Analysis of Letter Frequency Distribution in the Voynich Manuscript

Grzegorz Jaśkiewicz

Warsaw University of Technology
The Faculty of Electronics and Information Technology,
ul. Nowowiejska 15/19 00-665 Warsaw Poland
grzegorz@jaskiewi.cz

Abstract. The Voynich manuscript is one of the biggest mysteries in linguistic science. Although a lot of researches are being made, the author, the origin and the content of the manuscript still remain unknown. In this work letter frequency distributions of about 300 languages were compared to one of the language in the Voynich manuscript. The study shows the most similar languages according to this characteristics of a natural language.

Keywords: Letter frequency distribution, Voynich, statistics, linguistics, Wikipedia

1 Introduction

The Voynich manuscript is a book handwritten on 240 vellum pages, rich in illustrations. The book has its name after the Polish-Lithuanian-American book dealer - Wilfrid Michael Voynich, who acquired it in 1912. Despite many studies on the Voynich manuscript, the author, the content of the script and even the language remains unknown. Handwritten letters in Voynich manuscript do not resemble any alphabet known to human (see figure 1).

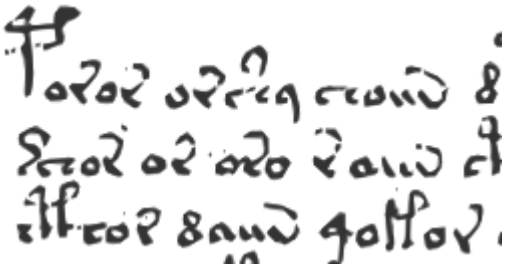

Fig. 1. Handwritten letters in Voynich manuscript

CONCURRENCY, SPECIFICATION AND PROGRAMMING
M. Szczuka et al. (eds.): Proceedings of the international workshop CS&P 2011
September 28-30, Pu tusk, Poland, pp. 250-261

The research performed with the C_{14} dating on manuscript's vellum shows that the manuscript was created between 1404 and 1438. There are many hypothesis about the possible content of the Voynich manuscript. All of them can be roughly divided into 3 categories [7].

1. Ciphertext - the manuscript is ciphered with some cipher.
2. Synthetic language - the manuscript is written in a synthetic language (like Esperanto).
3. Exotic natural language - the manuscript is written in a natural language in plain with an invented alphabet.

In this study the third hypothesis is explicitly used. Text samples of many different languages were compared to the Voynich manuscript in order to designate the languages which are the most similar to the language that is used in the manuscript.

The previous research on the language in the Voynich manuscript based on "exotic language" hypothesis, carried out by Zbigniew Banasik, suggested that it may come from north-eastern Asia and is a plaintext written in the Manchu language [2]. The author of this study proposed a translation of several words into English.

Other research shows that the manuscript has a linguistic nature [6] [9]. It conforms to the Zipf's law [12].

Dr. Leo Levitov's analysis suggests that the Voynich manuscript may be a liturgical manual for the Cathar religion written in ciphertext [8] in order to deceive the Inquisition. However, this hypothesis has been strongly criticized.

Dr. Jacques Guy, a linguist, suggested that the Voynich manuscript has got a similar word structure to many language families of central and east Asia. Those languages include Sino-Tibetan and Tai language family [4].

In spite of a fact that the volume of the Voynich manuscript is relatively large, it doesn't contain much text. Therefore it is impossible to use many algorithms based on statistics, data mining or artificial intelligence. Such algorithms were used to extract information from documents were knowledge about a language was only partial e.g. a Sumerian cuneiform script. Simple algorithms that do not require much data, can be used to analyze the manuscript. In this work a statistical analysis of a letter frequency distribution was used to find similar languages to the one that was used in the Voynich manuscript.

The research involving linguistic studies often compare an unknown language to the well-known language by its structure and syntax. Such comparison is precise, however it is limited by the knowledge of a researcher. In this work a simple characterization of language was used to compare many languages. This allowed to automate the whole process and to increase the scope of comparisons at cost of accuracy of a single comparison.

