Linguistic Modeling as a Basis for Creating Authorship Attribution Software[1]

Khomenko Anna
HSE University
Nizhny Novgorod, Russia
akhomenko@hse.ru

Baranova Yulia
CarrierX
Nizhny Novgorod, Russia
ligros7@gmail.com

Romanov Alexander
Tomsk State University of Control Systems and Radioelectronics
Tomsk, Russia
alexx.romanov@gmail.com

Zadvornov Konstantin
HSE University
Nizhny Novgorod, Russia
zadvornovk@mail.ru

Abstract

This paper discusses approbation of an integrative attribution method for texts in the Russian language. The methodology goes after (Koppel, Schler 2003): computer program tries to imitate human expert work. So, it is based on interpretative language study with its objectification through mathematical statistics. The choice of parameters describing the author's individual style is rooted to considering text to be a product of an authentic language personality. Language personality is described using psycholinguistic (Yu.N. Karaulov), sociolinguistic (M.Coulthard, R. W.Shuy) methods and the methodology of forensic linguistics (S.M. Vul, D.Wright). On the basis of the principles above, the software for attribution is created: http://khorom-attribution.ru/#/. As output the resource displays mathematical models of persons' individual styles and the metrics for null hypothesis evaluation: Pearson correlation coefficient, linear regression and Student's t-test. The functionality of the resource is aimed to solve an identification problem of text attribution for «closed class» (Juola 2008) with pair-wise comparison, but the resource can also be used in the personality diagnostics in forensic, philological and cultural researchers.

Keywords: authorship attribution; language personality; linguistic model; mathematical model
DOI: 10.28995/2075-7182-2021-20-1063-1074

Лингвистическое Моделирование как Основа для Создания Электронного Атрибуционного Ресурса[2]

Хоменко Анна
НИУ «Высшая школа экономики»
Нижний Новгород, Россия
akhomenko@hse.ru

Баранова Юлия
CarrierX
Нижний Новгород, Россия
ligros7@gmail.com

[1] The authors are grateful to the members of scientific research group "Automatic Systems in Natural Language Structures Modelling" of HSE University (Nizhny Novgorod): Babiy Alexander, Kazyulina Marina, Kondakova Veronika, Vaskina Lidia, Klimova Margarita, Troshina Anastasia, Merkulova Ekaterina, Mikhailova Anna, Rashchupkina Irina, Ivleva Polina, Mufazalova Irina, for help in article preparing.

[2] Авторы благодарят членов научно-учебной группы «Автоматические системы в моделировании естественного языка» НИУ ВШЭ (Нижний Новгород): Бабия Александра, Казюлину Марину, Кондакову Веронику, Васькину Лидию, Климову Маргариту, Трошину Анастасию, Меркулову Екатерину, Михайлову Анну, Ращупкину Ирину, Ивлеву Полину, Муфазалову Ирину, за помощь в подготовки статьи.

Романов Александр
Томский государственный
университет систем управления
и радиоэлектроники
Томск, Россия
alexx.romanov@gmail.com

Задворнов Константин
НИУ «Высшая школа экономики»
Нижний Новгород, Россия
zadvornovk@mail.ru

Аннотация

В статье речь идет об апробации интегративного атрибуционного алгоритма. Он основан на анализе идиостиля автора письменного текста методами интерпретативной лингвистики с последующей объективацией полученных данных с помощью математической статистики. Алгоритм решает идентификационную проблему атрибуции. Выбор параметров, описывающих индивидуальный стиль автора, основан на рассмотрении текста как продукта аутентичной языковой личности. Языковая личность описывается с использованием психолингвистических (Ю. Н. Караулов), социолингвистических и судебно-лингвистических (С. М. Вул, M.Coulthard, R. W.Shuy) методов. Для проверки гипотезы о том, что именно интегративная методика является наиболее эффективной при решении идентификационной задачи атрибуции, было создано электронное приложение «ХоРом», кумулирующее в себе описанные выше подходы к анализу языковой личности: http://khorom-attribution.ru/#/. С помощью ресурса можно сравнить две модели языковой личности и определить уровень их сходства посредством следующих метрик: коэффициента корреляции Пирсона, коэффициента детерминации линейной регрессии и t-критерия Стьюдента.

