# Statistical Localization of Bibliographic Descriptions in Unstructured Full-Texts Documents

Leonid Grashenko[1][0000-0002-7972-1358], Alexander Modin[2][0000-0001-7356-055X], Nikita Kuzmin[3][0000-0001-7242-5778]

Russia FSO Academy, Orel, Russia
[1]graschenko@mail.ru, [2]modin.schura2011@outlook.com, [3]kuzyan15@gmail.com

**Abstract.** The article describes the results of experiments in the field of automatic localization of bibliographic descriptions (single and group as part of lists) drown up according to GOST 7.0.100-2018 (or close standards). The experiments were performed on the set of unstructured full-text Russian-language documents of various styles. The proposed solution is based on several parameters of bibliographic descriptions: lengths distribution (in characters), the frequency of prescribed punctuation characters and autocorrelation factors. The use of these features in an explicit form during simple classifiers creation made it possible to obtain criteria of Recall and $F_1$-scores, comparable to previously obtained one using structural recognition methods.

**Keywords:** Bibliographic description, Bibliographic data, Text mining, Named entity recognition, Statistical features, Natural language processing.

## 1 Introduction

One of the actual tasks in the Natural Language Processing (NLP) is the bibliographic information extraction, which is used for both identification of source documents and bibliographic search. This task is a special case of metadata extraction. Primarily, such functionality is used in automatic quality assessment systems of scientific and educational papers, their peer review, in applications to Bibliometrics and Plagiarism detection. Bibliographic information extraction from scientific books, theses and articles eventually helps in saving researcher's time while performing literature acquisition. In addition, «metadata information within citation also carries immense importance especially in the domain of Scientometrics» [1]. Bibliographic information can be used to perform variety of other tasks including article recommendation and citation analysis, etc.

This task also can be characterized as a special case of Text Mining and Entity extraction within the NLP. More precisely, we can talk about extracting objects represented by a set of named entities. The process of extracting bibliographic information consists of three stages: (i) localization of bibliographic data, (ii) definition of object

boundaries and its type, (iii) identification of fields, extraction and interpretation bibliographic data. This paper presents the results of experiments to determine informative features for the statistical localization of single and group (as part of a list) bibliographic descriptions in full-text unstructured documents in Russian.

## 2 Overview

The majority of the texts in Russian with the appearance of bibliographic references and descriptions is subject to the standards GOST 7.1-2003, GOST 7.0.5-2008 and GOST 7.0.100-2018. Primarily, we are talking about scientific, academic and educational publications, technical reports, patents, official documents, etc. Of course, other formats (styles) of presentation of bibliographic descriptions are also in use. However, the article considers the localization of bibliographic descriptions in the notation of the GOST standards. Predominantly we considered bibliographic descriptions of books and articles as the most traditional types of publications.

So, based on 10 thousand random bibliographic descriptions from the Russian State Library web-portal of the [9], it was found that 61% of them represented by books, 17% by articles, 15% by dissertations (theses), 5% by electronic documents on the optical disk storage and 2% by other sources.

According to the above standards, the source of bibliographic information is presented by bibliographic records, bibliographic descriptions (BD) and bibliographic references (citations). At the same time, the bibliographic description is the main part of the bibliographic record, and differs from the bibliographic reference (BR) in that it clearly specifies the sequence and formatting of fields (elements) containing bibliographic data. Also in relation to bibliographic description defined the term «prescribed punctuation» – a set of characters used to separate fields within the BD (see Fig. 1).

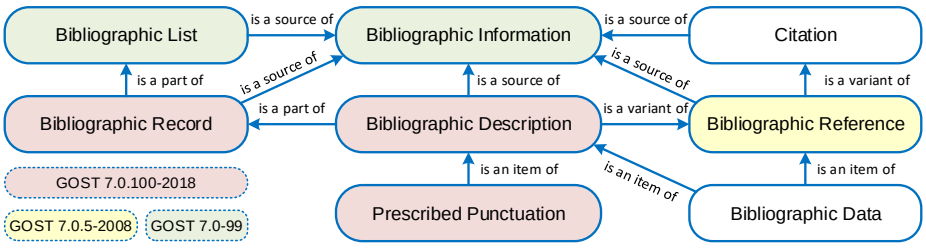


**Fig. 1.** The semantic network for the subject area

Thus, the extraction of bibliographic information from texts involves the identification of such objects as bibliographic records, bibliographic descriptions and bibliographic references and the interpretation of their constituent fields containing bibliographic data. Taking into account that BD may be considered as an extended version of BR, attention should be focused on identifying signs of localization of this object. It should also be noted that in full-text documents bibliographic descriptions may occur in various presentation forms, according to the classification (see Fig. 2). The classification

items on the diagram are ordered in such a way that the complexity of identifying such BDs increases from right to left. Wherein, the greatest difficulty is the automatic recognition of short single-level single BDs, which format does not match any known styles. Accordingly, it is easiest to identify multilevel full group (two or more in a row) bibliographic descriptions drawn up according to GOST.