2 Letter Frequency Distribution

The letter frequency distribution for a given text sample is a function which assigns each letter a frequency of its occurrence in that the text sample. The

text sample D is a sequence of letters over alphabet Σ.

$$D = (l_i)_0^n \in \Sigma^\star$$

The letter frequency distribution could be defined as

$$f_D(l) = \frac{card(l' : l' \in D \wedge l' = l)}{card(D)}$$

A letter frequency distribution analysis has got various applications in different domains:

- cryptanalysis - a letter frequency distribution is a tool used to break simple ciphers like substitution ciphers or transposition ciphers,
- data compression - a study of a letter frequency distribution is also used in modern data compression techniques e.g. the Huffman coding,
- usability design - the Dvorak keyboard placement is based upon the letter frequency distribution in English language,
- computational linguistics - a distribution of pairs and triples of letters may be used to automatically recognize a language of an unknown document.

It is easy to observe that there are countably many symbols in all alphabets all over the world. Therefore any letter frequency distribution has got a discrete domain and it could be described by a single sequence of real-valued numbers. Such sequences can form a Banach space with a well-defined distance function, depending on additional assumptions about analyzed sequences [10].

ℓ_1 is the space of sequences $(a_n)_{n=1}^{\infty}$ which satisfy the condition

$$\sum_{i=1}^{\infty} |a_i| < \infty$$

in this space, the distance is defined in following way

$$dist(a, b) = \sum_{i=1}^{\infty} |a_i - b_i|$$

We could assume that any single language has the finite set of letters. Therefore the distribution for a single text sample in given language is zero almost everywhere. The corresponding space of sequences is known as a_{00} space. It is easy to check that

$$a_{00} \subset \ell_1 \tag{1}$$

Therefore distance function from ℓ_1 space is still valid in a_{00} space. In the a_{00} space distance could be introduced in many other ways, however this is not a goal of this study and the knowledge of (1) is enough.

3 Experiment Set-up

The goal of the experiment was to find the languages known to human that may be similar to the language used in the Voynich manuscript. A large corpora of texts in different languages was needed to conduct such experiment. This corpora was built upon random texts retrieved from the Wikipedia.

The Wikipedia is an online encyclopedia containing knowledge on various topics. It has got a great number of human-made translations in nearly 300 languages including dead languages like Latin or Old Church Slavonic and artificial languages like Esperanto or Volapük. There are even non-official Wikipedias written in Klingon language by Star Trek fans - those were not considered, as their structure differs significantly from regular Wikipedia.

The amount of articles in various language versions of Wikipedia differs significantly. The biggest language version is an English one containing over 3.5 mln of articles. The second language version of Wikipedia is German and the third one is French. The Polish Wikipedia is relatively big - it's on the fifth place in terms of article count with more then 800.000 of articles. The distribution of Wikipedias sizes has been shown on a figure 2. It can be clearly seen that language versions of Wikipedia having article count in range 100 - 10.000 represent vast majority of language versions. Due to this fact the decision was made to sample each language version with 100 randomly selected articles and combine them, in order to create single text sample for selected language.

