Recognition of english and russian-language texts based on frequency characteristics

Y Kotov¹ and O Sanina²
¹Department of Information Security, Novosibirsk State Technical University, Novosibirsk, Russia
²Department of Computer Engineering, Novosibirsk State Technical University, Novosibirsk, Russia
E-mail:kotov@corp.nstu.ru

Abstract. Distinction of texts in one language from texts in others is necessary to solve the problems of automated text analysis. The paper presents criteria and critical values for recognizing English-language and Russian-language texts. The obtained criteria are estimated by experiments. The paper describes the methods to estimate the size of character codes and to identify a space character in a text. The algorithm for recognizing texts in the English and Russian languages with arbitrary encoding is studied and its accuracy is estimated experimentally.

1. Introduction

Language recognition is required to solve the problems of multi-coding communication, steganography, cryptanalysis and other problems of formal text analysis [1–4]. Nowadays, to solve the problem of language recognition, neural networks and Naive Bayes Classifier [5–7] are used. However, if the text encoding is unknown, the mentioned methods cannot be applied.

The problem solution with an arbitrary encoding being used assumes comparing certain characteristics of the text under investigation with sample characteristics of texts in a known language. In general cases it means that an arbitrary character sequence must be studied to answer the question if the sequence is a text in a given language (languages) or otherwise not a text at all. Such definition of language recognition differs from the known [5–7] cases and requires forming a new approach to solve the problem considered in the paper.

2. Problem definition and solution steps

In the paper we assume that each letter is represented by $k$ characters ($k \geq 1$) and consider texts in the Russian and English languages. In this regard, to recognize a text in an arbitrary character sequence it is necessary to tackle the following tasks:

1. To estimate the size of a character code $k$.
2. To make a primary distinction between English-language and Russian-language texts.
3. To substitute characters if $k > 1$.
4. To identify spaces in the texts.
5. To compare frequency characteristics of the texts with the language samples.
6. To identify letters in the texts.
7. To substitute characters of the texts for language letters.
3. The size of character codes and primary text distinction

If the size of character codes is known to be not greater than 2 ($k \leq 2$), the conjunction index can be used for estimating the size of character codes. The conjunction index is calculated based on the digram characteristics of a text [8]. If the value of the conjunction index tends to zero it can be concluded that combinations of two characters represent the letters of the alphabet; if it tends to one then one character stands for one letter. The critical value of the conjunction index for English-language texts is 0.816, and for Russian-language texts the index is equal to 0.821 [8]. If the current value is smaller than the critical, then two characters represent each letter in the text, otherwise one character represents one letter.

In the general case to estimate the size of a character code $k$, the alphabet cardinality of letters in a given language can be used as a criterion [9]. Measured for each point of the scale that represents the text size $x$, the minimal and maximal amount of the alphabet letters being used can be taken as critical values. Figure 1 shows the ranges of the letters’ number being used in the English and Russian languages at each scale point.

![Figure 1. Ranges of the alphabet cardinality in Russian-language and English-language texts.](image)

In this regard, the size of character codes can be estimated as follows:
1. To fragmentize a text – i.e. to represent a text as sequences of characters with the size $k$;
2. To compare the alphabet cardinality of the obtained code fragments with the sample number of the alphabet letters used in natural language texts (Figure 1).

If the alphabet cardinality of the obtained code fragments falls within the range presented in Figure 1, the code size is considered as accepted. There may be cases when several possible code sizes for one text fall within the range of the accepted number of letters. In particular, if the encoding sequences consist of the alphabet letters in the language under consideration or of single texts in a natural language, and the alphabet cardinality meets the criterion shown in Figure 1. Brute force can be employed to find all possible values for the size of the character code.

As a result, after all the possible values $k$ from one up to $x/200$ being brute-forced, the vector with accepted values of character codes $k$ is formed. In addition, if the cardinality of character codes fits the language, a flag $L$ must be assigned to indicate this language. The flag $L$ can have the following values:

- **0** – the cardinality of character codes does not fit any language (and the current value of $k$ is rejected);
1 – the cardinality of character codes fits the range of the Russian language;
2 – the cardinality of character codes fits the range of the English language;
3 – the cardinality of character codes fits the range both of the Russian and English languages (x < 1200).

If k > 1, the text is reencoded so that one letter of the language is presented by one character, and the analysis moves to the next step.

Stating with the greatest, the accepted values of the obtained vector are checked. The first size of a character code for which a text is identified as a text in a natural language is taken as the size of character codes.

At this step, English-language texts can be distinguished from Russian-language texts based on the cardinality of used alphabet letters. The unambiguous distinction between Russian-language and English-language texts based on this criterion is possible for texts with size x more or equal to 1200 characters. The distinction of texts with the size up to 1200 characters is possible only if the number of letters being used in the text does not fall within the intersection area of the languages.

