IPM named after M.V. Keldysh RAS

Future 2021 • Proceedings of the Conference

Yu.N. Orlov

Text language recognition methods on the example of the Voynich manuscript

Recommended form of bibliographic reference

Orlov Yu.N. Methods of text language recognition on the example of the Voynich manuscript // Designing the future. Problems of Digital Reality: Proceedings of the 4th International Conference (February 4-5, 2021, Moscow). - M .: IPM im. M.V. Keldysh, 2021. - S. 220-235. — **https://keldysh.ru/future/2021/20.pdf https://doi.org/10.20948/future-2021-20**

There is also a video of the speech

Text language recognition methods on the example of the Voynich manuscript

Yu.N. Orlov

Institute of Applied Mathematics. M.V. Keldysh RAS

Annotation. Statistical patterns of frequency distribution of letters in texts in European languages are studied. The level of reliability of the logarithmic approximation of the ordered distribution of frequencies for texts without vowels written in one alphabet in one and two languages is analyzed. Variants of the languages in which the Voynich Manuscript could have been written have been proposed. Spectral portraits of matrices of conditional probabilities of two-letter combinations for texts without vowels and the Voynich Manuscript are constructed.

Keywords: frequency distribution of letter combinations, groups European languages, Voynich Manuscript, spectral portrait

Language recognition methods and Voynich Manuscript analysis

Yu.N. Orlov

RAS Keldysh Institute of Applied Mathematics

Abstract. The statistical properties of letters frequencies in European litera ture texts are investigated. The determination of logarithmic dependence of let ters sequence for one-languge and two-language texts are examined. The pare of languages are suggested for Voynich Manuscript. The internal structure of Manuscript is considered. The spectral portraits of two-letters distribution are constructed.

Keywords: letters frequency distribution, European languages groups, Voynich Manuscript, spectral portrait

1. Introduction

The Voynich Manuscript (hereinafter MV) [1] is a manuscript dated by researchers of the 16th century. It consists of a sequence of characters treated as letters, from which the transcriptors extract 22 different characters. These characters are not elements of any known alphabets.

6. Mathematical models of the digital world

The volume of the manuscript is about 170 thousand characters. Currently, the manuscript is stored in the Yale University Library and has the status of a cryptographic puzzle.

Numerous studies to decipher this text have been carried out for more than a hundred years, but without success. The existing versions about the authorship, content, and language of the manuscript, a review of which can be found in [2–4], are not sufficiently convincingly supported by full-fledged statistical studies. The aim of the work, however, is not to decipher the manuscript. The vocabulary is not analyzed, so the semantic component of the text is not discussed in the work. The question is: is the MW a meaningful but encrypted text, and in what language, if so, is it written, or is it a hoax, i.e. meaningless character set? It may seem that deciphering the text is required for the answer, but this is not at all necessary. First, one should find out whether meaningful texts have some general statistical properties, without knowing which it is impossible to carry out the desired simulation. Studies carried out in [5] show that such properties

are available.

There is also no consensus on how many and what signs are in the MV. There is a so-called "European transcription" (EVA
[6]) mapping characters of the manuscript into the Latin alphabet. In addition, there is a transcription of Takahashi [7] - also in Latin, but with different frequencies.
lazy characters.

Numerous hypotheses about the structure of MFs have been proposed by various researchers. The Manuscript was believed to:

- written with a rearrangement of letters;
– two characters of some known alphabet correspond to one character of the manuscript;
– there is a manuscript-key, without which it is impossible to read the text, because the same symbols in different parts of the manuscript correspond to different letters;
– the manuscript is an encrypted bilingual text;
- initially, vowels were removed from the meaningful text;
– the text contains false spaces between words.

It seems that the idea of deciphering the MW through the analysis of "words" without identifying the language is not productive. It is possible to discuss meaningfully only the statistics of individual characters, assuming that the characters are letters, or, if their statistics are “non-alphabetic,” that they are
syllables, or at least some of them are syllables.

Studies carried out in [5] showed that the distribution of text symbols by frequency of occurrence is a stable characteristic not of the author or the subject of the text, but of the language. It is assumed that the distributions of a mixture of texts in different languages will turn out to be just as stable, according to the level of determination of which relative to a certain model

distribution, it will be possible to judge the share participation of different languages in the writing of such bilingual texts.

This work is devoted to the study of the invariant properties of European languages. The following statistics are used to find the invariants: the distance between the distributions of ordered empirical frequencies of letter combinations in the L1 norm; the level of determination of the logarithmic approximation of one-letter distributions for texts without vowels; the Hurst exponent for a series of the number of letters enclosed between the two most frequently occurring identical letters; spectral portrait of the matrix of two-letter combinations. The listed indicators made it possible to carry out a formal clustering of the languages of the Indo-European family according to the language groups that coincided with the groups that were formed on the basis of historical and linguistic research.

2. MW symbol statistics and frequency approximation

The subsequent analysis will be aimed at constructing the distribution of MV symbols by frequency of occurrence, comparing it with similar distributions in European languages, identifying deviations from the level of determination of the approximating dependence, and determining how large the distance between the actual frequency distribution and its approximation in the histogram norm L1.