Ключевые слова: атрибуция; языковая личность; лингвистическая модель; математическая модель

1 Credits

The problem of text attribution in modern linguistics is becoming increasingly relevant. Since L. Campbell [5] and V. Lutoslawski [26] in the West and N.A. Morozov [32] in Russia, attribution linguistics has followed two parallel paths: stylometry (Mendenhall T. [30]; F. Mosteller, D. L. Wallace [33]; C. Labbe, D. Labbe [21]; J. Burrows [4]; T. Merriam [31]; P. Juola, J.Sofko, P. Brennan [15]; G. Ya. Martynenko [27]; T. Litvinova, P. Seredin, O. Litvinova [24]; D. Wright [50]; S.V. Ionova, I.V. Ogorelkov [8] etc.) and interpretative text analysis (A.Yu. Komissarov [18]; G.R. McMenamin [29]; E.I. Goroshko [13]; E.I. Galyashina [10]; M. Coulthard [6]; S.M. Vul [49]; I.I. Rubtsova, E.I.Ermolaeva, A.I.Bezrukova [44], etc.). Nevertheless, nowadays a trend towards fully automated systems that use different models, algorithms and metrics is being formed. They include those based, for example, on n-grams (B. Murauer, M. Tschuggnall, G. Specht [34]; L. Muttenthaler, G. Lucas, J. Amann [35]), POS-tags (T. Litvinova, A. Sboev, P. Panicheva [23]), sentence, word lengths (J.E. Custódio, I. Paraboni [7]), using clustering (P. Panicheva, A. Mirzagitova, Y. Ledovaya [38]) and vector (A.Bacciu and others [1]) approaches, traditional (A. Gomzin, A. Laguta, V. Stroev, D.Turdakov [12]) and modified (M. Korobov [20]) Python libraries. Many of the quantitative approaches are productive and show high level results, but they consider the individual style to be a series of linguistic probabilities, not a product of individual's speech ability and comp-tence. Thus, using only quantitative approaches based on the collection of traditional stylometric features, even in a large number of them (M. Bhargava, P. Mehndiratta, K. Asawa [2]), it is impossible to create a complete model of the author's individual style that adequately reflects an author's language personality. Psycholinguistic, sociological and cognitive approaches to an individual style certainly help to make the model of an author's language personality more complete. There has been a successful at-tempt to use the integration of approaches above (quantitative and qualitative) and vector text representation in the research by (E. Pimonova, O. Durandin, A. Malafeev [39]). From our point of view, the idea of integration is quite relevant.

2 Introduction

In our work, we propose an integrative approach based on understanding the individual style both as a combination of language probabilities (quantitative approach based on stylometry), and as a result of a specific language personality representation (qualitative approach based on interpretative linguistics). It allows to create a fairly complete, comprehensively imitating the original and adequate model of author's individual style. Interpretative linguistics methods reveal information about the personality's thesaurus, pragmaticon, grammar and lexicon, and stylometry allows to objectify the interpretative analysis results. Moreover, such approach should be universal and solve attribution problems both for scientific and pragmatic purposes, including forensic ones. At the same time, it must solve the problem of authorship attribution for texts of different genres and lengths.

3 Research Methods

3.1 Theoretical basis

The most suitable parameters for attribution model should reflect language personality as a result of cognition process, identify the author's individual style and at the same time could be extracted from the text automatically with minimum preprocessing (tokenization, lemmatization, POS-tag annotation). To define them an expert manual study of 10 multi-genre text blocks (116 thousand words) was conducted. They are universal for text of any genre or length and easy to be extracted using some predefined rules. The parameters are distributed over three levels of language personality in Yu. N. Karaulov's [17] conception:

