Automated Determination of the Type of Genre and Stylistic Coloring of Russian Texts

Vladimir Barakhnin[1,2,*], *Olga* Kozhemyakina[2], and *Ilya* Pastushkov[1]

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

Abstract. In this paper we propose the algorithm of automated definition of the genre type and semantic characteristics of poetic texts in Russian. We formulated the approaches to the construction of a joint ("two-dimensional") classifier of genre types and stylistic colouring of poetic texts, based on the definition of interdependence of the type of genre and stylistic colouring of the text. 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 analyzed. The computational experiments with a corpus of texts of the Lyceum lyrics of A.S.Pushkin 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.

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 analysys 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 coloring of Russian texts are of a pioneer nature. In the present work we implement to develop the approaches to the construction of a joint classifier of the types of genre and stylistic colouring of poetic texts as well as we made the comparative analysis of algorithms of determination of these characteristics.

We underline that our purpose is not the creation of new theories of genre and stylistic relationships within literary works but the development of the analyzer that allows to correlate correctly the stylistic coloring 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 building of the joint classifier of genre types and stylistic colouring of poetic texts

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. The multidimensionality of the structure itself implies some

* Corresponding author: bar@ict.nsc.ru

error in the final result of the analysis, because the more we put into the system than the more potential probability of the occurrence of many variants at each stage of the analysis is possible. So, 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, what is crucial in the case of unstructured text analysis.

We are working on the joint classifier of genre types and stylistic colouring of poetic texts. Such classifier is created for the first time (at least for texts in Russian).

The creation of a "universal" classifier of literary texts according to their genre and stylistic characteristics requires the synthesis of a vast empirical material, so we decided to stay on the classification of lyrics of A.S.Pushkin of Lyceum period. The rationale of this choice is given below.

The most efficient approach to automatic determination of the genre types is the usage of algorithms with learning. However, the formation of the training sample is not a trivial task. Our attempt to use as a training sample the lyrics of Pushkin's of mature period (1828-1831), was failed at an early stage, since the variety of genres of Pushkin's work of this period is very subjective, in correlation with the stylistic features of the works of Pushkin in a special manner, herewith it does not follow the generally accepted laws. This feature is formulated by V.A.Grekhnev: "The genres and style does not confront each other as hostile, dening fundamentals, but there is always internal tension between them. This tension is increasing, where the power and scope of the writer's personality are increasing" [5. P. 234]. This fact causes the genre and stylistic varieties and variants, the "internal tension" between style and genre originate the non-canonical genres, and for training samples this becomes critical, as there are features that do not fit into the system and, therefore, are contrary in essence to the material for the construction of the system. As a result, we decided to stay at the Lyceum lyrics (1814 – 1817), as it has the most strict genre forms, stylistic unity, and adherence to the rules of grammar of that period: "Almost all the Lyceum lyrics belongs to the sublime style, except a few poems. Even many of the satirical poems are written entirely in the sublime style. It can be argued that in the early poems of Pushkin, there is the influence of the rigid rules of "Grammar" of his Lyceum teacher N.F.Koshanskiy" [6. P. 24].

In turn, the usage of the Lyceum lyrics as material for the creation of training samples, is justified by stylistic dimension, since the stylistic differentiation of lexemes is the development stage of the classifier. For texts in the Russian language there is the upward to the works of M.V.Lomonosov [7] the division of the texts (primarily literary) for the relevant to high, neutral and low styles. Historically, each of them is characterized by the ratio of the usage of the old Slavonic (Church Slavonic) and the Russian words (a group of words common to the Slavonic and Russian languages is separately considered), a share of archaisms, and the usage of certain syntactic constructions. In the classic theory the genre strictly dictates the choice of a particular style. So, with regard to the poetic texts this dependence is discussed in detail in the monograph by D.M.Magomedova [8].

So, for implementation of given task, we are going from practice, making a selection of the works of Pushkin Lyceum period, from 1813 to 1817, as material, on which we can build the most accurate system, what certainly makes the end result of analysis the most accurate and allows to develop the most appropriate classifier relating to the stylistic aspect. We exclude the poems, the tales, the translations, Dubia, and we do the list, statistically including the poems, as relevant to the genre system given in [8] and also not included in this system. So this system includes the canonical and non-canonical genres, the latter, however, occur after the considered period, and therefore they are not described in the list.

As the result of analysis of this list of works, we distinguish the following groups of genres.

Canonical: ode – 4 poems, elegy – 27 poems (including one historical elegy – "Napoleon on Elba", idyll – 2 poems, epistle – 55 poems, ballad – 3 poems, non-canonical (a fragment, a story in verse) – no.

Also, we're adding the genres, which are not in the system of canonical and non-canonical: epigram - 18 poems, madrigal – 4 poems, sonnet – 1 poem, romance – 1 poem, anecdote –1 poem, Parable – 2 poems.

In addition, the poem "Unbelief" (1817) is defined as philosophical ode and elegy [9]. But for the analysis we defined it as philosophical ode.