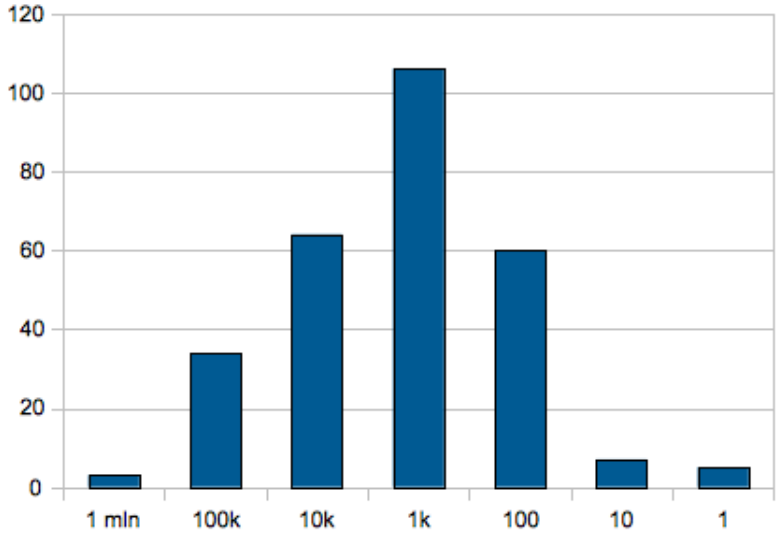

Fig. 2. A distribution of article count in different language versions of Wikipedia

Each language version of Wikipedia utilizes the same underlying software framework to manage content. Wikipedean articles are presented to users in the same fashion regardless of the language version. Even a structure of HTML page has got common elements for each version. It is very convenient to use this fact during creating a screen-scraper. Following assumptions about HTML structure of Wikipedean page were made:

- Content of an articles is always in a HTML div with the same identifier regardless of the language version;
- Each Wikipedia has a button for selecting a random article. This button resides in a HTML div with the same identifier regardless of language version.

With those assumptions the screen-scraper was written in Java in order to retrieve random articles from each language version of Wikipedia. The HtmlUnit library was used to mimic a human clicking on hyperlinks.

While running a data retrieval procedure, it turned out that one small fraction of all considered Wikipedias failed to satisfy the assumptions. This fraction was tiny (approx. 20 instances) and screen-scraper was modified to accommodate them. However, only one of Wikipedias has completely no *Random Article* button. This version of Wikipedia was rejected.

There are available to the public several transcriptions of the Voynich manuscript into the ASCII encoding. Those transcriptions differ slightly as it is not always clear if some glyph is a new letter or a ligature. The transcription used in this research is freely available on the Internet [3].

4 Experiments

The downloaded articles are not always entirely written in a desired language. They are usually contaminated by the English language. This phenomenon is also visible in the spoken language. After having all necessary data retrieved from the Wikipedia, the quality of the results was tested by sampling the text corpora in a random language. Some languages that contained no latin characters were selected for the evaluation, so any latin character was treated as undesired one. The ratio of the undesired characters to all text was presented in figure 3.

It is possible that the languages, which have contact with the western culture can assimilate more foreign words. The perfectly pure sample of any language could not be obtained from Wikipedia due to the culture assimilation process - visible especially well on a worldwide communication medium like the Internet. Languages evolve throughout centuries and a character frequency distribution may change. It would be perfect to have samples of all languages from the era when the Voynich manuscript has been written. However, for the sake of this experiment available samples have been used and an error have been estimated by mean calculated on tested languages, which was 3%.

A character frequency distribution should converge to some function as the amount of evaluated text tends to the infinity. However, the estimation of speed of this convergence is a big problem as a probabilities of occurrence particular