The logarithmic symbol distribution model was developed by S.M. Gusein-Zade in [8] under the assumption of a constant distribution density

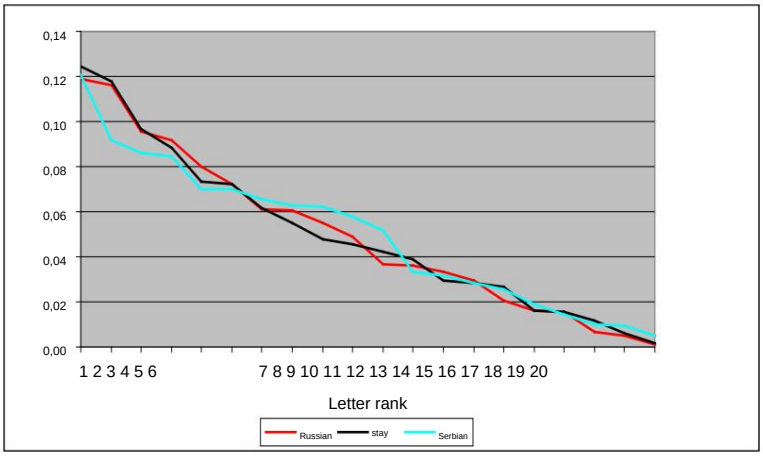
division of a random point (p is the ith n - dimensional simplex $\ddot{y} = \sum_{i=1}^{n} p_i$ on the

estimate the completeness of the text alphabet in texts in which it is modified and applied to frequency of the letter of the text [5], this model was modified and applied to

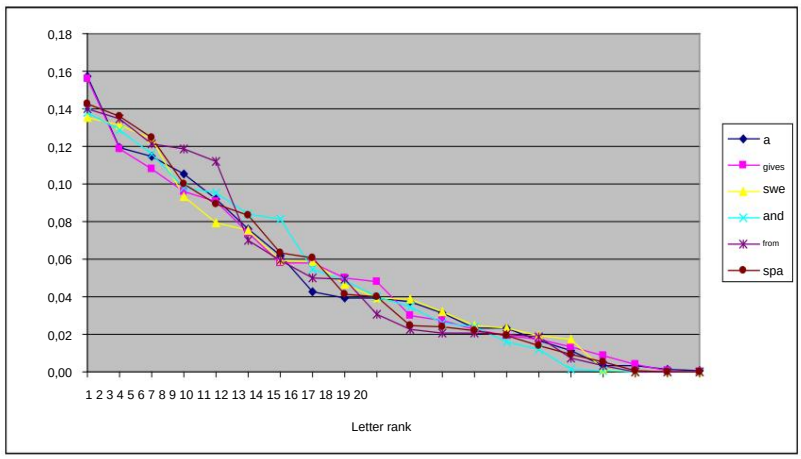
$$f(k) = \frac{1}{n + o} \log \frac{n \ddot{y}!}{k^{n \ddot{y}}} + 1 \ddot{y} \quad (1)$$

In this formula , n is the number of letters in the alphabet, and the parameter o is the nearest integer to the selected value, which corresponds to the smallest error in the approximation of the actual distribution according to formula (1). The meaning of this parameter is that for the text under consideration the most adequate alphabet is the number of characters in which is $n + o$. For the Russian, German, English, and Hungarian languages, the empirical dependencies in Figs. 1-2 are best modeled by dependence (1), in which $o = 0$. For French, Spanish and Italian $o = \ddot{y}3$, for Danish and Swedish $o = \ddot{y}1$. For Finnish $o = \ddot{y}6$ Estonian $o = \ddot{y}4$.

for



Rice. 1. Ordered character frequencies in Cyrillic texts without
advertisements



Rice. 2. Ordered character frequencies in Latin texts without
advertisements

The distances between the distributions of characters in the Cyrillic texts for the Slavic group show that the Russian, Bulgarian, and Serbian languages are related: Russian and Bulgarian are closest (distance 0.06), Russian and Serbian, like Bulgarian and Serbian, are one from the other by 0.12. Note that the Greek language in Cyrillic transcription is more than 0.20 apart from them and in this sense is not similar
none of the Slavic languages.

For texts in Latin, the distances between the distributions of ordered frequencies form clusters in the sense of proximity to each other in the L1 norm in accordance with the language groups. So, for example, the statistics of Danish and Swedish are quite close, but they differ from French and Italian, which are also close to each other. Czech and Croatian have similar statistics, which differ from those of the other mentioned groups. This shows that the languages of the Indo-European family, united in groups or subgroups, have close