- pragmaticon level (the level of speech strategies and tactics): sentences with homogeneous parts, sentences with appositions, parenthetic words and phrases explicating subjective modality; purpose, emphasizing and comparative syntactic structures representing the level of the author's competence in writing and attitude towards reality; syntactic blends giving an idea of the functional style of the text; verbal mononuclear sentences explicating the representation of reality; complex sentences; address forms as a phatic element – in total 11 constructions and 107 custom-built rules for extracting them from the text;
- thesaurus level (cognitive worldview): this section includes the most frequent combinations of words that describe grammatical and semantic features of the text (word bigrams and trigrams); key lexemes; explicators of axiological text dominants of the ʻus/them' dichotomy – in total 3 standard algorithms and 1 authentic, rule for extracting linguistic information;
- lexicon level (lexical and grammatical competence): parts of speech (the number of independent parts of speech and their ratio: B. N. Golovin' s coefficients including coefficient of connexion and others [11], Gunning fog index, Flesch-Kincaid readability tests with coefficient for the Russian language [47: 679], etc.), hyphenated words; modal particles, interjections, presence/absence of the modal postfix '-то', preferred intensifiers; number of misspelled words and typos – in total 10 standard algorithms and 32 authentic, custom-built rules for extracting linguistic information from the text.

3.2 The principles and peculiarities of search and modeling procedure

To create the rules morphological tagging, information about semantic valence of words, information on structural schemes of the Russian language and information on punctuation are used.

Extracting of pragmaticon parameters is connected with principles of semantic syntax [37], grammar of Russian [45], and based on POS-tags and punctuation. For example, the formalized rule (search algorithm) for finding explicators of subjective modality is the following:

- a dictionary of subjective modality explicators is created;
- a punctuation rule, which allows to overcome homonymy is prescribed:

1. __, Prnt, __
2. Prnt, __

where Prnt is any part of speech; __ - some part of a sentence.

Purpose syntactic structure rule is based on semantic valence and structural principles of linguistic constrictions [22]. Compound prepositions 'с целью/из расчёта' (for the purpose of/in order to) require an infinitive implementing purpose semantics. Thus, the rule is: 'с целью/из расчёта' + INFN, where INFN – infinitive.

Defined personal sentences could be found with the help of the algorithm below:

1. + V, 1per / 2per, sg / pl, praes / fut, indic
2. + V, sg / pl, imper

3. – N / SPRO, nom, sg / pl
4. – NUM, nomn _+ N в gen/ gen2, pl
5. – «много/мало/несколько/немного/немало» _ + N в gen/ gen2, pl.

where: «+» at the beginning of the scheme – the presence of an element in the sentence; «+» between the elements – the presence of both elements in the scheme; «-» at the beginning of the scheme – the absence of an element in the sentence; «/» - designation of "or"; «_» – possible presence of one or two words; V is a finite verb; 1per / 2per – the first and the second person respectively; sg / pl - singular and plural, respectively, praes / fut - present and future tense, respectively; indic - indicative; imper - imperative; N is a noun; SPRO - pronoun-noun (a pronoun that has the semantics and syntactic function of a noun); nom – nominative case; NUM - numeral; gen / gen2 - genitive and second genitive, respectively. Nomenclature is taken from the Russian National Corpus: https://ruscorpora.ru/new/corpora-morph.html.

Constructing rules for units of author's lexicon is based on morphological annotation. Modal postfix '-то' rule is the following: POST-то, other than SPRO or APRO in any case in plural or single form, where POST – any part of speech, SPRO/APRO – pronoun with noun/adjective semantics and syntactic function.

An intensifier refers to a lexeme used to determine the degree of the semantic category of intensity. Most often, intensifiers are adverbs, the number of which is large yet limited. Nevertheless, the category of intensity is not limited exclusively to adverbs, for example:

(1) *Какая красота!*
What a beauty!
'What a beauty!'

— in this case, the pronoun *какая* (what) serves as an intensifier. Thus, in this study, we have made a set of rules to search for structures containing intensifiers (in total 16 rules); the list of intensifiers includes adverbs, some adjectives and pronouns (in total 93 units) in relevant grammatical structures, such as: ADJ in direct cases in singular or plural form + NOUN, where ADJ – adjective:

(2) *Настоящий бардак.*
Real mess.
'Real mess.'

To extract linguistic structures Pymorphy2 is used for morphological information finding and NLTK with a model for the Russian language are used for syntactic information analysis. NLTK is also used for creating unique, custom-built rules for text structures search because of its convenience for this purpose. After extracting the above morphological and syntactic information from the texts, the absolute frequencies of each parameter occurrences are converted into relative frequencies (ipm is used) that allow to compare texts of different lengths. Instance per million for lexical material is carried out in the standard way, for syntactic parameters, the value of each parameter is divided by the number of sentences in the text.