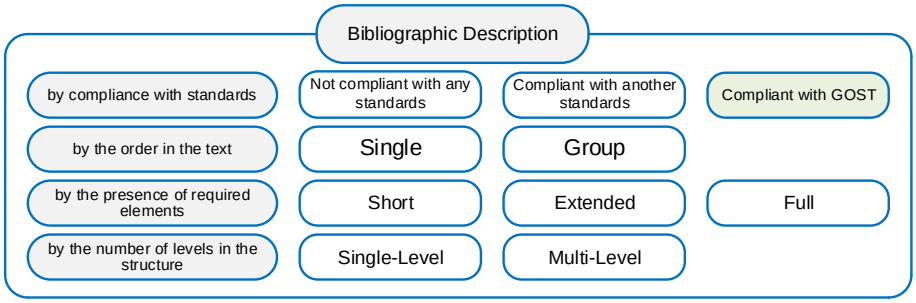


**Fig. 2.** The classification of the bibliographic descriptions

Taking into account the above classification, the following aspects should be listed that affect the solution of the of the problem of identifying BD in the text:

1. Bibliographic descriptions may be located both in groups (in lists) in certain places, and by one in arbitrary places within the text.
2. Generally, bibliographic descriptions can be multilingual information objects.
3. Texts of documents can be structured and unstructured [2]. Structured texts contain special markup (tags) or dedicated fields for bibliographic descriptions. In unstructured texts, bibliographic descriptions are not especially distinguished in any way.
4. Bibliographic descriptions may be incomplete (one or several fields are missing) and corrupted (the necessary fields or markup elements, as well as their sequence are distorted). In addition, sometimes references to literature contain erroneous data, for example, incorrect page numbers, publisher name, etc. First and foremost, the source and cause of distortions is a human.

## 3 Previous and Related Works

A review of the available publications showed that the following approaches to extracting bibliographic information from texts of various documents in Russian were implemented in practice:

- «Antiplagiat» system [3] includes a module for extracting bibliographic information. However, unlike other modules, there is no publicly available description of its algorithm.
- The work [4] presents a method for extracting bibliographic information from full-text patent descriptions based on a library of patterns (regular expressions). After searching in the text for all templates, the procedure of integrating the text fragments

selected by each template is performed. The authors point to the achieved accuracy of selection of bibliographic references – 88% (another 9% is allocated partially).

- In work [5], an approach to the identification of bibliographic descriptions in the bibliography based on the modified shingle algorithm is implemented. The low speed of the developed system and the dependence of accuracy on the composition (completeness) of the required bibliographic database are noted.
- The article [6] presents a toolkit for extracting elements of bibliographic lists from scientific texts based on automatic generation of regular expressions. The author points out that template generation was successful in 76% of cases (100 bibliographic references to electronic resources were used).
- Finally, work [7] discusses the extraction of bibliographic descriptions using regular expressions. They are used to determine the correspondence of a block of text to the characteristics of a named entity in accordance with certain templates, which include special characters. The authors point out that for correct and accurate extraction of bibliographic information, tens to hundreds of regular expressions are required, which significantly slows down the text processing. In their opinion, it is advisable to combine the regular expressions with statistical recognition methods to localize text fragments where bibliographic information may be contained.

For metadata extraction, various researchers have considered approaches based on Hidden Markov Models (HMM), Conditional Random Fields (CRF), Support Vector Machine (SVM), decision trees and neural networks [1, 2, 7]. Some researchers use heuristic approaches; others prefer Machine Learning methods. There are studies that combine the advantages of each of these approaches. However, in the vast majority of cases, practical systems first parse the document to establish its structure, and then proceed to extract metadata. In [8] it is indicated that various practical systems demonstrate the recall of the bibliographic information extraction within 57–90%, and the F1-score within 67–93%. In this case, used classifiers based on tens or hundreds of features. Machine learning models are widely used due to their ability to adapt to different structures and text styles.

