However, due to the magnitude of the character set of the VM, these errors are negligible and should not interfere with major analyses.

The primary difficult task when transcribing the VM is determining which handwritten characters are ligatures (combined characters such as æ) and which are simply complex characters, and each of the below transcribers had a different opinion of this[3].

One of the first known attempts to transcribe the VM came from a group of individuals led by William Friedman, The First Study Group, but they only created an alphabet for certain parts of the manuscript they wanted to focus on, so their alphabet is incomplete and not applicable for this study[3]. Prescott Currier and Mary D'Imperio also created the "C-D" or Currier alphabet, which used the A through Z alphabet and the digits 0-9[3]. However, because they represented some characters that looked like multiple characters as single characters, there is inaccuracy in their alphabet[3]. Jacques Guy with his alphabet Frogguy used characters to represent each stroke as a symbol, which though precise, meant that single characters from the VM were made of multiple characters in the Frogguy alphabet[3]. The Eva Alphabet created by Gabriel Landini and Rene Zandbergen was designed to capture the entire manuscript and create pronounceable words[3]. This is so humans could draw connections and patterns, as this is more easily done when words are pronounceable and memorable[3]. Ultimately, because the alphabet Voynich101 [v101] created by Glen Caston was one of the most recent transcriptions, uniquely represented characters that were slight variants as separate characters, and spanned the entire text, it was used for the purpose of this study[3].

Evolutionary Linguistics

Latin comes from the Indo-European family, and can be broken down into Classical Latin (used for texts and also spoken by upper class members) and Vulgar Latin (spoken Latin), from the latter of which Continental Romance languages are derived[4]. From the Proto-Italian branch comes Italian, and from Western Romance languages come the Ibero-Romance languages of Portuguese and Spanish and the Gallo-Romance language of French[4].

Another branch of the Indo-European family is the Germanic Languages[5]. English and German were derived from this specific branch[5].

Irish was derived from the Celtic branch under the Indo-European family[6].

Greek was derived from the Indo-European languages[7].

Hebrew was derived from the Semitic language family[8].

Belarusian is another Indo-European language, but it is specifically from a branch of the East Slavic languages[9]. Czech is from the West Slavic languages[10].

On the other hand, Indonesian is derived from branches under the Austronesian family, not Indo-European, languages[11]. Languages from this family are used more in Asian countries[11].

Research Methods

Previous researchers have tried various strategies, from making qualitative observations to applying algorithms, but none have been able to provide a solution to decrypting the text that is applicable for its entirety. Therefore, the purpose of this study is not to decipher the text, but to provide a starting point for future decryptions through a quantitative analysis of it.

Creating a Data Set For the VM and Acquiring Letter Frequency Tables

Before performing any analysis on the manuscript, it was essential to convert the manuscript into a data set. Specifically, a set of character frequencies was needed to compare with the letter frequencies of languages commonly known today.

To start my work, I acquired a transcribed form, and its corresponding transcription alphabet of the VM, namely Voynich 101 created by Glen Caston[12]. I input the entire text of the manuscript into the tool WordCreator™ (provided by Stefan Trost Media[13]) that is equipped with a character counter, and set exclusions for irrelevant characters (question marks representing unknown letters, periods representing spaces, etc.)[3]. The output was the relative frequency of each character in the manuscript, and was used as the data set representing the VM's letter frequencies.

From Stefan Trost Media's website, I also acquired the letter frequency tables for Latin[14], Italian[15], Indonesian[16], Portuguese[17], German[18], Spanish[19], French[20], English[21], Irish[22], Hebrew[23], Greek[24], Belorussian[25], and Czech[26]. These languages are all modern-day versions, which future work may improve upon, as this is a starting point for comparison (see "Limitations to this Study").

Interestingly, in the process of testing, I realized that the Voynich Manuscript and other languages do not match in terms of alphabet sizes (there are 73 unique characters in the Voynich manuscript, which is much larger than other alphabets). However, after the 18th most frequent character, the rest each appeared less than 0.63% of the time, and after the 26th most frequent character, each character appeared less than 0.27% of the time. Additionally, each other alphabet was composed of at least 18 characters and about 26. Therefore, when I applied the test to an alphabet and the VM, I truncated the Voynich Alphabet to match the length of the tested alphabet, as required by the K-S test. I first sorted the characters in the VM into descending order of frequency, truncated the alphabet to match the tested alphabet's length, and disregarded the rest of the characters, which would not significantly affect the accuracy of my results due to their low frequencies.