The thesaurus level of the language personality is the most difficult for formalization. It is not difficult to create a material explication of the author's thesaurus [3]. Nevertheless, to determine how the units in thesaurus "are arranged in a hierarchical system indirectly reflecting the structure of the world" [translation ours] [17] is extremely difficult. That is why this level is represented by the smallest number of parameters.

The most frequent word combinations in the texts are identified by the absolute frequencies of their occurrence next to each other taking into account that the lexeme must not be included in the list of stop words. Key lexemes are determined using the logarithmic likelihood algorithm when comparing the text with a large reference corpus (Opencorpora, URL: 1,540,034 words). As a result, for each text we obtain a list of keywords with the loglikelihood score (LL) numerical explication. The final list only includes words with LL greater than 50.

Before analysing key lexemes and the most frequent word combinations, the combinations with personal and proper names are removed from the lists, as these lexemes mark the text theme rather than the peculiarities of the author's individual style.

Explicators of axiological textual dominants of ‘us/them’ groups in this study refer to the dispersion of the pronouns ‘I/we-groups’, ‘you/they-groups’, i.e. pronouns of all categories in direct and indirect cases are counted for the relevant groups.

3.3 Application architecture and tools

The algorithm for parameters extraction has the following architecture:

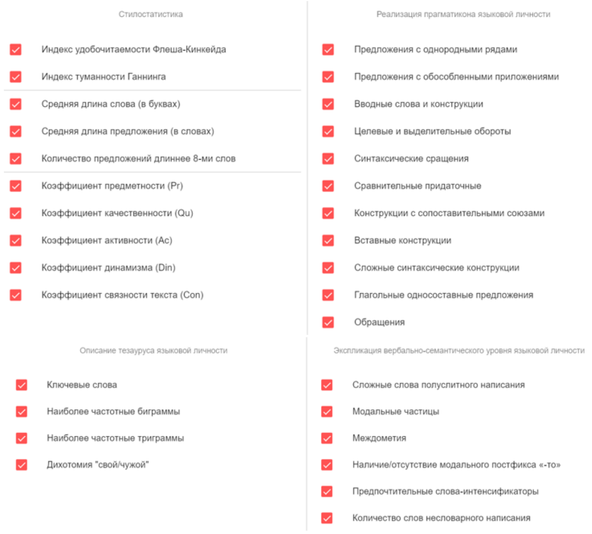
1. Input
2. Automatic extraction of parameters describing author’s individual style
2.1. Preprocessing
2.1.1. Sentence-splitting
2.1.2. Tokenization
2.1.3. Morphological parsing
3.2. Processing
3.2.1. Stylometry block
3.2.1.1. Calculation of basic metrics (number of words, sentences)
3.2.1.2. Search for traditional stylometric textual data (n-grams, indices)
3.2.2. Cognitive block
3.2.2.1. Search for parameters by preset rules
3.2.2.2. Assigning weight to each parameter
4. Building mathematical models of the texts being compared: attribute presentation as a sequence of numerical features
5. Comparing the mathematical models calculating the degree of similarity between them using Pearson correlation coefficient, coefficient of determination of linear regression and Student’s t-test to prove or disprove the hypothesis H0 that the two texts have the same author.
6. Sending the results to a client

The resource is a single-page application built on client-server architecture with the interactions carried out through HTTP-requests. The user interface interacts with the backend sending data to the server and displaying it in the way convenient for user. Back-end is based on principle of RESTful web API and is responsible for text preprocessing and parameter calculating. Back-end is programmed using Python 3.6, Flask Restplus framework and some other components: NLTK, Pymorphy2, Requests, Pyaspeller (a Python shell for Yandex.Speller), Scipy. Back-end receives a HTTP-request containing texts and set of parameters to conclude about their similarity degree. Then it calculates the needed results and sends them back to front-end in an HTTP-request. Front-end visualizes results from the request. The application is developed on a virtual machine powered by Ubuntu operation system and operated by uwsgi and nginx web-server. Front-end is programmed using an open access Javascript-framework Vue.js with the following packages and libraries: Axios, Vue-Material, Vuex, Vue-Router, Webpack, Yarn.