Note that in [7] the authors define BD as a sequence of named entities. Indeed, some standardized fields of bibliographic descriptions contain indications of persons, geographical names; others are represented only by numeric values (see Fig. 3).

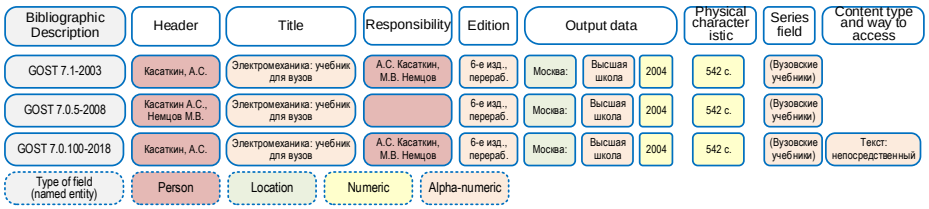


**Fig. 3.** Representing a bibliographic description/reference as a set of named entities

Recognition and selection of named entities is a rather trivial task for structured text, primarily because at its creation certain conventions and patterns are used. However, in general we are dealing with poorly structured texts. Therefore, direct recognition of BD

by structural methods may be preceded by the localization of places in texts where the presence of BD is most likely. For this, it is reasonable to use a statistical classifier based on a compact set of simple features. Therefore, this paper investigates the main statistical properties of bibliographic descriptions. In contrast to the previously presented data [9], here the research is carried out on a more representative information base.

## 4 Computational Models for Localization of the Bibliographic Descriptions

Let us consider universal alphabet Ω – the set of all possible characters that can occur in natural language texts, including the empty character ε. In practice, this can be a Unicode character set. Also given a set of prescribed punctuation characters (PPC) Σ of size |Σ|. In addition, over the alphabet Ω, a string of characters A = a0a1 ... an-1 of length n = |A| is defined.

The statistical features of bibliographic descriptions considered below are based on the following set of statements. BD in the general case is a text fragment with limited length. This fragment is composed by arbitrary sequence of the alphanumeric substrings (fields of bibliographic data), separated by characters from the set Σ. In addition to the usual grammatical punctuation marks, there are PPCs between the BD fields. Therefore, on the length of the BD there should be an increased occurrence of prescribed punctuation characters relative to the text as a whole. Then the localization of bibliographic descriptions can be performed using the sliding window. And the first localization model is specified by three parameters: $M_1 = <S, O, T_{PPC}>$, where $S$ is the window size, $O$ is the offset step for window sliding relative to the previous one, $T_{PPC}$ is the threshold of the relative frequency of the PPC set Σ. It is applicable to Bayesian classification scheme.

Further, because BDs are structured entities, they may demonstrate some cyclic or periodic properties with respect to the positions of prescribed punctuation characters. Therefore, there may be increased values of the corresponding indicators in the areas of BD localization in the text. As such an indicator, consider the autocorrelation.

Let φ: Ω → {ε, 0, 1} be a homomorphism where for character $x$ from Ω

$$\varphi(x) = \begin{cases} \varepsilon, if\ x = \varepsilon, \\ 1, if\ x \in \Sigma, \\ 0, otherwise. \end{cases} \quad (1)$$

The autocorrelation (periodic) of a string $A$ length $n$ is defined by

$$c_k(\mathrm{A}) = \sum_{j=0}^{n-1} \varphi(a_j)\varphi(a_{j+k\ mod\ n}), k \in \mathbb{N}, 0 \le k \le \left\lfloor \frac{n}{2} \right\rfloor. \quad (2)$$

Averaged autocorrelation coefficient (AACC) is defined by

$$AACC_K = \frac{1}{K}\sum_{i=0}^{K} \mathrm{c}_i(A), 0 \le K \le \left\lfloor \frac{n}{2} \right\rfloor. \quad (3)$$

Accordingly, the second localization model is also specified by three parameters M2 = <S, O, TAC>, where S is the window size, O is the offset step for window sliding relative to the previous one, TAC is the threshold value of AACC. It is also applicable to the Bayesian classification scheme.

# 5 Experiments and Evaluation

## 5.1 Datasets Overview

The following datasets were used for the experiments. A sample of about 350,000 bibliographic descriptions was obtained from marked-up lists of references to texts posted on the websites of the Russian State Library [10], the National Electronic Library [11] and the Scientific Electronic Library [12]. The original texts were divided into three categories: books (monographs), articles and theses (with auto-abstracts) in the following percentage: 50%, 28%, and 22%. Bibliographic descriptions were extracted with the use of Selenium, PhantomJS and Scrapy libraries in Python.