Genre types of these works formed the basis of the classifier: along one axis we have placed the genre types in order of ascending "the sublimity": an ode, the elegy, the idyll, the epistle, etc., and along another axis - the traditional styles (see Table 1). On this empirical material an obvious correlation between genre and stylistic characteristics of texts can be seen: an ode, the elegy, the idyll are usually written in high style, they do not use lexical content corresponding to the low style, but the epigrams, on the contrary, are characterized by the usage of elements of the vocabulary of the low style. In general, the style of a text is determined by the most "low" of its lexemes, what is especially typical for epigrams: the presence of high vocabulary, which is often used in an ironic way, must not mislead, because the usage of one or two words of spoken or obscene vocabulary characterize immediately the author's intention. However, for the genres that traditionally imply the sublime form, especially, a madrigal, we do not consider to refer the poems of these genres to reduced style if the several "reduced" (but not obscene!) words are used in them with ironic purposes.

It should be noted that a style is characterized by the lexemes much more than a genre, although, in our experiment, in particular, and to the global literary categories processes and, in general, the number of genres is much more than a number of styles. This complicates the choice, as because of direct factors, and also because of given training sample of works. The features of training sample include works written in the

genre of the parable, one of them ("Riders") belongs to high style, the second ("Truth") – to the neutral, although, as we know, the parable, being the closest to the genre of the fable, suggests the possibility of writing it in different styles, as evidenced, in particular, by the parable "The Cobbler", which can be attributed rather to the low ("colloquial") style.

Table 1. The statistics on the genre and stylistic compliance.

	High	Neutral	Low
Ode	4	-	-
Parable	1	1	-
Madrigal	4	-	-
Epistle	-	55	5
Idyll	-	2	-
Elegy	-	37	-
Romance	-	1	-
Ballad	-	3	-
Epigram	-	-	18
Anecdote	-	-	1

3 About the creation of the dictionary of stylistically differentiated lexemes

Before the choice of algorithms of determination of stylistic and genre characteristics of poetic texts, it is necessary to answer the question: is it possible to use a priori compiled dictionaries of lexemes with a particular stylistic or genre painting?

Great attention is paid on the stylistic differentiation of words in the monograph of O.S.Akhmanova "Essays on General and Russian lexicology" [10]. The list of "spoken" words, with "reduced" stylistic characteristics and with "increased" stylistic characteristics. However, these lists are not complete and rather more illustrative, moreover, the author concedes that "not all included words are equally convincing (many of them, undoubtedly, will seem disputable)", and finally, the stylistic colouring of some of the lexemes have changed over time, i.e. this feature, taken from the monograph [10], could be other for the language of the XIX century, as for the modern language. Therefore, in the same monograph, for relation of the words with a particular style it is proposed to use the analysis of their structural-semantic forms. So, nouns with the suffix -к-а in a variety of structural-semantic variants, and also the various suffixes meaning "a person" refer to "conversational" or "reduced" vocabulary; for "spoken" in contrast to "reduced", the vocabulary is characterized by a large number of dialects; for "book" vocabulary the borrowed words are typical, and for the "sublime" – Slavic words with a complex structure, as well as archaisms, etc.

However, all these observations are of a rather private nature. So, the words with the suffix *-к-а*: *пытка*, *речка*, *шутка*, etc., are found in Pushkin's verses which are not related to "low" or "conversational" style, the same applies to the words *бочка*, *кружка*, *пушка*, etc., in which *-к* is a part of the root, but the establishment of this fact requires a non-trivial etymological analysis, which is difficult to automation. Similarly, the borrowed words over time have become available to all styles, and it's not only about "ancient" borrowings like *лошадь* or *собака*, but also about new: *велосипед*, *танк*, etc.. The slavonicisms, including the words with complex structure, could be used, in particular, to give the poem an ironic tone (e.g., Pushkin's "Ode to his Excellency Graf D.I.Khvostov" and numerous satirical verses by A.K.Tolstoy).

The situation is complicated as by the fact that often not all the word's variants are belong to "conversational" or "reduced" style, but only one of its lexical-semantic variants, as well as because of the gaining the word of a particular stylistic colouring only at the entry to the idiom. Thus, the occurrence in the text of certain lexemes cannot serve as a sufficiently reliable criterion for attribution of the text to a particular stylistic type because most of the words are of polysemantic colouring.

Moreover, a strict definition of genre colouring of separate words seems quite a unpromising task to us and we don't know any satisfactory attempt at its resolution at least on a theoretical level. That is why it seems to us that the most appropriate way is to define the stylistic and genre characteristics of poetic texts on the basis of the occurrence in them of a set of lexemes defined by the training samples.

4 Description of the numerical experiment

For the experiment we used the above-described massive of the texts of Pushkin's lyrics of Lyceum period, comprising 121 poems, marked by an expert on genres and styles.