The K-S test itself requires that both alphabets be sorted in ascending frequency, so after I truncated the alphabet of the VM, I sorted both sets by this standard. Of course, I could not truncate the VM after it was sorted in terms of ascending frequency, because I then would be disregarding the most frequent characters and only be using obscure characters with frequencies of less than 1%.

The K-S Test For Two Samples

There are many statistical tests that can be used to compare two data sets, but with the limited information from the manuscript, few can be applied. I chose to apply the K-S test because this test only required the data sets I'd obtained, allows different sample sizes, was a robust means of comparison, and would calculate a dissimilarity value between the VM and each language. The advantage of this last fact was that I could compare each dissimilarity value and gage which languages are most similar to the VM.

The K-S test for two samples is a statistical test that compares the similarity of the distributions of two data sets, which shall be referred to as the observed and theoretical sets[27]. By comparing an alphabet's distribution to the distribution of characters in the VM, I was basically comparing how similar the language and the VM were. Similar distributions between the two means that the language behind the VM is likely a form of, or influenced by, that language.

For the K-S test to apply, the data does not need to have a certain distribution, making the test ideal for the VM[28]. In this case, the theoretical set is the dataset from the VM and the observed set is the distribution from whichever language we currently decide to test. The null hypothesis, or H_0, assumes the datasets have the same distribution, and the opposing hypothesis, H_a, assumes the datasets have different distributions[29]. Through the K-S test, one computes a K-S statistic that represents the dissimilarity of the distribution of two datasets. Then, one compares this statistic to a critical value to determine if there is convincing evidence to reject the null hypothesis.

Although statistical software can be used to carry out this test, I have conducted this test manually through Microsoft Excel (Microsoft Office Professional Plus 2019, Version 16.0.12624.20424). My reasons for using this software are detailed further below (see "Applying the K-S Test"). The procedure I used is described by Charles Zaiontz in "Two-Sample Kolmogorov-Smirnov Test" [29], but I am describing it below for context.

In a new sheet for each round of the test, I listed out each dataset in a column in ascending order of frequency. For reference, I am referring to these as columns B for the theoretical dataset and column E for the observed dataset.

Next, I calculated the cumulative distribution for each set. Going down the frequency columns, at each symbol, I simply took the sum of the values of the frequency of the symbol and all the symbols' frequencies before it, ending at the first symbol's frequency. Then, I moved to the next symbol, until the last character was reached. This ultimately produced a column representing the cumulative distribution for that particular dataset. To store the cumulative distributions in my work, I used column G for the theoretical set and column H for the observed set.

Then, I calculated the absolute difference at each point along both cumulative distribution functions by taking the absolute difference of the entry in column G and the entry in column H as for every point along the datasets. I stored these values in column 5.

Next, I calculated D_n, the K-S statistic, the maximum absolute difference at any point in Column 5. I also refer to this as the dissimilarity value of the two datasets.

Then, I compared the dissimilarity value to a critical value. At a 95% confidence level, this value is $D_{crit,\ 0.05} = 1.36(1/n_{observed} + 1/n_{theoretical})^{0.5}$, where “n” represents each sample size[26]. One and thirty-six hundredths [1.36] is the standard constant when using an alpha value [α] of 0.05, as in this case[26].

If $D_n > D_{crit}$, then one rejects the null hypothesis; in other words, there is enough evidence to assume the two datasets have different distributions. Otherwise, we have insufficient evidence to reject the null hypothesis, and there isn’t enough evidence to assume the datasets have different distributions. This does not mean that the null hypothesis is necessarily true, it just means we cannot disprove it for the time being.

One can also calculate the p-value (which I did in the cell below the one with the critical value) and compare it to alpha, the significance level of Type 1 error. To calculate the p-value, I used an online calculator by University of Baltimore[30] with parameters of the cells of D_n, $n_{observed}$, and $n_{theoretical}$, where n represents the sample size of the languages from which frequencies were calculated. In the case of the VM, the Word Creator tool counted 157,918 characters. I inputted the entire Voynich Manuscript as the sample to avoid bias from any sections of the text, so the sample count actually represented the entire population. In the case of other languages, Stefan Trost provided the number of characters he had chosen to calculate frequencies from, which I applied as the sample size.