A sample of full-text documents was taken from the Russian National Corpus posted on the Internet and marked up by experts [13]. The sample size is 2000 texts of five functional styles (scientific, official-business, journalism, spoken, and literary). The total length of the sample texts is more than 20 million characters.

## 5.2 Parameter Estimation

The analytical software package Statistica 10 was used to determine the parameters of the distributions. Based on the Kolmogorov-Smirnov test with a significance level of 0.95, it was found that a lognormal law with various parameters for articles, books and theses describes the BD lengths distribution (see Fig. 4).

As shown, in articles and books, short BDs are usually used (mean – 220 and 244 characters, respectively), while in theses, extended and full BDs are often used (mean – 437 characters). Due to this, the mathematical expectation and variance of BD lengths are noticeably higher for dissertations. For convenience and optimization of further calculations, the width of the BD localization window is taken as a value near the average lengths for books and articles – 256 (as $2^8$).

Next, the assumption that the frequencies of occurrence of prescribed punctuation characters (according to GOST 7.0.1-2018) differ within the boundaries of bibliographic descriptions and in the text as a whole was checked.

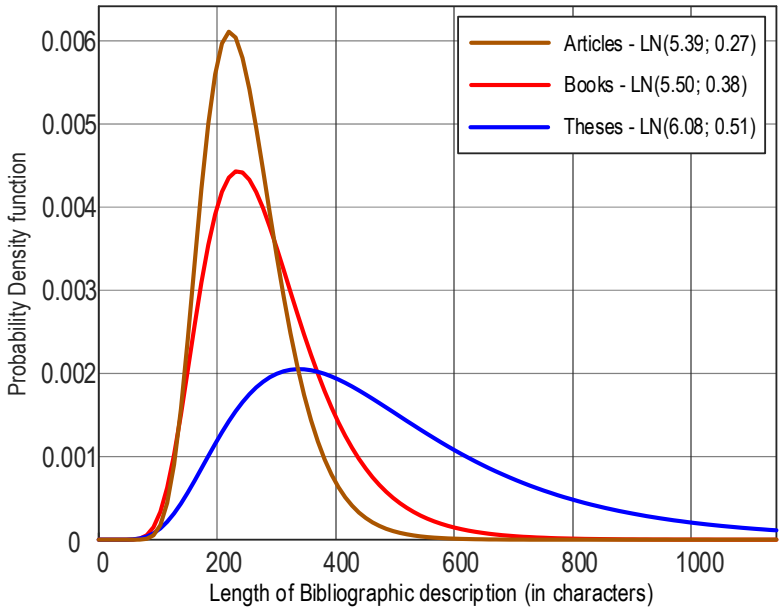


**Fig. 4.** The distributions of the bibliographic descriptions lengths

The following Table 1 summarizes data on occurrence frequencies of each prescribed punctuation character (PPC) in texts and within bibliographic descriptions.

**Table 1.** The relative occurrence frequencies of prescribed punctuation marks in the texts and in the bibliographic descriptions (the sum over column is 1.0)

| Character | Unicode# (Hex) | Plain text | BD |
|---|---|---|---|
| . | 0x002E | 0.24 | 0.32 |
| , | 0x002C | 0.35 | 0.38 |
| : | 0x003A | 0.07 | 0.05 |
| ; | 0x003B | 0.08 | 0.04 |
| / | 0x002F | 0.02 | 0.10 |
| ( | 0x0028 | 0.07 | 0.03 |
| [ | 0x005B | 0.06 | 0.03 |
| + | 0x002B | 0.03 | 0.00 |
| - | 0x002D | 0.01 | 0.05 |
| = | 0x003D | 0.04 | 0.00 |
| … | 0x2026 | 0.03 | 0.00 |

Since the use of PPC in texts and BD differs (for some characters significantly), this allows the use of relative frequencies of such characters as predictors in the logistic regression model for classifying text blocks. From the initial dataset, using a sliding window with a size of 256 characters and with an offset of 128 characters, blocks of two types were selected: those containing no BDs (text blocks) and containing BDs

(bibliographic blocks). Then the parameters of the frequency distributions of the full set of PPCs were determined (see Fig. 5).

