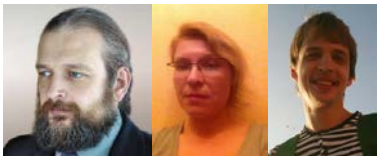
The search for additional features for the improving of a joint ("two-dimensional") classifier of genre types and stylistic colouring of poetic texts

V Barakhnin[1, 2*], O Kozhemyakina[1], I Pastushkov[1]

[1]*Institute of Computational Technologies of SB RAS, Lavrentiev av., 6, 630090, Novosibirsk, Russia*
[2]*Novosibirsk State University, Pirogov str. 1, 630090, Novosibirsk, Russia*

**Corresponding author's e-mail: bar@ict.nsc.ru*



Abstract

In this paper we propose the usage of additional features for the improving of a joint ("two-dimensional") classifier of genre types and stylistic colouring of poetic texts in Russian. We expand a range of the features to increase the precision of classifier. On the basis of these approaches the principles of formation of the training samples for the algorithms for the definition of styles and genre types were analysed. The computational experiments with a corpus of texts of the lyrics of A.S.Pushkin from 1818 to 1825 and of the whole lyrics of K.N.Batyushkov were implemented, which showed good results in determining the stylistic colouring of poetic texts and sufficient results in determining the genres. The proposed algorithms can be used for automation of the complex analysis of Russian poetic texts, significantly facilitating the work of the expert in determining their styles and genres by providing appropriate recommendations.

Keywords: automated analysis, computational experiments, "two-dimensional" classifier, genre, style

1 Introduction

In the tasks of automated text analysis in natural language, the problem of determination of their genre and stylistic characteristics is determined. The researcher can get this problem in a wide range of situations: from the problems of automation of the complex analysis of poetic texts, for which the type of genre and stylistic characteristics are the important attributes used in determining of the impact of lower levels on higher levels of the verse (see for example [1]), to the tracking of messages in social networks to identify the terrorist threats, the determination of marketing preferences of buyers, etc.

The researches in the field of automated determination of the genre type of texts were started recently – in early 2010-ies. So, in work [2] the algorithms of determination of genre types of odes, songs, epistles, elegies and epitaphs are based on the works of English poets-sentimentalists of the XVIII century. The time period in this study was not chosen by chance: in the poetry of the XVIII century the classicism with its strict genre rules dominated, and this greatly facilitated the development of algorithms.

The paper [3] describes the method of text classification (for certain genres and authors) based on the analysis of statistical regularities of letter distributions, i.e. the probabilities of occurrence of letters and letter combinations, along with this a solution is found without the "invasion in the sphere of literature, i.e., without the analysis of syntax, literary techniques and patterns of character interactions". However, in [4], the authors build an original counterexample to the statistical method of identification that shows the necessity of using, at least, the methods of morphological analysis.

As for the automation of determination of stylistic characteristics of the texts, we don't know the researches in this area, at least for the texts in Russian. Thus, our researches on computer joint definition of the type of genre and stylistic colouring of Russian texts are of a pioneer nature.

In the present work we suggest and use the additional features for the improving of a joint ("two-dimensional") classifier of genre types and stylistic colouring of poetic texts. Our purpose is not the creation of new theories of genre and stylistic relationships within literary works but the development of the analyser that allows to correlate correctly the stylistic colouring of the text with its genre identity what has relevance for researches in the field of Informatics, because we are talking about the tools used not in the strictly linguistic space.

2 The choice of training data

While we built the joint ("two-dimensional") classifier of genre types and stylistic colouring of texts, we took into account that the classifier itself is a multidimensional structure, based on the totality of parameters, which define the object of study. When we construct the multidimensional classifiers associated with such difficult (for unequivocal definition) categories like genre and style, the phased development of each analysis parameter is required in order to exclude possible errors and the variability of results. Such classifier is created for the first time (at least for texts in Russian). For the analysis we take the lyrics of A.S.Pushkin from 1818 to 1825 and the whole lyrics of K.N.Batyushkov. We confirm the results received on the material of the lyrics of A.S.Pushkin of Lyceum period (that we took as training sample on the first stage of

the researches), and we make the experiments with classifier using new features.