We used the standard method of support vectors machine (support vector machine) [11] with a linear core and the RBF nonlinear core, in addition, for comparison, the results of calculations were carried out with the use of neural network based on multilayer perceptron [13]. When training the dictionary of all used words was created, except service words, and each text was coded by sequence of the symbols 0 and 1 corresponding to the dictionary in word order: 0 was set if a word is not in the text, 1 – if the word is in the text. Also we used the linear regression to determine the styles, our hypothesis was that as styles can be unambiguously ranked: low – 1, neutral – 2, high – 3, the regression can give a numeric value which will be close to the value of the style, and a divergence with the value will be a mistake. The experimental results are following (see Table 2): we calculated the average, the minimum, and the maximum of the proportion of correct predictions of the method with 100 runs, the sample is divided into 80% training and 20% test, the division into which is random every time, each run is independent from the previous ones (algorithm was implemented in the language python using the library scikit-learn). As it is difficult to rank a

neutral along with the others, than in each method there is the experiment with it and without it.

As can be seen from the obtained data, the high style is not practically recognized – probably because of its insufficient representation in the sample. The method of support vector machines is the best in this case. It is worth to note that in the case of non-linear core the high style was recognized, but by the common parameters, the case of linear core is better than the multilayer perceptron and logistic regression.

Table 2. The statistics on the genre and stylistic compliance.

	Average value	Max	Min
SVM, neutral is ignored	0.76	0.92	0.58
SVM	0.80	0.96	0.57
SVM, rbf core	0.62	0.85	0.11
Multilayer neural network	0.77	0.96	0.46
Logistic regression	0.76	0.96	0.46
Linear regression, neutral is ignored	0.70	0.82	0.45
Linear regression	0.70	0.45	0.58

	High	Neutral	Low
SVM, neutral is ignored	0.0	0.86	0.72
SVM	0.0	0.86	0.70
SVM, rbf core	0.10	0.75	0.13
Multilayer neural network	0.0	0.96	0.33
Logistic regression	0.0	0.85	0.72
Linear regression, neutral is ignored	-	-	-
Linear regression	-	-	-

Similarly, we carried out the experiment on definition of the genre (one series of experiments was carried out under the simplified scheme, when the historical elegy and philosophical ode was not seen as separate genres). From Table 3 it is seen that the definition of genre has fared worse than the definition of styles as each genre is represented by a relatively small number of samples. The lexical signs are not enough for genres, we need poetic features (rhyme, size, number of accented syllables) which should strengthened, for example, with the help of the AdaBoost algorithm [13].

Table 3. Experiment with the definition of the genre.

	Average value	Max	Min
SVM. simplified types	0.45	0.65	0.22
SVM	0.45	0.65	0.27

5 Conclusion

The paper proposes the approaches to the construction of a joint ("two-dimensional") classifier of genre types and stylistic colouring of poetic texts, based on the definition of interdependence of the type of genre and stylistic colouring of the text. On the basis of these approaches we analyse the principles of formation of the training samples for the algorithms to define styles and genre types. We implement the computational experiments with a corpus of texts of the Lyceum lyrics of A.S.Pushkin, which showed good results in determining the stylistic colouring of poetic texts and sufficient results in determining the genres. Thus, the proposed algorithms showed their efficiency and 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.

Work is executed with partial support of the Presidium of RAS (project 2016-PRAS-0015) and of the Presidential programme "Leading scientific schools of RF" (grant 7214.2016.9).

References

1. V.B. Barakhnin, O.Yu. Kozhemyakina, CEUR Workshop Proceedings, 934 (2012) (in Russian)
2. M.A. Lestsova, Bulletin of the Chelyabinsk State Pedagogical University, 4 (2014) (in Russian)
3. Yu.N. Orlov, K.P. Osminin, Applied Informatics, **26,** 2 (2010) (in Russian)
4. Yu.N. Orlov, K.P. Osminin, *Methods of statistical analysis of literary texts* (Editorial URSS, Moscow, 2012) (in Russian)
5. V.A. Grehnev, *Lyrics of Pushkin. About the poetics of genres* (Gorkiy, 1985) (in Russian)
6. M.V. Lomonosov, The preface about the advantages of Church books in Russian language, In: M.V.Lomonosov, Complete Collection, **7 (**Moscow, Leningrad, 1952) (in Russian)
7. V.B. Barakhnin, O.Yu. Kozhemyakina, Vestnik of Tomsk State University. Philology, **13**, 2 (2016) (in Russian)
8. D.M. Magomedova, *Linguistic analysis of a lyric poem*. (Moscow, 2004) (in Russian)
9. S.F. Svobodina, Pushkin Museum: an almanac, **6** (2014) (in Russian)
10. O.S. Akhmanova, *Essays on General and Russian lexicology* (Moscow, 1957) (in Russian)
11. N. Cristianini, J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods* (Cambridge University Press, 2000)
12. S. Haykin, *Neural Networks: A Comprehensive Foundation* (Prentice-Hall, 1999)
13. Y. Freund, R.E. Schapire, Journal of Japanese Society for Artificial Intelligence, **14**, 5 (1999)