However, the sole knowledge of the alphabet cardinality is not enough to make a final decision if the text belongs to any language. For that, frequency characteristics of a text should be compared with sample frequency characteristics of the chosen language. Prior to obtaining the frequency characteristics, the presence or absence of a space character must be confirmed, and a space character should be identified.

4. Identification of space characters in a text
A variety of criteria can be employed to identify a space character in a text [10]. The most accurate is the criterion based on the deviation of the expected value and the variance of the Poisson distribution, since the distribution of words from dictionary over the length is close to it [10]. The critical values for Russian-language texts are determined as (1):

\[ \delta(x) = \begin{cases} 
2.76 \cdot 10^{-4}x + 2.09, & 200 \leq x \leq 6000 \\
3.75, & x > 6000 
\end{cases} \]

For English-language texts, a constant value equal to 1.35 can be taken as a critical value for texts of any size (2):

\[ \delta(x) = 1.35 \]

If the index of words’ length is used as a criterion, the critical values for Russian-language texts are considered as (3):

\[ y(x) = \begin{cases} 
0.00087x + 9.206, & 200 < x < 1200 \\
10.25, & x \geq 1200 
\end{cases} \]

For English-language texts, a constant value equal to 23 [10] can be assumed as a critical value for texts of any size (4):

\[ y(x) = 23 \]

5. Recognition of English-language and Russian-language texts
The following frequency characteristics should be used to recognize texts in a natural language: the coincidence index and the number of bigrams being used in a text [8].

The range of English-language text’s size can be divided into 4 intervals, and in each interval, the following ranges for the critical values of the coincidence index can be accepted (with 95% of texts fall within the range) (5):
The accuracy of the step for identification of English-language texts without considering the cardinality of the alphabet being used has been estimated by the experiment with the validation sample of the Russian-language and English-language texts. The results are shown in Table 1, where \( N \) is the number of texts with the size \( x \), \( m \) stands for the number of errors in texts with the size \( x \), \( P_e \) is the error frequency, \( P_{\alpha \text{AV}} \) is the average error frequency in texts, \( \alpha \) – type I errors (English-language texts are rejected), \( \beta \) – type II errors (Russian-language texts are identified as texts in English).

The average error is 7.9%. The minimal error is zero at several scale points. The maximal error is 38% at the scale point \( x = 200 \).

The amount of error reduces as the text size is increasing: at the point \( x = 200 \) the average error achieves 38%, at the point \( x = 400 \) – 25%, the error is 16.3% at the interval (400; 800], at the interval (800; 2000] – 11.3%, and at the interval (2000; 350000] – 1%.

The accuracy of the step for identification of Russian-language texts without considering the cardinality of the alphabet being used has been estimated by the experiment with the same validation sample.
samples. Results are provided in table 2, where $\alpha$ – type I errors (Russian-language texts are rejected), $\beta$ – type II errors (English-language texts are identified as texts in Russian).

| $x$ | $N$ | $m$ | $P_E$ | $N$ | $m$ | $P_E$ | $P_{AV}$ | $x$ | $N$ | $m$ | $P_E$ | $N$ | $m$ | $P_E$ | $P_{AV}$ |
|-----|-----|-----|-------|-----|-----|-------|----------|-----|-----|-----|-------|-----|-----|-------|----------|
|     |     |     |       |     |     |       |          |     |     |     |       |     |     |       |          |
| 200 | 100 | 4   | 0.04  | 100 | 72  | 0.72  | 0.38     | 10000| 100 | 0   | 0    | 100 | 3   | 0.03  | 0.015   |
| 400 | 100 | 9   | 0.09  | 100 | 41  | 0.41  | 0.25     | 30000| 100 | 0   | 0    | 100 | 0   | 0     | 0        |
| 600 | 100 | 4   | 0.04  | 100 | 28  | 0.28  | 0.16     | 50000| 99  | 0   | 0    | 100 | 0   | 0     | 0        |
| 800 | 100 | 2   | 0.02  | 100 | 31  | 0.31  | 0.165    | 70000| 99  | 0   | 0    | 100 | 0   | 0     | 0        |
| 1000| 100 | 1   | 0.01  | 100 | 24  | 0.24  | 0.125    | 90000| 98  | 1   | 0.01| 100 | 0   | 0     | 0.005    |
| 1200| 100 | 3   | 0.03  | 100 | 23  | 0.23  | 0.13     | 110000| 98  | 2   | 0.02| 100 | 0   | 0     | 0.010    |
| 1400| 100 | 1   | 0.01  | 100 | 18  | 0.18  | 0.095    | Total 3| 594 | 3   | 0.005| 600 | 3   | 0.005 | 0.005    |
| 1600| 100 | 0   | 0     | 100 | 21  | 0.21  | 0.105    | Group 4|     |     |      |      |     |      |          |
| 1800| 100 | 4   | 0.04  | 100 | 17  | 0.17  | 0.105    |         |     |     |      |      |     |      |          |
| 10000| 100 | 0   | 0     | 100 | 4   | 0.04  | 0.02     | Total 2| 2270| 37  | 0.016| 2400| 340 | 0.142| 0.079   |