If the p-value is greater than alpha, then the null hypothesis is not rejected. Else, there is enough evidence to reject the null hypothesis.

Applying the K-S Test

Because I would apply the K-S test between 13 languages and the VM and needed to apply formulas hundreds of times, spreadsheet software was the ideal tool to use. One round of the test consisted of applying the K-S test once between one language and the VM, so for my 13 rounds total, I created 13 sheets in Microsoft Excel, in addition to a couple more sheets to summarize my results.

I started by comparing the VM to Italian. I copied the letter frequencies for the VM in ascending order, and repeated the procedure in a new column for Italian. Using the rest of the procedure detailed in “The K-S Test For Two Samples”, I found the maximum value of the difference of cumulative relative frequencies, or the dissimilarity value, between Italian and the VM.

As part of the K-S test, I compared the dissimilarity value to a critical value. If the dissimilarity value exceeds the critical value, the two data sets’ distributions are too dissimilar, and I excluded the language from future comparisons, as it was likely not relevant to the VM at all. Else, I noted the dissimilarity value between the VM and that language. Finally, I computed the p-value of each round of the test and input the values in a table as shown in Figure 5. In either case, rejection of the null hypothesis or not, this marked the completion of the K-S test. I repeated this test for 12 additional languages, input the dissimilarity values in a table, and sorted the dissimilarity values in ascending order.

A	B	C	D	E	F	G	H	I
Voynich Letter	Frequency		Italian Letter	Frequency		Sum	Sum	Difference
(	0.0025359		É	0.0006		0.0025359	0.0006	0.0019359
5	0.0025801		Ì	0.0009		0.005116	0.0015	0.003616
z	0.0032235		Ò	0.0011		0.0083395	0.0026	0.0057395
3	0.003255		Ù	0.0012		0.0115945	0.0038	0.0077945
j	0.0035578		À	0.0015		0.0151523	0.0053	0.0098523
A	0.004851		È	0.0042		0.0200033	0.0095	0.0105033
H	0.0054629		Q	0.0045		0.0254662	0.014	0.0114662
g	0.0055449		Z	0.0085		0.0310111	0.0225	0.0085111
K	0.0057405		F	0.0101		0.0367516	0.0326	0.0041516
p	0.0062388		B	0.0105		0.0429904	0.0431	0.0001096
n	0.0111403		H	0.0143		0.0541307	0.0574	0.0032693
s	0.0170512		G	0.0165		0.0711819	0.0739	0.0027181
7	0.0171269		V	0.0175		0.0883088	0.0914	0.0030912
C	0.0179406		M	0.0287		0.1062494	0.1201	0.0138506
2	0.021183		P	0.0296		0.1274324	0.1497	0.0222676
m	0.0259395		U	0.0316		0.1533719	0.1813	0.0279281
4	0.0341149		D	0.0339		0.1874868	0.2152	0.0277132
k	0.0386064		C	0.043		0.2260932	0.2582	0.0321068
y	0.0423787		S	0.0548		0.2684719	0.313	0.0445281
h	0.0626155		L	0.057		0.3310874	0.37	0.0389126
8	0.065265		R	0.0619		0.3963524	0.4319	0.0355476
e	0.0672205		T	0.0697		0.4635729	0.5016	0.0380271
1	0.0697249		N	0.0702		0.5332978	0.5718	0.0385022
c	0.0867445		O	0.0997		0.6200423	0.6715	0.0514577
a	0.0914505		I	0.1018		0.7114928	0.7733	0.0618072
9	0.1100093		A	0.1085		0.8215021	0.8818	0.0602979
o	0.1580528		E	0.1149		0.9795549	0.9967	0.0171451
							Difference	**0.0618072**
			Count	27			Critical D	0.003583195
							p-value	0.00

Fig. 3. Sample Calculations for the VM and Italian.

Results and Analysis

Language	Disimilarity Value
Italian	0.0618
Portugese	0.0645
Indonesian	0.0677
Latin	0.0734
Spanish	0.0742
German	0.0750
French	0.0758
English	0.0962
Irish	0.1091
Hebrew	0.1176
Greek	0.1438
Belarusian	0.2110
Czech	0.2280

Fig. 4. Dissimilarity values to the VM for thirteen languages, sorted in ascending order of dissimilarity. Blue text indicates that the dissimilarity values were greater than the critical value; all languages demonstrate this. The least dissimilar value is 0.0618.