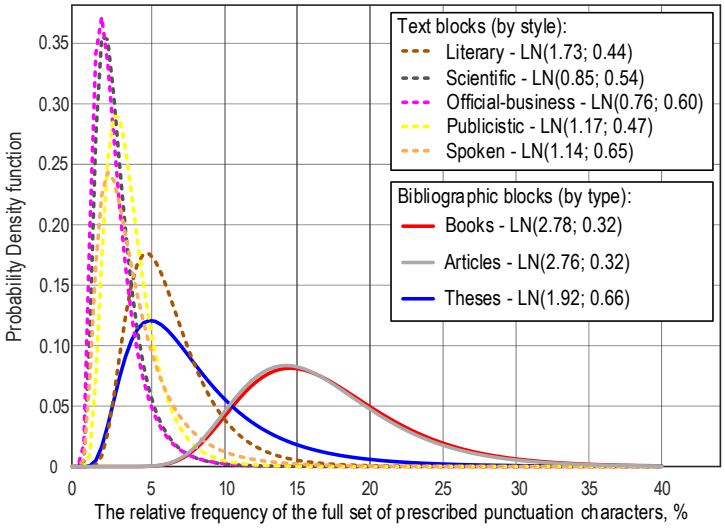


**Fig. 5.** Probability distributions of the relative occurrence frequencies of the complete set of prescribed punctuation characters in the text and bibliographic blocks (length 256)

As a result of averaging the data over blocks of texts of all styles, the threshold of the relative frequency of the full set of PPCs in various bibliographic descriptions was selected TPPC = 0.045. In addition, the probabilities of Type I errors (α, «false positive») and Type II errors (β, «false negative») are 0.0122 and 0.0054 respectively (see Fig. 6).

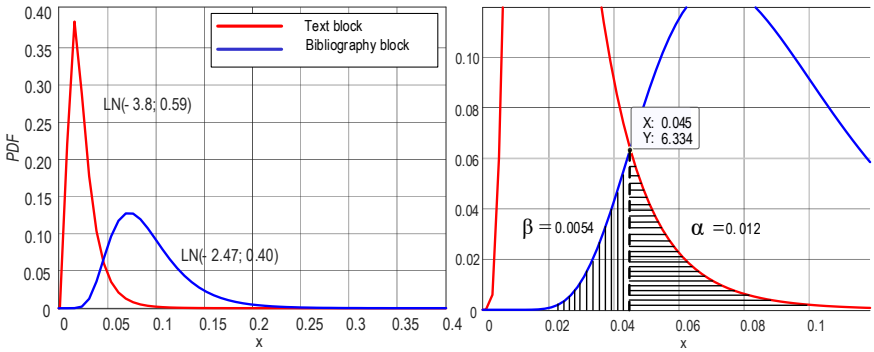


**Fig. 6.** Parameters estimation for Naive Bayes classifier for text and bibliographic blocks

Evaluation of the first classifier with parameters M1 = <256, 128, 0.045> were carried out on samples of 70,000 text blocks of various styles and 150,000 bibliographic blocks obtained by pulling a sliding window. It was found that a Naive Bayesian classifier at the specified threshold, although it allows identifying nearly all of BD locations, but

also captures some syntactic constructs. These are addresses, listings of surname-name groups, tabular data, mathematical and formal expressions, legislative acts.

### 5.3 Autocorrelation Features

Similarly, the parameters for the Bayesian model of recognition of bibliographic blocks were determined based on the value of the average autocorrelation coefficient (AACC) calculated on a string with a length of 256 characters (see Fig. 7).

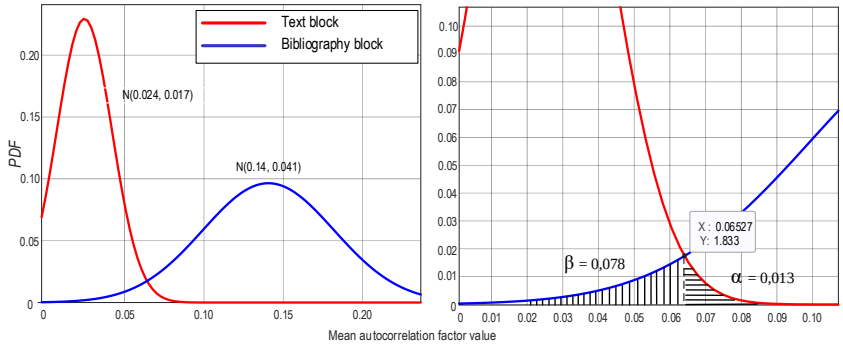


**Fig. 7.** Parameters estimation for Naive Bayes classifier for average ACF coefficient

Since the calculation of autocorrelation is a computationally expensive process, we optimized such a parameter as the maximum index of the autocorrelation coefficient used for the calculation. It is founded that the minimum value of the classifier errors sum is reached with the shift $K = 30$ (see Fig. 8).