3.4 The functionality and interface of the software

The algorithm presented above has been implemented in an open access prototype of attribution software named ‘KhoRom’, URL: http://khorom-attribution.ru/#/. The user module has the following functions: two texts A and B are used as input; the user can pre-select the text genre. This option is included because rules for finding parameters may vary in different discourses (for example, they are very specific for sentence ending in business e-correspondence).

The user can not only build a model based on the preset parameters (Figure 1), but also has the option to select those he/she finds the most relevant for a certain pair of texts. This functionality sets apart our software from similar attribution algorithms, for example, from those based on machine learning [2; 39: 40], where all parameters are preset by the developer, not by the exact user. This also makes the resource not fully automatic, which can be important for forensic authorship analysis in Russia where full automation of the identification process is unacceptable according to the methodology [44] and the law [36; 9].



Стилостатистика

☑ Индекс удобочитаемости Флеша-Кинкейда
☑ Индекс туманности Ганнинга
☑ Средняя длина слова (в буквах)
☑ Средняя длина предложения (в словах)
☑ Количество предложений длиннее 8-ми слов
☑ Коэффициент предметности (Pr)
☑ Коэффициент качественности (Qu)
☑ Коэффициент активности (Ac)
☑ Коэффициент динамизма (Din)
☑ Коэффициент связности текста (Con)

Описание тезауруса языковой личности

☑ Ключевые слова
☑ Наиболее частотные биграммы
☑ Наиболее частотные триграммы
☑ Дихотомия "свой/чужой"

Реализация прагматикона языковой личности

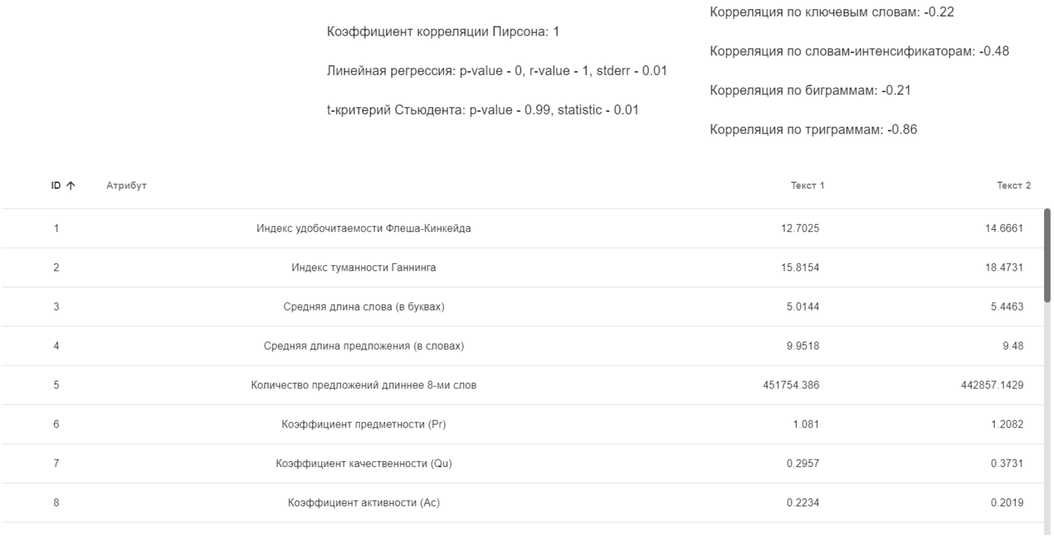
☑ Предложения с однородными рядами
☑ Предложения с обособленными приложениями
☑ Вводные слова и конструкции
☑ Целевые и выделительные обороты
☑ Синтаксические сращения
☑ Сравнительные придаточные
☑ Конструкции с сопоставительными союзами
☑ Вставные конструкции
☑ Сложные синтаксические конструкции
☑ Глагольные односоставные предложения
☑ Обращения

Экспликация вербально-семантического уровня языковой личности

☑ Сложные слова полуслитного написания
☑ Модальные частицы
☑ Междометия
☑ Наличие/отсутствие модального постфикса «-то»
☑ Предпочтительные слова-интенсификаторы
☑ Количество слов несловарного написания

Figure 1: User module of the ‘KhoRom’ resource: preset parameters

As output the system displays Pearson correlation coefficient, values of linear regression, and Student’s t-test for the models of the two texts being compared, as well as the values of each parameter for the two texts (Figure 2).