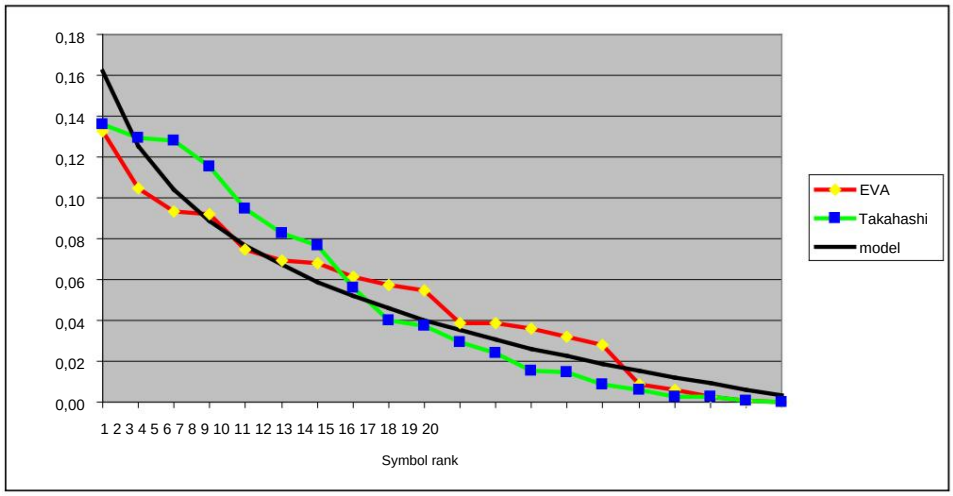
statistical properties. Distances in the L1 norm between distributions from the same language group vary within a rather narrow range of 0.08-0.13, and between different subgroups they are 0.14-0.22.

Although these distributions are close in terms of the level of determination of the approximating dependence (0.93), they differ significantly in details. The EVA graph (red broken line) is typical for the Germanic group of languages, more precisely for the West Germanic subgroup, and the Takahashi graph (green broken line) is for Slavic and Romance languages, as well as for Germanic, but for the North Germanic subgroup. The distance between these two transcriptions, whose frequencies are ordered in descending order, in the L1 norm is 0.26, which is about three times greater than between the distributions of unvoiced texts from the same language group, and 10 times greater than between texts with complete alphabet. This means that each of these transcriptions corresponds to a fundamentally different reading of the MV, so both of them cannot be used simultaneously to refine the statistics. Differences in transcriptions are connected, apparently, with the problem of recognizing the signs of the Manuscript, because not all of them can be interpreted unambiguously. The extent to which the signs of the manuscript are correctly recognized is not discussed here; only the statistical properties of the presented transcriptions are studied.

Most of the modern languages of the Indo-European family are characterized by a logarithmic dependence of the frequency of a letter on its rank with a reliability of more than 0.98. The level of determination of texts without vowels is somewhat lower, but also quite high - at the level of 0.96 (Fig. 2).

The actual distribution of ordered character frequencies for most texts differs from the logarithmic approximation in the L1 norm within 0.08-0.13, the distances between real distributions in the same language are in the same interval, regardless of which language it is . In this case, the 90% confidence interval is [0.085; 0.115].

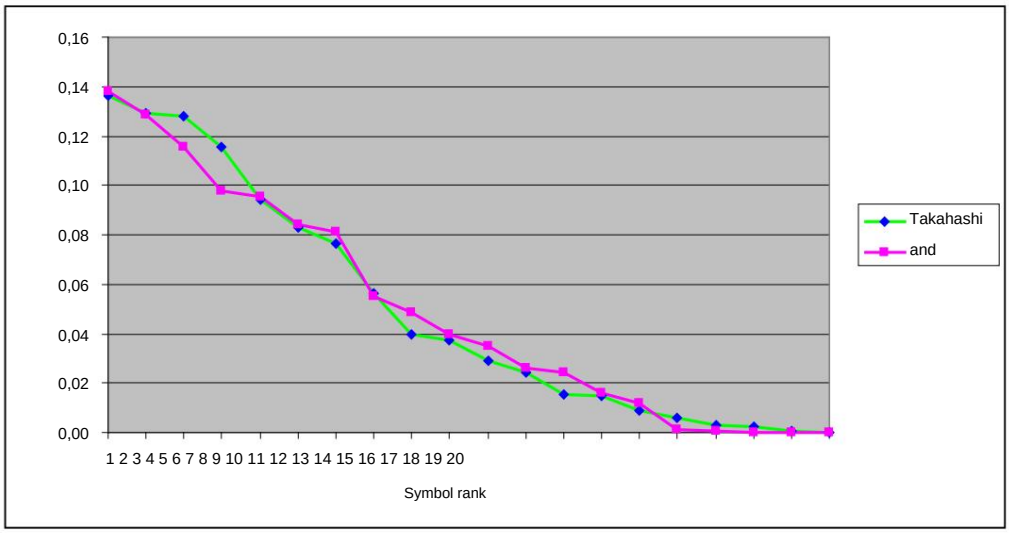
In relation to the transcription of EVA, for which $n = 22$, the best the approximation is achieved at $o = ÿ2$. This means that only 20 characters are actually used in the MV. Eliminating the two rarest symbols, we obtain a logarithmic approximation with a determination of 0.93 and a deviation in the L1 norm from the actual distribution at the level of 0.167 (see Fig. 3). The same observation is true for the transcription of Takahashi.



Rice. 3. Ordered frequencies of two MV transcriptions and logarithmic approximation

However, the deviations in the L1 norm of the corresponding approximations for both transcriptions of the Manuscript are approximately the same and equal to 0.17, which indicates the insufficient adequacy of the logarithmic model in relation to the text under consideration.