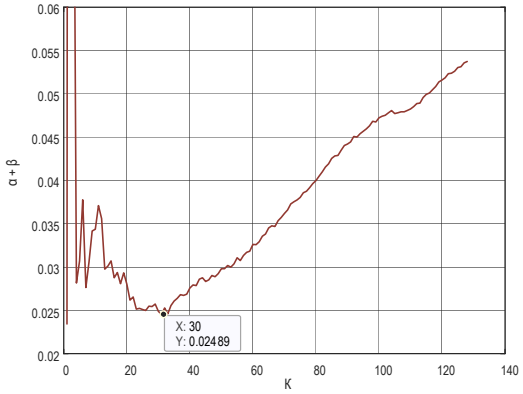


**Fig. 8.** Minimization of the sum of errors of the Naive Bayesian classifier by the maximum shift for calculating the average autocorrelation factor

In general, the proposed method for taking into account autocorrelation properties for localizing bibliographic references, with some refinement, can be used to visualize the structure of documents (see Fig. 9).

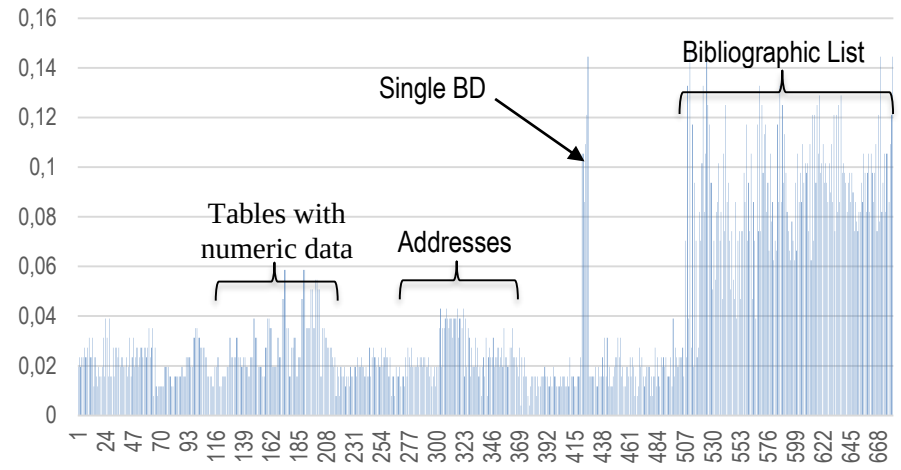


**Fig. 9.** An example of the AACC values for a text containing a bibliography

### 5.4 Spectral Features

In addition, after applying the Fast Fourier transform (FFT) to the autocorrelation sequences sized 128 symbols was investigated the form of the power spectrum. The experiment was carried out on a sample of 45000 blocks: 15000 plain-text blocks, 15000 blocks with mathematical expressions selected from textbooks of physics, chemistry and mathematics and 15000 bibliographic blocks. Based on the results of averaging the spectra, an integral result was obtained that demonstrates significant differences in the spectral characteristics between the indicated types of content (see Fig. 10).

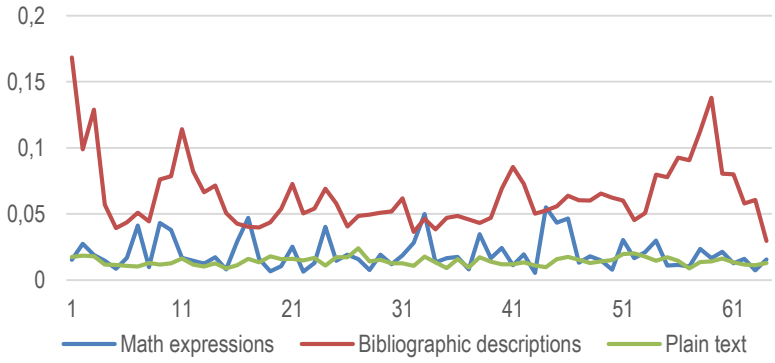


**Fig. 10.** Average FFT-spectrum of ACF for three types of text block

### 5.5 Preliminary Results

Based on the considered statistical features of bibliographic descriptions in the first approximation, an attempt was made to use them as features (in their original form,

without any transformations) when constructing simple classifiers separating text fragments on the width of a sliding window into two classes - text blocks and bibliographic blocks. Thus, it becomes possible to localize bibliographic descriptions in an unstructured full-text document. The first two classifiers are built on the basis of a naive Bayesian approach – the decision is made based on comparison with the threshold of a single feature value.