Language	p-value
Italian	0.00000
Portugese	0.00000
Indonesian	0.00000
Latin	0.00000
Spanish	0.00000
German	0.00000
French	0.00000
English	0.00000
Irish	0.00000
Hebrew	0.00000
Greek	0.00000
Belarusian	0.00000
Czech	0.00000

Fig. 5. The table shows p-values from the K-S test for thirteen languages. The languages are in the exact same order as in Fig. 2. All languages show $p < \alpha$, where $\alpha = 0.05$.

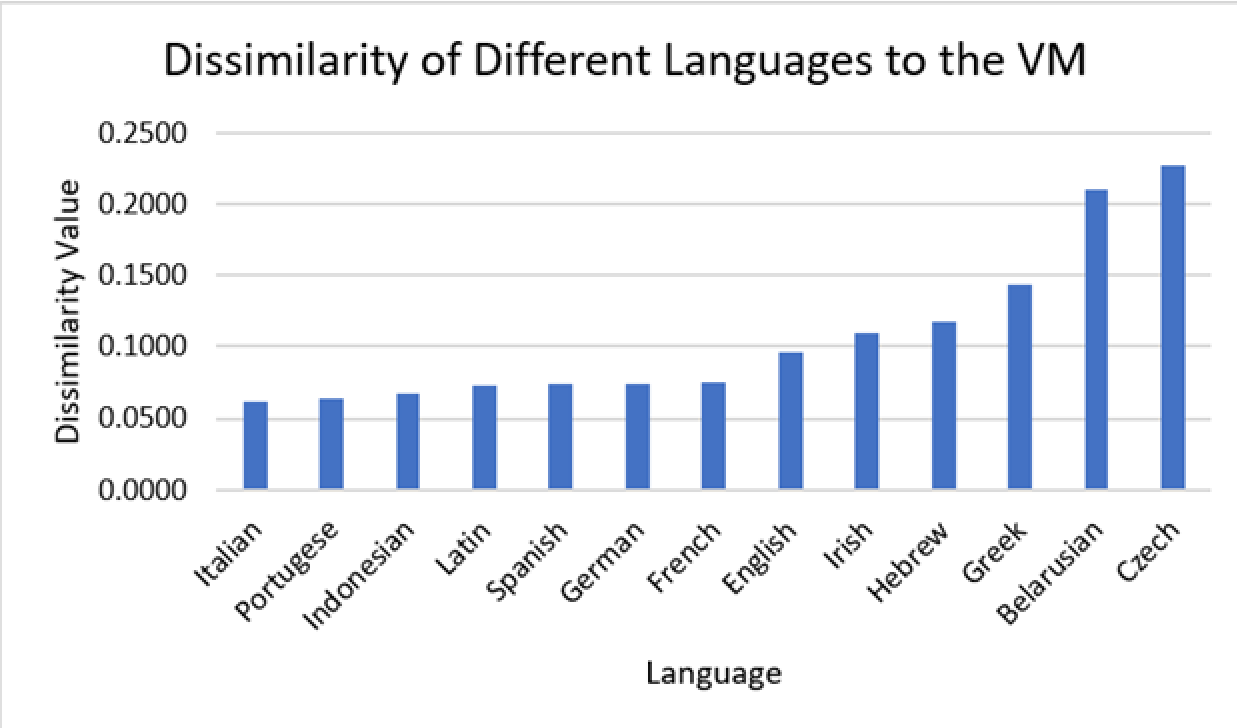

Figure 6. Dissimilarity of each language to the VM.

After calculating each p-value, since $p < \alpha$, it seemed that each test produced sufficient evidence to reject the null hypothesis. In other words, it seemed that there was not enough evidence that the distribution between any language was similar to the distribution of language of the VM. In all cases, $D_n > D_{crit}$ as well. This seems to support the first idea, and it would seem that no language is similar at all to the VM. However, these next few paragraphs explain why this interpretation is incorrect.

First, no two languages have the exact same distributions, so it is not expected for the VM to match any language distribution perfectly. This is important to note especially since the VM is from centuries before these languages.