Table 2. The errors in identifying Russian-language texts.

| $x$ | $N$ | $m$ | $P_E$ | $N$ | $m$ | $P_E$ | $P_{AV}$ | $x$ | $N$ | $m$ | $P_E$ | $N$ | $m$ | $P_E$ | $P_{AV}$ |
|-----|-----|-----|-------|-----|-----|-------|----------|-----|-----|-----|-------|-----|-----|-------|----------|
|     |     |     |       |     |     |       |          |     |     |     |       |     |     |       |          |
| 200 | 100 | 11  | 0.1   | 100 | 61  | 0.61  | 0.355    | 10000| 100 | 0   | 0    | 100 | 0   | 0     | 0        |
| 400 | 100 | 13  | 0.13  | 100 | 66  | 0.66  | 0.395    | 30000| 100 | 0   | 0    | 100 | 0   | 0     | 0        |
| 600 | 100 | 8   | 0.08  | 100 | 78  | 0.78  | 0.43     | 50000| 100 | 0   | 0    | 100 | 0   | 0     | 0        |
| 800 | 100 | 1   | 0.01  | 100 | 76  | 0.76  | 0.385    | 70000| 100 | 0   | 0    | 100 | 0   | 0     | 0.005    |
| 1000| 100 | 2   | 0.02  | 100 | 62  | 0.62  | 0.32     | 90000| 100 | 0   | 0    | 100 | 0   | 0     | 0.005    |
| 1200| 100 | 0   | 0     | 100 | 51  | 0.51  | 0.255    | 110000| 100 | 0   | 0    | 100 | 0   | 0     | 0.005    |
| 1400| 100 | 1   | 0.01  | 100 | 39  | 0.39  | 0.2      | Total 3| 600 | 5   | 0.05| 594 | 0   | 0     | 0        |
| 1600| 100 | 1   | 0.01  | 100 | 52  | 0.52  | 0.265    | Group 4|     |     |      |      |     |      |          |
| 1800| 100 | 1   | 0.01  | 100 | 48  | 0.48  | 0.245    |         |     |     |      |      |     |      |          |
| 2000| 100 | 1   | 0.01  | 100 | 47  | 0.47  | 0.24     | Total 2| 2000| 50  | 0   | 0.04| 300 | 214 | 0.428| 0.217   |
|     |     |     |       |     |     |       |          |     |     |     |       |     |     |       |          |
| 2000| 100 | 1   | 0.01  | 100 | 52  | 0.52  | 0.265    | 25000| 50  | 1   | 0.02| 300 | 50  | 0.02  | 0.001    |
| 4000| 100 | 0   | 0     | 100 | 57  | 0.57  | 0.285    | 30000| 79  | 0   | 0    | 29  | 0   | 0     | 0        |
| 6000| 100 | 0   | 0     | 100 | 52  | 0.52  | 0.26     | 35000| 21  | 1   | 0.048| 27  | 0   | 0     | 0.024    |
| 8000| 100 | 0   | 0     | 100 | 51  | 0.51  | 0.255    | Total 4| 300 | 6   | 0.02| 176 | 0   | 0     | 0.010    |
| 10000| 100 | 2   | 0.02  | 100 | 2   | 0.02  | 0.02     | Total 4| 2400| 48  | 0.02| 2270| 794 | 0.350| 0.185   |

The average error is 18.5%. The minimal error is zero at several scale points. The maximal error is 43% at the scale point $x = 600$.

The amount of error reduces as the text size is increasing: at the interval [200; 1000] the error is 37.7%, at the interval (1000; 10000) – 22.9%, at the interval (10000; 350000) – 0.6%.

As the language of the text is identified, the text analysis can move to the next step, where the language letters are identified and substitution of character codes for letters of the language follows the identification. The techniques described, for instance, in [11] can be employed for letter identification.
6. Conclusion

The flowchart of the algorithm for recognizing English-language and Russian-language texts with unknown encoding is shown in Figure 2.

![Flowchart of the algorithm for language recognition](image)

**Figure 2.** The flowchart of the algorithm for language recognition.

The accuracy of the suggested algorithm has been estimated experimentally with the validation samples of Russian-language and English-language texts. The results are provided in table 3, where \( N \) is the number of texts with the size \( x \), \( P_E \) – the number of all errors (both type I and type II), \( P_{AV} \) is the average error frequency in texts.