In addition, two binary logistic classifiers (BLC) were built with a decision threshold equal to zero (two classes are given: -1 for text block, +1 for bibliographic block). The first classifier based on the normalized relative frequencies of the six most commonly used prescribed punctuation characters: comma ($x_1$), dot ($x_2$), colon ($x_3$), semicolon ($x_4$), hyphen ($x_5$) and forward slash ($x_6$). All six factors are significant.

$$F_1(x) = -0{,}94 - 9{,}40 \cdot x_1 + 18{,}14 \cdot x_2 + 22{,}48 \cdot x_3 + 9{,}55 \cdot x_4 + 3{,}58 \cdot x_5 + 62{,}94 \cdot x_6. \quad (4)$$

The second BLC-classifier is based on same set of features, but the seventh feature is added – the normalized value of the averaged correlation coefficient AACC30 (x7). All seven factors are significant.

$$F_2(x) = -1{,}06 - 9{,}67 \cdot x_1 + 15{,}49 \cdot x_2 + 19{,}53 \cdot x_3 + 9{,}84 \cdot x_4 + 2{,}25 \cdot x_5 + 47{,}02 \cdot x_6 + 4{,}03 \cdot x_7. \quad (5)$$

For experimental verification of the presented solutions, in addition to the described set of initial data, with the involvement of volunteers, 500 full-text documents were annotated – one hundred texts of each style. Each document contained a list of references from 80–150 bibliographic descriptions, and 121 documents contained a single BD, 230 contained paired BDs, and 149 contained single and paired BDs in different parts of the text.

To compare and evaluate different classifiers major evaluation metrics include Precision, Recall and $F_1$-score. The test results of classifiers are summarized in Table 2.

**Table 2.** Precision, recall, $F_1$-score and area under curve (AUC) values for bibliographic descriptions localization models

| Model (classifier) | Recall | Precision | $F_1$-score | AUC-ROC |
|---|---|---|---|---|
| Naive Bayesian ($M_1$ = <256, 128, 0.045>) | 0.821 | 0.647 | 0.734 | 0.821 |
| Naive Bayesian ($M_2$ = <256, 128, 0.008>) | 0.830 | 0.701 | 0.760 | 0.848 |
| Binary Logistic Regression ($F_1$) | 0.839 | 0.684 | 0.754 | 0.844 |
| Binary Logistic Regression ($F_2$) | 0.692 | 0.791 | 0.738 | 0.807 |

It should be noted that the obtained values of the AUC-ROC indicator can be described as good, which indicates the prospects of the statistical indicators considered as classification signs for solving the problem of extracting bibliographic information from full-text documents. The obtained values of Recall, Precision and F-score are comparable with the results achieved to the results obtained in the above studies using structural recognition methods. At the same time, a compact model of BD localization on simply calculated features is implemented.

To implement stages of the presented study specialized software was developed. It provides the ability to vary various parameters of the models (width of the sliding window, parameters for calculating autocorrelation, symbol sets, etc.) and iterate over them in the specified ranges. In this case, the user can visually view the results of BDs localization – text fragments selected from the input documents (see Fig. 11).