Second, the formula for calculating the p-value is based on the sample sizes, which were extremely (more than usual) large in this case. For example, Italian had a sample size of 1,641,351 characters. As the sample size increases to this level, the p-value behind the K-S test expects that the two distributions will be more and more similar, i.e. D_n will be extremely close to the critical value (as is usual for statistics tests). This explains why the p-values were so small for this test, yet why D_n (whose formula depends on the frequencies rather than sample sizes of two samples), was greater than the D_{crit}. As highlighted more in "Limitations to this Study" (see below), there was no choice but to use large sample sizes to avoid bias, and these were simply factored largely into the p-values.

Third, the D_n values were extremely low, and statistically close to the D_{crit} values, as one can see in the sample calculations for the VM and Italian in Figure 3. This strengthens the idea that while the VM is not *perfectly* similar to any language, it is still relatively similar in distribution with some of them.

Overall, because of the large sample sizes, the p-values are not the best indicator of similarity in this case. They simply reaffirm that languages are not perfectly similar to one another in distribution. The D_n values are also greater than the D_{crit} values, but not significantly so. They show that the VM does not have the exact same distribution of any other language, but is still similar to quite a few of them. The takeaway are the relative D_n values, which show relative similarity and do not depend on the large sample sizes.

There are clearly some languages that are significantly more similar to the VM than others. As shown by the table, Italian shows the least dissimilarity to the VM. Portuguese is also similar to the manuscript, but the language does not quite make sense geographically based on the "Introduction" (see above). However, it is important to note that Portuguese is extremely closely related to Italian and is derived from Vulgar Latin. In fact, though German and Indonesian seem to have similar distributions to the VM, of the top seven languages with the most similar distributions to the VM, five are derived from Vulgar Latin. This strongly suggests that a Vulgar Latin derivative is behind the manuscript.

Belarusian, Greek, Hebrew, and Irish were the most dissimilar languages to the VM (besides Czech, which has already been discussed). The dissimilarity value for Czech is nearly five times the dissimilarity value for Latin. Based on this evidence, it is probable that these five languages are not as relevant to the language behind the VM. It is important to note that these languages have roots other than Vulgar Latin, again supporting the notion that the language behind the VM comes from Latin.

These results indicate that Latin and Italian, or some language derived from these Romance languages, is the principal language behind the VM. Because Classical Latin and Vulgar Latin are both branches of Latin and Italian is derived from Vulgar Latin, the evidence strongly suggests that a Vulgar Latin derivative is behind the VM.

Limitations to this Study

In this study, I used modern-day versions of all 13 languages for the purpose of getting accurate letter frequencies from a text of many genres. These letter frequencies have all been taken from one website, implying consistency when sampling across languages. It is difficult to get letter frequencies from these languages from the Medieval Times, and it is even more difficult to get unbiased letter frequencies for each alphabet, because letter frequencies differ based on the type of text used. Therefore, to avoid bias and be as consistent as possible, I used a modern-day version of these languages from which letter frequencies from all genres of text have been calculated. Because there is now a direction to pursue, future studies can use medieval alphabets of these suggested languages to conduct further analyses.

Conclusion

The language behind the VM is usually seen in a category of its own, but this study shows that there is great similarity between it and Vulgar Latin. Through an application of the K-S test, I was able to quantitatively compare the source language behind the Voynich manuscript to thirteen [13] languages commonly known today: Classical Latin, Italian, Indonesian, Portuguese, German, Spanish, French, English, Irish, Hebrew, Greek, Belorussian, and Czech. Of the top seven least dissimilar languages, five were Vulgar Latin derivatives, suggesting that the language behind the VM is a form of one of these Romance languages and/or Vulgar Latin. Czech, Belorussian, Greek, Hebrew, and Irish were the least similar to the manuscript, reinforcing the previous statement. Ultimately, this research may guide future endeavors at deciphering the Voynich Manuscript by providing a logically sound starting point, Vulgar Latin.

Acknowledgements

I would like to thank Principal Investigator Mr. Robert Downing for his guidance on this project. I would also like to thank the Aspiring Scholars Directed Research Program at Fremont STEM.