Let us now note that in most European languages the number of consonant letters is 20. It can be assumed that the manuscript under study was written in one of them, but without vowels. A necessary in a statistical sense, but, of course, not a sufficient condition for this is, firstly, the proximity of one of the distributions of transcriptions to the distribution of the selected language (the deviation in the L1 norm does not exceed 0.10) and, secondly, an approximately equal distance from transcription, and from the chosen language to the approximating model dependence (approximately 0.17). Of the languages of the Indo-European family, there is only one suitable in this regard - namely, Danish. It deviates from the model logarithmic dependence by 0.172, the Takahashi transcription deviates from it by 0.167, and the empirical distributions of the Manuscript and Danish languages deviate from each other by 0.083 (see Fig. 4).



Rice. 4. Frequency distributions of symbols in Takahashi transcription and in Danish texts without vowels

At the same time, the determination of the logarithmic approximation of the Danish language without vowels, as well as the transcription of Takahashi, is 0.93. The languages close to Danish, Swedish and Norwegian (Bokmål), are much less suitable for the role of the original MW language, since the distances between all the languages of the North Germanic group are the same and equal to 0.11 (differences appear only in the third decimal place), and the difference between Swedish and Norwegian from the indicated transcription is 0 .14 instead of 0.08 for Danish. For the transcription of EVA, there was no suitable language among the examined European ones.

So, one argument is given in favor of the fact that MW is written in some language without the use of vowels. An indirect confirmation of the thesis put forward is the presence in the manuscript of chains of three identical characters in a row, which were not found in the texts in Latin with the full alphabet, but turned out to be present in them after the removal of vowels. So, for example, in an English text without vowels, the frequency of occurrence of "bbb" is 3 10ÿ6, "lll" 4 10ÿ4, respectively , and "ttt" 8 10ÿ4, respectively. The Takahashi transcription also contains strings of triplets with similar frequencies: “ttt” 5 10-6, “lll” 2 10-5, “ooo” 5 10-5 and “eee” 8 10-4.

3. Using the Hurst indicator

Consider the text as a time series of a random variable (letter), and the letter takes values from a set called "alphabet". The length of a string of text characters that do not contain a pair of specific characters inside is a very important characteristic of the language, since its distribution is also stable.

The Hurst indicator is introduced as an indicator of the volatility of a time series and is defined as follows.

For a given time series *b(t),* a series *x(t)* = b(t +1) ÿ *b(t)* is constructed first differences and a moving average of increments over a sample of length *k* is introduced :

$$\bar{x}_t(k) = \frac{1}{k}\sum_{i=t-k+1}^{t} x(i)$$

Then the accumulated deviation from the mean (range) is calculated:

$$R(t,k) = \max_{1\le j\le k}\left(\sum_{i=t-k+1}^{t-k+j}(x(i)-\bar{x}_t(k))\right) - \min_{1\le j\le k}\left(\sum_{i=t-k+1}^{t-k+j}(x(i)-\bar{x}_t(k))\right).$$

The moving variance of the considered time is also calculated. variable series on a sample of length *k*

$$\sigma_x^2(t,k) = \frac{1}{k}\sum_{i=t-k+1}^{t}(x(i)-\bar{x}_t(k))^2$$

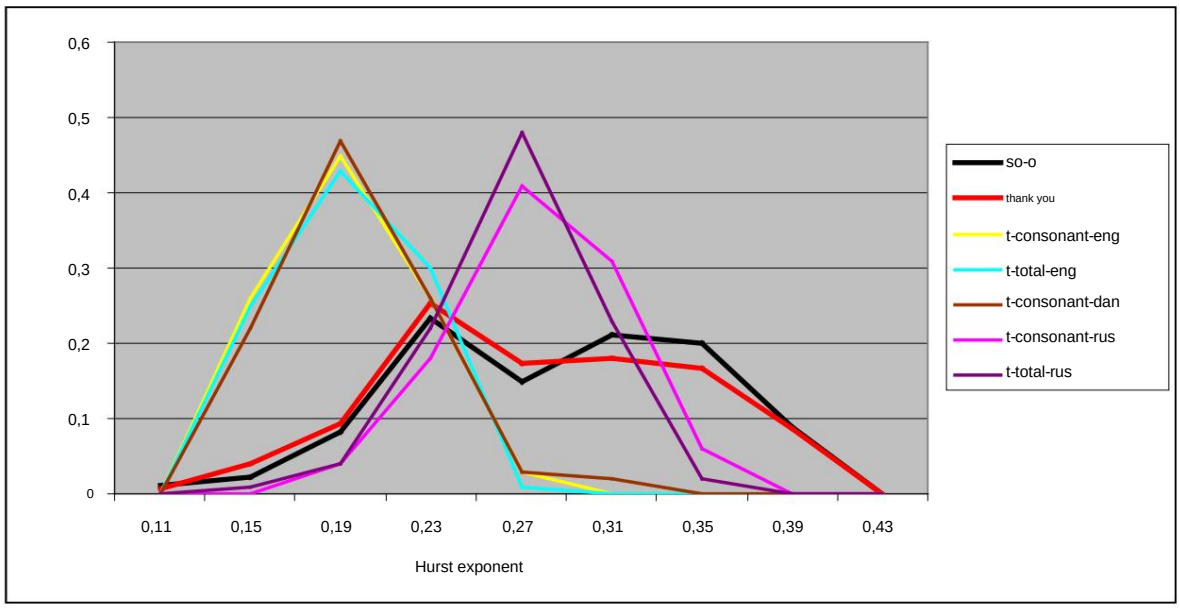
logarithm of the peak-to-noise ratio and its sample mean:

$$\psi(t,k) = \ln\frac{R(t,k)}{\sigma_x(t,k)}, \quad \bar{\psi}_N(t) = \frac{1}{N}\sum_{k=1}^{N}\psi(t,k)$$

The Hurst exponent $H_N(t)$ for a sample of length N at step t is determined as the regression coefficient of the value ÿ(t,k) on the logarithm of the sample length and is calculated by the formula:

$$H_N(t) = \frac{\sum_{k=1}^{N}(\psi(t,k)-\bar{\psi}_N(t))(\ln k - \ln N)}{\sum_{k=1}^{N}(\ln k - \ln N)^2} \qquad (2)$$

It turned out that the distances between the same letters, regardless of vowels, for all the languages under consideration form the so-called antipersistent series, since the Hurst exponent for a series of these distances is significantly less than the critical value of 0.5 corresponding to white noise. The distributions of the Hurst exponent, built on a sample of length *N* = 5000, are shown for some languages in fig. 5.



Rice. 5. Distributions of Hurst exponents for series of distances between the most frequently occurring letters in texts

It can be seen that the distributions in Fig. 5 for Russian and English have maxima at distinctly different points, while the distributions for Danish and English are practically the same. The distribution of the Hurst exponent is an indicator of the language group. It follows that the variant put forward in section 2 with Danish as the original language of the MS should be excluded, since the distributions of the Hurst exponent for the manuscript differ significantly from similar distributions for ordinary texts. For MW, this distribution is flatter and shifted to the right (black and red lines in Fig. 5), which indicates a greater randomness in the arrangement of characters than for texts in one of the European languages. This means that the statistics of the language of the manuscript differ from texts written in the same language. Consequently, the manuscript contains entries in various languages.

4. Analysis of bilingual texts

The dermination of texts without vowels at the level of 0.96 observed in most European languages can be reduced to 0.93, as in MW, if we assume that the text is written in two languages that have the same alphabet (for example, Latin), and this text after the removal of vowels and recoding turns into what we know as the Voynich Manuscript. At the same time, we assume that the same letters in different languages were not indicated in the manuscript by different symbols, which, of course, greatly narrows the search field. Note, however, that since the Takahashi transcription determination is higher than 0.9, the possibility of using different scripts is unlikely. For this use, you need to know

frequency of use of symbols in each of the alphabets and group the redesignated symbols in an appropriate way, which for the 16th century. seems not very realistic, especially considering that the regression analysis needed for this was invented much later. For the same reason, it should be assumed that the text of the manuscript is meaningful, otherwise the deviation from the statistics of letters specific to the natural lexicon would be much greater.

Thus, in this section, we accept the following working hypotheses regarding MV:

1. The manuscript is a bilingual text with a common alphabet.
2. Vowels were removed from the text before recoding.
3. Recoding consisted in the unambiguous replacement of a letter with a symbol.
4. Spaces in text are not considered characters.

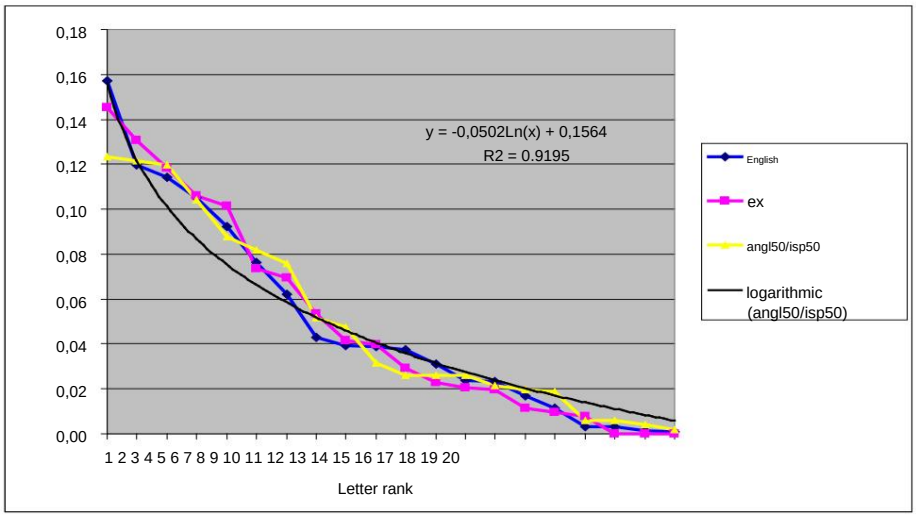
Then it is necessary to find out which pairs of languages with a common alphabet and in what proportion could be considered as languages of the Manuscript, whether they are from the same language group or from different ones and which ones, and also how strongly their thematic orientation affects the statistical properties of texts. With regard to texts in Russian, the influence of the genre on the alphabetical (but not on the ordered by frequency) distribution was considered in [9], where a certain dependence was marked.