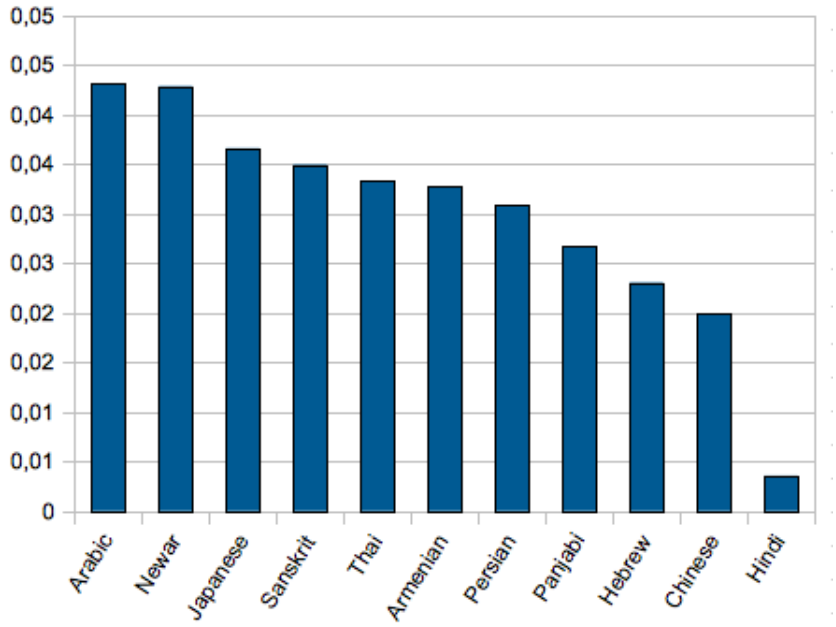

Fig. 3. Ratio of latin characters in sample to sample size

letter on a particular position in any text are not independent. Therefore, the speed of this convergence was checked empirically. The ℓ_1 measure (2) was used to measure the distance between two distributions.

$$d(f, g) = \sum_{x \in U} |f(x) - g(x)| \tag{2}$$

The English language with letter frequency taken from [1] was chosen as a benchmark. To test the convergence of the letter frequencies, the consecutive prefixes of a sample text were taken, the character frequency distribution was evaluated on those prefixes and results were compared to benchmark by the ℓ_1 standard distance function. Three books were downloaded from the *Project Gutenberg* site to conduct this test. These books were:

- *The Adventures of Sherlock Holmes* written by Sir Arthur Conan Doyle
- *20000 Leagues Under the Seas* written by Jules Verne
- *Father Goriot* written by Honore de Balzac (English translation)

The result of this test is shown on figure 4. It can be clearly seen that the distribution of letter frequencies is close to the benchmark. However, it is not exactly the same distribution - for a sample which is big enough there is a difference, which could be bounded by 5% under the ℓ_1 norm. We will assume that this error margin will hold for data retrieved from the Wikipedia.

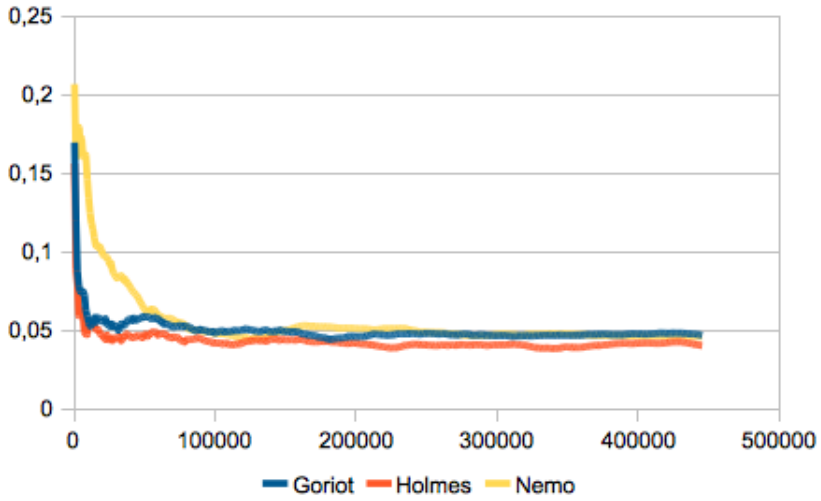

Fig. 4. The difference in letter frequency distribution between the benchmark and a sample

The measure ℓ_1 itself does not work well when it comes to comparing two different languages, because both can have different charsets. It may even happen that two text samples in the same language can have different charsets, e.g. texts in Serbian language can appear in a latin alphabet as well as in cyrillic one. An alphabet in the Voynich manuscript is completely different from any known charset, so such approach would have failed completely. In order to accommodate this issue, the different measure was defined

$$d(f,g) = \sum_{i=1}^{\infty} a^i \cdot |f(\sigma_f(i)) - g(\sigma_g(i))| \tag{3}$$

where σ_f is such injective assignment $\sigma_f : N \to Dom(f)$ which satisfies

$$f(\sigma_f(k)) \geq f(\sigma_f(k+1))$$

This assignment is well defined as $card(Dom(f)) < \aleph_0$. In equation (3) a suppressing factor $a \in (0,1)$ was introduced to reduce the error caused by the differences in size of two different charsets and occurrence of letters from different charsets. It is easy to check that (3) is a still valid distance function.