References

1. Voynich Manuscript. (2020, April 1). Retrieved from https://beinecke.library.yale.edu/collections/highlights/voynich-manuscript
2. The Voynich Manuscript [digital image]. Retrieved from https://archive.org/details/TheVoynichManuscript/page/n23/mode/2up
3. Zandbergen, R. (2020, April 17). Text Analysis - Transliteration of the Text. Retrieved from http://www.voynich.nu/transcr.html
4. Lin. (2010). Historical Linguistics: a history of sound changes from Vulgar Latin to French and Spanish [PowerPoint slides]. Retrieved from http://udel.edu/~dlarsen/ling203/Slides/History%20of%20Romance%20Languages.pdf.
5. Moulton, W. G., & Buccini, A. F. (2020, February 13). Germanic languages. Retrieved from https://www.britannica.com/topic/Germanic-languages
6. Greene, D. (2017, April 18). Linguistic characteristics of the Insular Celtic tongues. Retrieved from https://www.britannica.com/topic/Celtic-languages/Linguistic-characteristics-of-the-Insular-Celtic-tongues#ref604017
7. Saxey, R. (n.d.). The Greek Language Through Time. Retrieved from http://linguistics.byu.edu/classes/Ling450ch/reports/greek.html

8. Christensen, D. (1999, August 7). The Semitic Language Family. Retrieved from https://linguistics.byu.edu/classes/Ling450ch/reports/semitic
9. Poulsen, T. M., & East, W. G. (2020, April 2). People. Retrieved from https://www.britannica.com/place/Europe/People#ref309461
10. The Editors of Encyclopaedia Britannica. (2017, April 20). Czech language. Retrieved from https://www.britannica.com/topic/Czech-language
11. Blaylock, N. (1999, September 6). The Austronesian Language Family. Retrieved from https://linguistics.byu.edu/classes/Ling450ch/reports/austronesian.html
12. [Transliteration file]. Retrieved from http://www.voynich.nu/data/voyn_101.txt
13. Trost, S. (2020). WordCreator [Computer software]. Retrieved from https://www.sttmedia.com/wordcreator-download
14. Trost, S. (n.d.). WordCreator. Retrieved from https://www.sttmedia.com/characterfrequency-latin
15. Trost, S. (n.d.). WordCreator. Retrieved from https://www.sttmedia.com/characterfrequency-italian
16. Trost, S. (n.d.). WordCreator. Retrieved from https://www.sttmedia.com/characterfrequency-indonesian
17. Trost, S. (n.d.). WordCreator. Retrieved from https://www.sttmedia.com/characterfrequency-portuguese
18. Trost, S. (n.d.). WordCreator. Retrieved from https://www.sttmedia.com/characterfrequency-german
19. Trost, S. (n.d.). WordCreator. Retrieved from https://www.sttmedia.com/characterfrequency-spanish
20. Trost, S. (n.d.). WordCreator. Retrieved from https://www.sttmedia.com/characterfrequency-french
21. Trost, S. (n.d.). WordCreator. Retrieved from https://www.sttmedia.com/characterfrequency-english
22. Trost, S. (n.d.). WordCreator. Retrieved from https://www.sttmedia.com/characterfrequency-irish
23. Trost, S. (n.d.). WordCreator. Retrieved from https://www.sttmedia.com/characterfrequency-hebrew
24. Trost, S. (n.d.). WordCreator. Retrieved from https://www.sttmedia.com/characterfrequency-greek
25. Trost, S. (n.d.). WordCreator. Retrieved from https://www.sttmedia.com/characterfrequency-belorussian
26. Trost, S. (n.d.). WordCreator. Retrieved from https://www.sttmedia.com/characterfrequency-czech
27. Massa, S. (2016). Lecture 13: Kolmogorov Smirnov Test & Power of Tests [PowerPoint slides]. Retrieved from http://www.stats.ox.ac.uk/~massa/Lecture%2013.pdf

28. Kolmogorov-Smirnov Test. (n.d.). Retrieved from http://www.physics.csbsju.edu/stats/KS-test.html
29. Zaiontz, C. (n.d.). Two-Sample Kolmogorov-Smirnov Test. Retrieved from http://www.real-statistics.com/non-parametric-tests/goodness-of-fit-tests/two-sample-kolmogorov-smirnov-test/
30. The P-values for the Popular Distributions [Online Calculator]. Retrieved from https://home.ubalt.edu/ntsbarsh/Business-stat/otherapplets/pvalues.htm#rkstwo