We present here the results of a statistical analysis of frequencies in modern texts written in two languages but in the same alphabet. Let us first consider texts from one language group.

It turned out that both “pure” and 50/50 mixed texts in Russian and Bulgarian have close distributions with the same determination equal to 0.96, and the deviation of the actual distribution from the model one is 0.10. This mixture obviously has different statistical properties than MB in either of the two transcriptions.

Also for the distributions of characters for other texts - English German, Franco-Italian and in general texts in the languages of one group or subgroup - the determination of the logarithmic approximation of the mixture approximately coincides with the determination of texts in one language and is the same 0.96.

Let us now consider examples of mixing texts from different language groups of the Indo-European family. On fig. 6 shows the frequency distributions in Spanish-English texts without vowels. Note that mixing English and Spanish texts in equal proportions leads to determination at the level of 0.92 with a deviation of 0.17 of the approximation in the L1 norm from the actual distribution. In terms of statistics, this mixing is similar to the Takahashi transcription, the distance between the two distributions was 0.09.



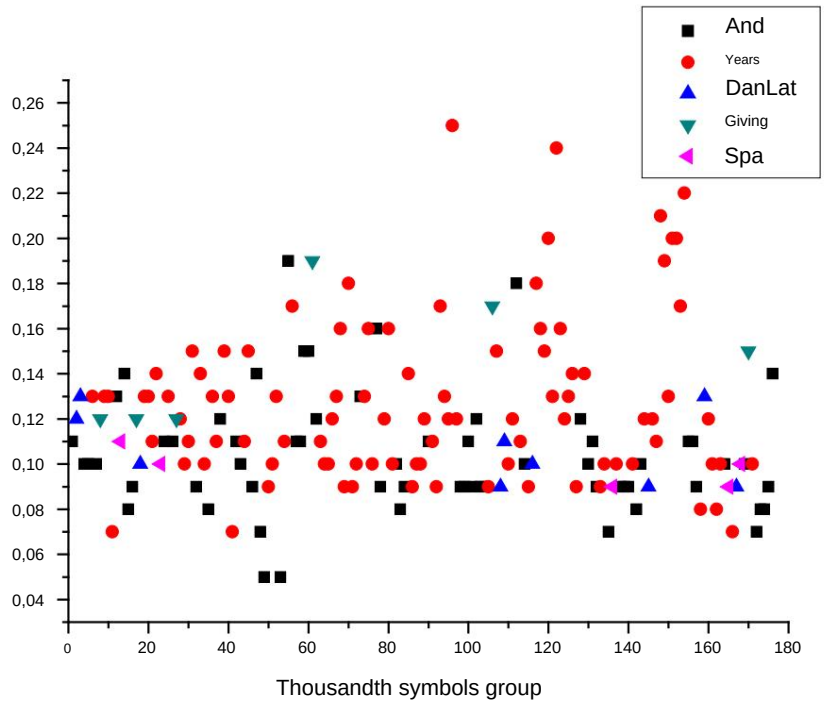
Rice. 6. Approximation of Spanish-English texts

It is characteristic that at a distance of 0.08-0.10 there are all distributions corresponding to different Spanish-English texts, taken in a 50/50 ratio. Therefore, it makes sense to look for the proportion between the volumes of texts in these languages, at which the determination will increase to 0.93. This proportion is found: approximately 60% of the English text and 40% of the Spanish correspond to it. The distance between the distributions of this mixture and the Takahashi transcription in normal L1 was 0.08. Thus, the statistical hypothesis about such a linguistic composition of the MW text can be considered quite acceptable.

The constructed reference distributions of symbols in texts on the definition divided languages can be used to answer the question where the text of the manuscript uses predominantly one language (for example, Spanish), and where it is mixed. To do this, it is necessary to apply the method of identifying sample distribution functions for small samples. The essence of the method is as follows. Let there be reference distribution functions (patterns) *F (x)* and some fragment of the time series, the sample distribution function of which is *G(x).* Then this fragment is considered to be a sample from the distribution *F (x)* with number

$$j \;\bar{a}\text{rgmin}\, \| F_j(x) - G(x) \| \qquad (3)$$

The results of the analysis of samples of length 1000 characters in the Takahashi transcription led to the following results (Fig. 7).



Rice. 7. Identification of the language of MV fragments

Periodically, the language of the text turns out to be close to one or another language of the groups under consideration.