According to the algorithm, the text is examined to check if it belongs to the English language and after that, if rejected, if it belongs to the Russian language. Examination of texts for Russian and then for English provides with worse results.

**Table 3.** The errors in language recognition.

| \( x \) | \( N \) | \( P_E \) | \( N \) | \( P_E \) | \( P_{AV} \) | \( x \) | \( N \) | \( P_E \) | \( N \) | \( P_E \) | \( P_{AV} \) |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| Group 1 | | | | | | | | | | | |
| 200 | 100 | 0.45 | 100 | 0.04 | 0.245 | | | | | | |
| 400 | 100 | 0.13 | 100 | 0.09 | 0.11 | 10000 | 100 | 0.03 | 100 | 0 | 0.015 |
| 600 | 100 | 0.08 | 100 | 0.04 | 0.06 | 30000 | 100 | 0 | 100 | 0 | 0 |
| 800 | 100 | 0.01 | 100 | 0.02 | 0.015 | 50000 | 100 | 0 | 99 | 0 | 0 |
| 1000 | 100 | 0.02 | 100 | 0.01 | 0.015 | 70000 | 100 | 0 | 99 | 0 | 0 |
| 1200 | 100 | 0 | 100 | 0.03 | 0.015 | 90000 | 100 | 0 | 98 | 0.010 | 0.005 |
| 1400 | 100 | 0.01 | 100 | 0.01 | 0.01 | 110000 | 100 | 0.01 | 98 | 0.020 | 0.015 |
| 1600 | 100 | 0.01 | 100 | 0 | 0.005 | Total 3 | 600 | 0.007 | 594 | 0.005 | 0.006 |
| 1800 | 100 | 0.01 | 100 | 0.04 | 0.025 | | | | | | |
| 2000 | 100 | 0.01 | 100 | 0.03 | 0.0 | 100000 | 100 | 0 | 30 | 0 | 0 |
| Total 1 | 1000 | 0.073 | 1000 | 0.031 | 0.052 | 150000 | 50 | 0.04 | 30 | 0.033 | 0.037 |
| Group 2 | | | | | | | | | | | |
| 2000 | 100 | 0.01 | 100 | 0.04 | 0.025 | 250000 | 50 | 0.02 | 30 | 0 | 0.01 |
| 4000 | 100 | 0.01 | 100 | 0 | 0.005 | 300000 | 79 | 0 | 29 | 0 | 0 |
| 6000 | 100 | 0 | 100 | 0 | 0 | 350000 | 21 | 0.048 | 27 | 0 | 0.024 |
| 8000 | 100 | 0 | 100 | 0 | 0 | Total 4 | 300 | 0.02 | 176 | 0.011 | 0.016 |
| 10000 | 100 | 0.02 | 100 | 0 | 0.01 | Total 2400 | 0.036 | 2270 | 0.018 | 0.027 |
The average error is 2.7%. The minimal error is zero at several scale points. The maximal average error is 24.5% at the point $x = 200$. Furthermore, the average error for Russian-language texts (3.6%) is found to be higher than for English-language texts (1.8%). The greatest number of errors is found in texts with small sizes (up to 600 characters). The average error at this interval is equal to 13.8%. As the text size is increasing, the distribution of frequency characteristics is getting stable and the amount of errors in language recognition reduces.

References
[1] Rubinstein-Salzedo S 2018 Cryptography (Luxembourg: Springer) p 259
[2] Banik B G and Bandyopadhyay S K 2018 IETE Journal of Research 66:3 384–395
[3] Bochkarev A I, Pepelyaeva M A and Skvortsova E B 2018 Proc. XIV International Scientific-Technical Conference on Actual Problems of Electronics Instrument Engineering 33–36
[4] Yang N and Mali A D 2016 Proc. 28th International Conference on Tools with Artificial Intelligence 165–168
[5] Zazo R, Lozano-Diez A, Gonzalez-Dominguez J, Toledano T D and Gonzalez-Rodriguez J 2016 PLOS ONE 11 1–17
[6] Simonchik K, Novoselov S and Lavrentyeva G 2016 Lecture Notes in Computer Science 9811 174–181
[7] Noraset T, Demeter D and Downey D 2018 Proc. of the AAAI Conference on Artificial Intelligence 5333–5341
[8] Kotov Yu and Sanina O 2018 Proc. XIV International Scientific-Technical Conference on Actual Problems of Electronics Instrument Engineering 175–179
[9] Kotov Yu and Sanina O 2019 Proc. VIII International Scientific Conference of Graduate and Postgraduate Students "Progress through Innovations" 69–71
[10] Kotov Yu and Sanina O 2020 Vestnik SibGUTI 60–72
[11] Alkazaz N R, Irvine S A and Teahan W J 2018 Information Security Journal 27 57–75