Genre types formed the basis of the classifier: along one axis we have placed the genre types in order of ascending "the sublimity" and along another axis - the traditional styles (see Table 1, Table 2).

TABLE 1 The statistics on the genre and stylistic compliance, Pushkin (1818-1825)

	High	Neutral	Low
Ode	4	-	-
Epistle	17	66	9
Madrigal	-	1	-
Satire	-	-	2
Idyll	1	1	-
Tale	-	-	3
Song	-	2	-
Elegy	28	49	1
Anacreontic	1	1	-
Epigram	-	21	24
Ballad	2	2	-
Anecdote	-	1	1

TABLE 2 The statistics on the genre and stylistic compliance, Batyushkov

	High	Neutral	Low
Ode	4	-	-
Epistle	9	15	1
Madrigal	-	1	-
Idyll	1		-
Elegy	22	20	-
Satire	1	1	3
Epitaph	3	-	-
Epigram	-	-	18
Anecdote	-	-	1

3 Description of the numerical experiment

For the new experiment we used the above-described massive of Pushkin's lyrics (1818-1825), comprising 247 poems, and of Batyushkov's lyrics, comprising 105 poems, marked by an expert on genres and styles.

All the corpus of texts was devided into three parts: the lyrics of A.S.Pushkin of Lyceum period, the lyrics of A.S.Pushkin from 1818 to 1825 and the whole lyrics of K.N.Batyushkov. For each poem we extract some features among the features of TF-IDF matrix, the usage of the words from a poem from the dictionary of Russian language of XVIII century, the metro-rhytmic features like a rhyme type, size, stanza, number of male and female endings and the year when a poem was written. The usage of old Slavonic and old church Slavonic was determined by the difference between the dictionary mentioned above and Russian Wikipedia corpus and was decoded as vector with quantity of usage for each of these words. There also the SMOTE algorithm and the random oversampling for solving problem of class minority were used. After the extraction of the features the logistic regression was learned and the features that more important than others for taking decision were used for the learning model of multilayer perceptron [5]. The achieved average f1-measure is about 95% by cross-validation of the 3 partitions for each author, but it's worth noting that some classes behave worse because of being represented by too few examples.

The experimental results are following (see Table 3): we calculated the average, the minimum, and the maximum of the f1-measure of correct predictions of the method with cross-validation (algorithm was implemented in the language Python using the library scikit-learn).

TABLE 3 Experiment with the definition of the genre on multilayer perceptron with features best for logistic regression

	Average f1-measure	Max	Min
Pushkin	0.93	0.95	0.92
Batyushkov	0.91	0.92	0.89

4 Conclusions

The search for additional features for the improving of a joint ("two-dimensional") classifier of genre types and stylistic colouring of poetic texts helps to obtain the best results. We implement the computational experiments to the corpus of texts of the lyrics of A.S.Pushkin from 1818 to 1825 and the whole lyrics of K.N.Batyushkov, and these experiments showed good results in determining the stylistic colouring of poetic texts and sufficient results in determining the genres. The using of large number of heterogeneous features requires a certain approach to the architecture of the classifier, but it gives the greater accuracy of predictions.

References

[1] Barakhnin V, Kozhemyakina O 2012 About the automation of the complex analysis of Russian poetic text *CEUR Workshop Proceedings* **934** 167-71 (*in Russian*)
[2] Lestsova M 2014 The determination of the core and the periphery of the genres of odes, songs, epistles, elegies and epitaphs on the works of English poets-sentimentalists of the XIX century *Bulletin of the Chelyabinsk State Pedagogical University* No 4 196-205 (*in Russian*)
[3] Orlov Yu, Osminin K 2010 The definition of the genre and the author of a literary work by statistical methods *Applied Informatics* **26**(2) 95-198 (*in Russian*)
[4] Orlov Yu, Osminin K 2012 *Methods of statistical analysis of literary texts* Editorial URSS: Moscow (*in Russian*)
[5] Haykin S 1998 Neural Networks: A Comprehensive Foundation 2nd Prentice Hall PTR Upper Saddle River, NJ, USA