5. Use of spectral portraits of bigram matrices

Consider the matrix *Pij of* empirical conditional probabilities

that in some place of the text there is a character *j ,* provided that on the left is . This matrix is one-letter *f (i)* of the probability distributionexpressed in terms of the two-letter symbol *i F(i,* j) and the

$$P_{ij} = \frac{F(fiji)}{(\,)} \quad (\,)\ fi\ F\ \ddot{y}j\ (\ ,\ ) \qquad (4)$$

It follows from (4) that the matrix *Pij* has one of its eigenvalues equal to 1, and the eigenvector *f* (i) corresponds to this value. Other eigenvalues of this matrix characterize the frequency stability of letter pairs for text fragments. According to S.K. Godunov [10], the number ÿ belongs to the ÿ -spectrum *(P)* ÿÿ of the matrix *P* such that ÿ ÿ ÿ *P* and det(ÿI ÿ *P* ÿ ÿ) = 0.

if there is such an outrage

matrix ÿ

When studying the location of the points of the spectrum, they represent the res closed smooth curves ÿ representing the isolines of the ÿ spectrum. into two parts, lying inside and outside it. The contour parameter *(P)* ÿÿ is estimated by the norm of the square of resolvent (9) on the given curve:

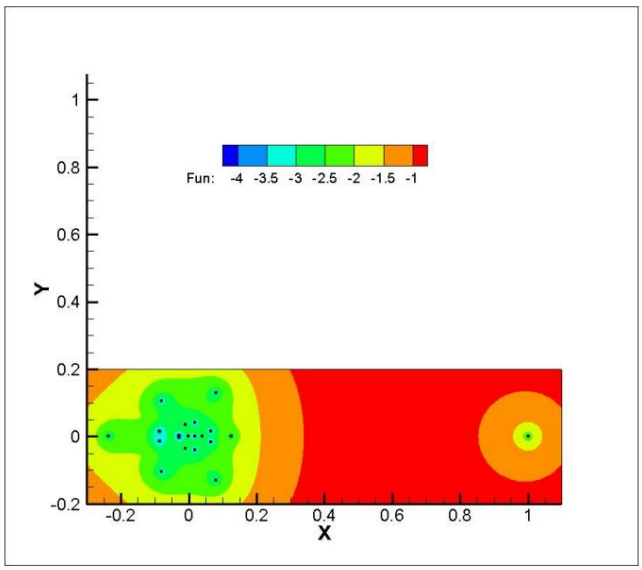
$$\ddot{y}\ddot{y} \pm (P) \quad \frac{\|P\|^2}{l_c} \ddot{y} \|R\ddot{y}(\ddot{y})d\ddot{y}\|^2 \qquad (5)$$

the spectrum is the length of the contour. The value *(P)* ÿÿ is chosen as an indicator. Here,

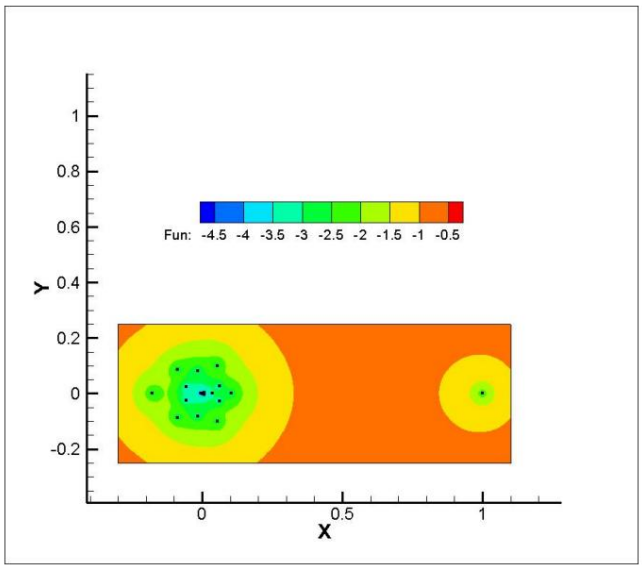
there are no points of the spectrum *ÿ(P),* then the norm of the resolvent on such a curve is finite: $\|R(\ddot{y})\ddot{y}\| < \ddot{y}$ as well as the integral of it over this curve.

If there are several eigenvalues ÿ , inside the region bounded by the curve, then in addition to specified accuracy ÿ .

It is of interest to compare the spectral portraits of matrices (4) for two MV transcriptions, as well as for the texts of the Germanic and Romance groups without vowels. The calculation results are shown in Figs. 8-113. The regions containing the eigenvalues of matrices are shaded with the same color if the elements of these matrices are known with the accuracy specified in the legend.