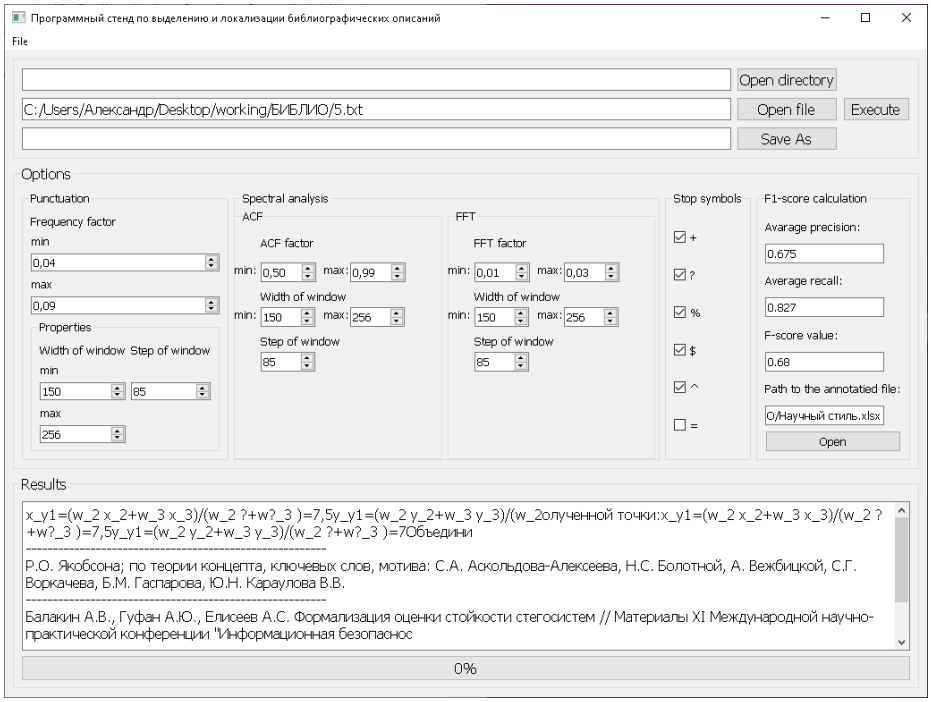


**Fig. 11.** Main window of the software research stand

## 6 Conclusions

- The lognormal law describes probability distributions of the lengths of bibliographic descriptions, as well as relative frequencies of the prescribed punctuation characters set.
- A large number of false positives were identified for scientific-style texts, which explained by an abundance of mathematical expressions with punctuation marks. Therefore, the relative frequencies of PPC is a poor predictor for long size BDs (full), which are typical for dissertations. It is not advisable to use this feature in its original form.
- The obtained values of Recall, Precision and $F_1$-score of classifiers, developed on the basis of the studied statistical features of BDs, are comparable with the results achieved in previous studies using structural recognition methods.

- It seems promising to increase the Recall value of the classification of the presented compact models in order to achieve the necessary precision at the next stage of processing – applying regular expressions to selected text fragments.
- Experimental data indicate that the studied statistical features and models are applicable not only to texts in Russian, but also to other alphabetic languages.
- Note also that there remains some potential for improving performance within the framework of the presented models by varying the sliding window size, changing the original text string encoding way into a binary sequence, changing character sets, using more advanced metrics. It is also worth considering the possibility of using neural networks to solve the stated problem.
- The presented results are preliminary and intermediate and will be further refined.

## References


1. Nasar, Z., Jaffry, S.W., Malik, M.K.: Information extraction from scientific articles: a survey. Scientometrics, 117, 1931–1990. Springer, Heidelberg (2018).
2. Chenet, M.: Identify and extract entities from bibliography references in a free text. Master thesis. University of Twente, Enschede (2017).
3. Antiplagiat service Homepage, www.antiplagiat.ru, last accessed 2020/09/21.
4. Zatsman, I.M., Havanskov, V.A., Shubnikov, S.K.: Method of bibliographic information extraction from full-text descriptions of inventions. Informatics and Applications, 7(4), 52–65. Russian Academy of Sciences, Moscow (2013).
5. Lutsenko, E.V.: The application of ASC-analysis and "AIDOS" intelligent system to solve, in general, the problem of identifying the sources and authors of the standard, nonstandard and incorrect bibliographic descriptions, http://ej.kubagro.ru/2014/09/pdf/32.pdf, last accessed 2020/09/21.
6. Sokolova, T.A.: An extraction of the elements from bibliography based on automatically generated regular expressions. In: Proceedings of the All-Russian conference with international participation «Information and telecommunication technologies and mathematical modeling of high-tech systems», 313–316. Peoples' Friendship University of Russia, Moscow (2019).
7. Kolmogortsev, S., Saraev P.: Extracting bibliography from texts with regular expressions. New Information Technologies in Automated Systems, 20, 82–88 (2017).
8. Tkaczyk, D., Szostek, P., Fedoryszak, M. et al.: CERMINE: automatic extraction of structured metadata from scientific literature. International Journal on Document Analysis and Recognition (IJDAR), 18, 317–335. Springer, Heidelberg (2015).
9. Graschenko, L.A., Cherkasov, N.V., Kuzmin, N.S.: Experience of automatic localization of bibliographic descriptions in Russian-language texts. New information technologies in automated systems, 22, 192–198 (2019).
10. Russian State Library Homepage, www.rsl.ru, last accessed 2020/09/21.
11. National Electronic Library Homepage, http://rusneb.ru, last accessed 2020/09/21.
12. Scientific Electronic Library Homepage, www.elibrary.ru, last accessed 2020/09/21.
13. Russian National Corpus, http://www.ruscorpora.ru/new/, last accessed 2020/09/21.