Before advancing to the final experiment, the test of measure (3) was made. 23 samples of different languages where compared all to all using (3) measure. Each comparison resulted in a single number describing similarity of two language samples. Lower number means more similar letter frequency distributions. Zero value means the exact match of two distributions. The results of those tests are shown in figure 5.

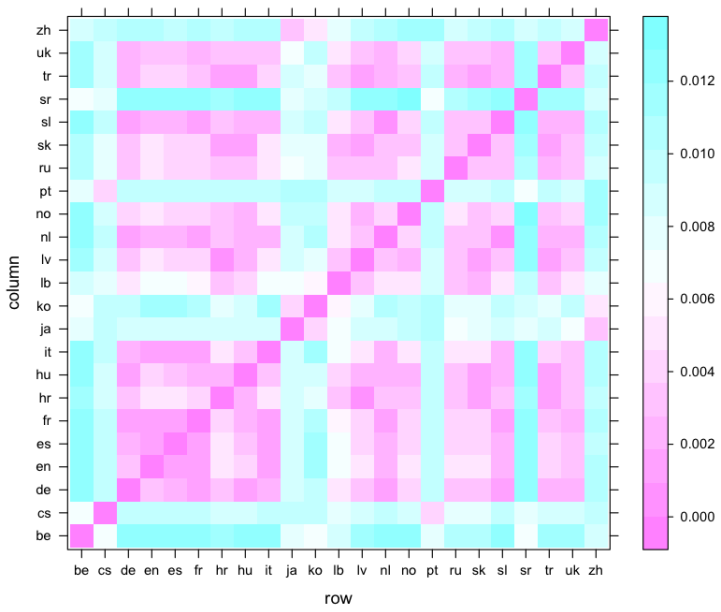

Fig. 5. Similarity of languages based on letter frequency

We can see that two languages, which are rich in vowels (e.g. French, Spanish), tend to have a lower (3) distance than two languages poor in vowels (e.g. Serbian, Czech). Languages from the same language family (e.g. Slavic) tend to have a lower distance. Unfortunately, it is not the rule - sometimes completely different languages are similar in measure (3).

5 Conclusion

The final test was carried out to compare the transcription of the Voynich manuscript to each text sample in different language using the measure (3). The top five matches are:

- Moldavian
- Karakalpak
- Kabardian Circassian
- Kannada
- Thai

The regions, where those languages exist, are marked on figure 6.

The first three results designate Caucasus region and other two the region of west Asia. The second match would explain similarity of the Voynich manuscript to Sanskrit and hypothesis stating that it has got its origin in far Asia. Both