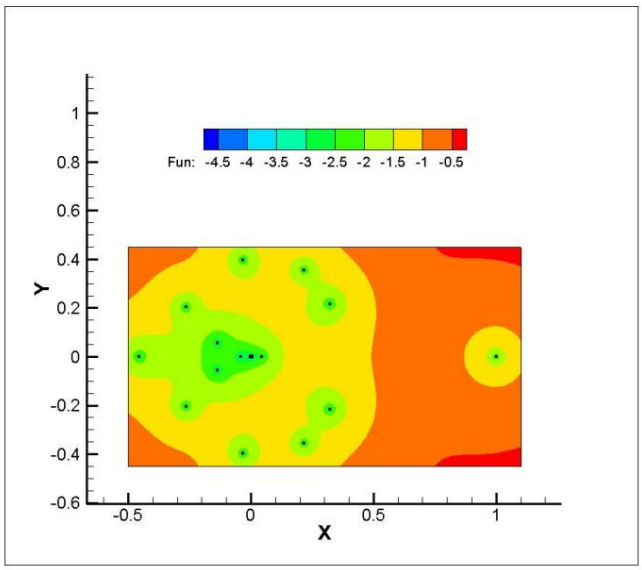


Rice. 8. Spectral portrait of the text without vowels in English

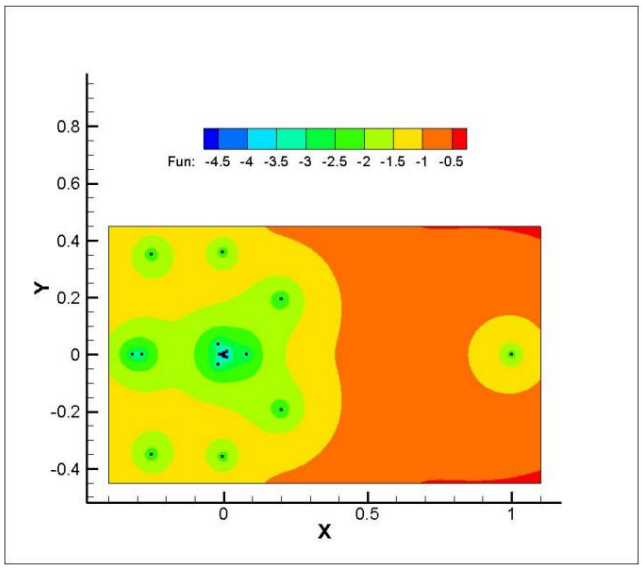


Rice. 9. Spectral portrait of the text without vowels in Latin

All matrices of the form (4) have one isolated eigenvalue equal to one. The remaining eigenvalues form a structure characteristic of a given language. Of interest are the real eigenvalues, the kernel near zero, and complex eigenvalues with large absolute values. For all European languages, the spectrum area is approximately limited to a circle with a radius of 0.2 (green area in Fig. 10-11). As it was found out in [5], for texts in the full alphabet, the region of the spectrum is not a circle, but an ellipse, the major semiaxis of which is approximately 0.5, and the minor one is still equal to 0.2.



Rice. 10. Spectral portrait of EVA transcription



Rice. 11. Spectral portrait of Takahashi transcription

Comparing Fig. 8-9 and fig. 10-11, we see that the areas of equal accuracy in finding the eigenvalues of matrices (4) for MV and ordinary texts (both in the full alphabet and without vowels) differ markedly. Of fundamental importance is the fact that for both MV transcriptions the circle (not an ellipse!) of the location of eigenvalues has a radius approximately twice as large as for natural languages. At the same time, the spectrum of EVA

shifted to the left, and the Takahashi spectrum to the right. The difference in the spectral portraits of transcriptions corresponds to the differences in the distributions of ordered frequencies, for which the convexities of these curves change in antiphase. Characteristically, both transcriptions have five disconnected spectral bands of equal accuracy of 10–2 (light green color in Fig. 11).

The fact that the eigenvalues of MV transcriptions lie in a circle rather than an ellipse distinguishes texts without vowels. The twice larger radius of this circle indicates that the possible neighborhoods of pairs of symbols are more variable than for one language. Thus, the results obtained in this section do not contradict the proposed concept of the MV compound language and supplement it with one more statistical argument. It seems important to emphasize that all these arguments are fundamentally different; express the features of independent statistics indicating that the interpretation of the MV as a composite manuscript is quite admissible.

with respect

The work was supported by the Russian Foundation for Basic Research (project no. 19-01-00602).

Literature

1. *Shailor B.A.* Voynich catalog record. Yale University Beinecke Rare Book & Manuscript Library.

2. *Pelling N.J.* The curse of the Voynich: the secret history of the world's most mysterious manuscript. – Surbiton, Surrey: Compelling Press, 2006. – 230 p.
3. *Barabe J.G.* Materials analysis of the Voynich Manuscript. Yale University Beinecke Rare Book & Manuscript Library.
4. *Levitov L.* Solution of the Voynich Manuscript: A liturgical Manual for the Endura Rite of the Cathari Heresy, the Cult of Isis. – Walnut Creek, Califor nia: Aegean Park Press, 1987. – 182 p.
5. *Orlov Yu.N., Osminin K.P.* Methods of statistical analysis of literary texts. - M .: Editorial URSS / Book House "LIBROKOM", 2012. - 326 p.

6. *Landini G., Zandbergen R.* A well-kept secret of mediaeval science: The Voynich manuscript // Aesculapius. 1998. V.18, p.77-82.
7. Transcription Takahashi. http://voynich.no-ip.com/folios/
8. *Huseyn-Zade S.M.* On the distribution of the letters of the Russian language by frequency of occurrence // Problems of information transmission, 1988. V.24, issue 4, p.102.
9. *Orlov Yu.N., Osminin K.P.* Determining the genre and author of a literary work by statistical methods // Applied Informatics, 2010. V. 26, no. 2, p. 95-108.

10. *Godunov S.K.* Modern aspects of linear algebra. - Novosibirsk: Scientific book, 1997. - 388 p.