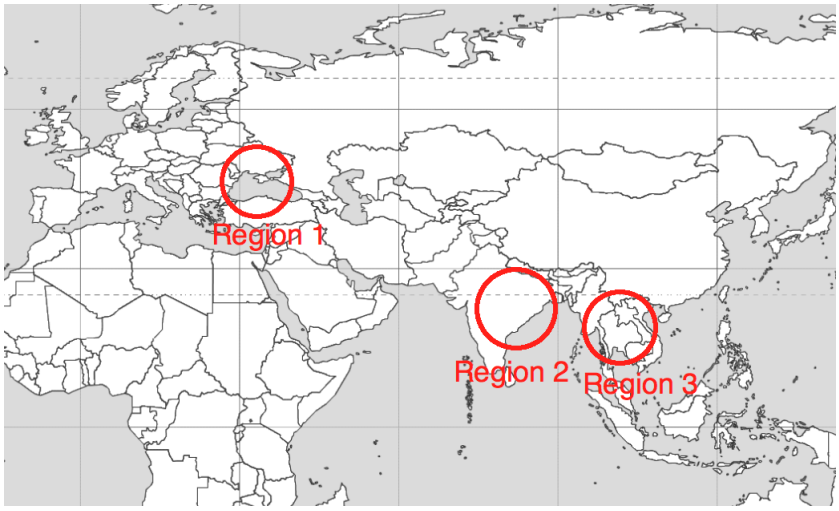

Fig. 6. Regions indicated by character distribution similarity

Asian matches designate the region near China, historically influenced by this country. Hypothesis stating that the Voynich manuscript may have Chinese roots would designate the same region. Also the fact that figures in manuscript are not typical for China could be explained - it could be created not in China, but nearby - in the region influenced by China, like Indochina region. Similarity to Thai language was also proposed by dr. Jaques Guy.

Those two matches are not distant in a world scale - it is possible that the manuscript may have been created somewhere between those regions. The data used for the purpose of this research depicts only the current state of the languages and it captures only the official status of a language - neither minor dialects nor historical language evolution are taken into consideration, so more precise region cannot be indicated.

Based on considerations from the section 3 a total difference between a letter frequency distribution for a given language and a letter frequency distribution calculated on text sample could be bounded by 8%. Therefore, difference between a letter frequency distribution of text sample and a letter frequency distribution of the language in Voynich manuscript could be bounded by 13% in ℓ_1 norm. To estimate this error under (3), some assumptions must be made how two distributions differs. In this study an assumption was made that each letter contributes the same value to overall error, which could be calculated by summing a geometric sequence.

$$e = \frac{b}{card(Dom(f_{voy}))} \cdot \frac{1 - c^{(card(Dom(f_{voy})))}}{1 - c} \tag{4}$$

where b is estimated error bound, c is suppressing factor equal $0,96$ and f_{voy} is a letter frequency distribution calculated for the Voynich manuscript. The transcription used in this study contains 26 glyphs, so

$$card(Dom(f_{voy})) = 26$$

and

$$e = 0,4\%$$

With such error estimate a list of possible languages reaches about 40. The most prominent matches indicate Asia, the other ones from the list indicate the languages existing in Europe.

Similarity to both language families could be just coincidence, but there is a theory that the Voynich manuscript was created by a traveller visiting China, who didn't know the Chinese language and alphabet [11]. Such traveller may have written down information that he learned in this region in invented alphabet. The traveller might have been European, as the manuscript was later discovered in Europe. Similarity to both Asian and European language fits this theory well. If an author of the manuscript was European, his language habits may have influenced a letter frequency distribution in the manuscript, resulting in some similarity to the author's native language.

The second conclusion drawn from this research would be the fact that language in the Voynich manuscript may be the language poor in vowels. When the letter frequency distribution from the manuscript is compared to the distributions of other languages, it behaves similarly to the languages poor in vowels. The observed values are significantly higher than those obtained by comparing a language rich in vowels.

To observe this fact the language from the manuscript was compared to languages rich in vowels - Swedish and French as well as languages poor in vowels - Serbian and Moldavian. Each comparison was carried out by comparing a letter frequency distribution of the selected language to distributions of all of the languages. In figure 7 there are shown cumulative histograms of obtained values for each language. Histogram for language from the Voynich manuscript is similar to histograms for languages poor in vowels.

This result acknowledges the outcome obtained by Jacques Guy using the Sukhotin algorithm [5] for vowel identification, where only 4 vowels were identified.

6 Future Works

With the list of languages similar to the Voynich manuscript it is worth to analyze deeper languages which are the most similar to language in the manuscript. Bigger and better text samples of selected languages could be obtained and more complex algorithms could be used to analyze and compare them.

Results of this study indicate that the language from the Voynich manuscript is based on Asian language - it is also possible that it was somehow influenced

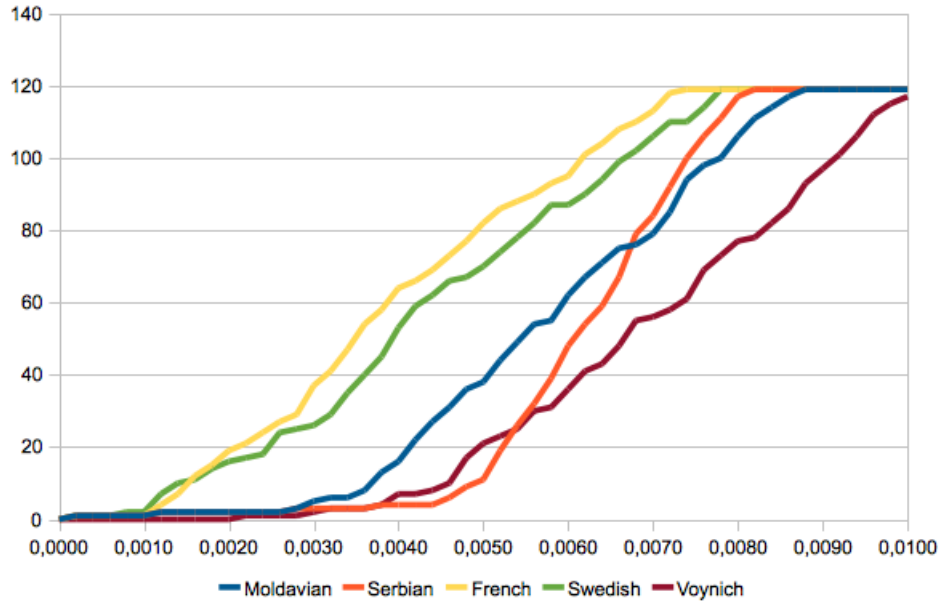

Fig. 7. Histograms of comparisons for all languages

by European languages. It still leaves many possibilities to consider, but with conjunction with historical research this area of speculations could be narrowed.

7 Acknowledgements

We would like to thank my mentor prof. Jarosław Arabas for advices provided while writing this article.

References

1. Beker, Henry; Piper, Fred, *Cipher Systems: The Protection of Communications*, Wiley-Interscience, p. 397, 1982
2. Zbigniew Banasik, Jorge Stolfi, *Zbigniew Banasik's Manchu theory*, `http://www.ic.unicamp.br/~stolfi/voynich/04-05-20-manchu-theo`, 2004
3. *Voynich Manuscript Transcription*, `http://voynichcentral.com/transcriptions/Voynich-101/index.html`
4. Jacques Guy, *Statistical Properties of Two Folios of the Voynich Manuscript*, Cryptologia, XV, number 4, pp. 207-218, July, 1991.
5. Jacques Guy, *Vowel identification: an old (but good) algorithm*, Cryptologia, XV, number 3, 1991
6. Gabriel Landini, *Evidence of linguistic structure in the Voynich Manuscript using spectral analysis*, Cryptologia, 2001

7. Landini, Gabriel, *A Well-kept Secret of Mediaeval Science: the Voynich manuscript*, Journal of the University of Birmigham Medical and Dental Graduates Society, 1998
8. Leo Levitov, *Solution of the Voynich Manuscript: A Liturgical Manual for the Endura Rite of the Cathari Heresy, the Cult of Isis*, Aegean Park Press, 1987
9. Sravana Reddy, Kevin Knight, *What We Know About The Voynich Manuscript*, The Natural Language Group at the USC Information Sciences Institute
10. Walter Rudin, *Functional Analysis*, PWN, 2009
11. *Voynich Manuscript mailing list archives*, `http://www.voynich.net/Arch/2004/10/msg00256.html`
12. Zipf G. K, *The Psycho-biology of Language*, Hought Mifflin Co, Boston, pp. 20-48, 1935