Коэффициент корреляции Пирсона: 1

Линейная регрессия: p-value - 0, r-value - 1, stderr - 0.01

t-критерий Стьюдента: p-value - 0.99, statistic - 0.01

Корреляция по ключевым словам: -0.22

Корреляция по словам-интенсификаторам: -0.48

Корреляция по биграммам: -0.21

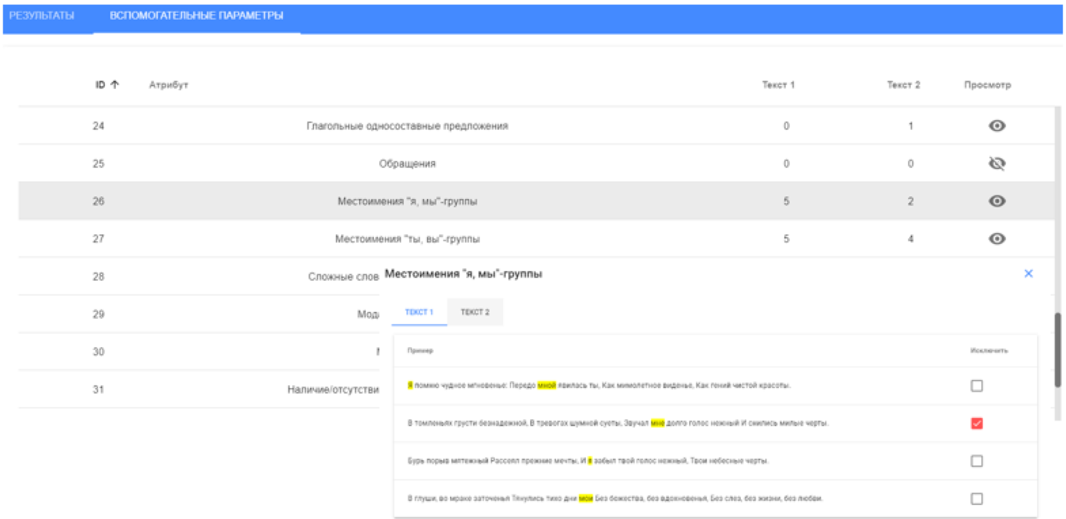
Корреляция по триграммам: -0.86

ID ↑	Атрибут	Текст 1	Текст 2
1	Индекс удобочитаемости Флеша-Кинкейда	12.7025	14.6661
2	Индекс туманности Ганнинга	15.8154	18.4731
3	Средняя длина слова (в буквах)	5.0144	5.4463
4	Средняя длина предложения (в словах)	9.9518	9.48
5	Количество предложений длиннее 8-ми слов	451754.386	442857.1429
6	Коэффициент предметности (Pr)	1.081	1.2082
7	Коэффициент качественности (Qu)	0.2957	0.3731
8	Коэффициент активности (Ac)	0.2234	0.2019

Figure 2: User module of the ‘KhoRom’ resource: output data

It is important that this block is not the final step in the developed methodology. Text statistics needs to be interpreted. Thus, for traditional mathematical statistics, Pearson correlation coefficient of more than 60% is considered significant; in the case of stylometry, it is necessary to talk about the similarity of models with a correlation coefficient of 86% and higher [40]. The software deliberately does not produce the result as a postulate “The author of the two texts is one person / The authors of the two texts are different people”. In the proposed methodology, it is the expert who makes the final decision on the text attribution, examining the statistical data according to scoring tables elaborated in the research.

The verification module is also programmed for the user: on the ‘Auxiliary Parameters’ tab, the user can view all the components, for which relative frequencies and other metrics have been calculated (Figure 3). Moreover, the user can make own corrections if he/she thinks that the program has selected a certain parameter implementation incorrectly; after making the corrections, it is possible to recalculate the parameter values and change the configuration of the final models.



ID ↑	Атрибут	Текст 1	Текст 2	Просмотр
24	Глагольные односоставные предложения	0	1	👁
25	Обращения	0	0	👁
26	Местоимения "я, мы"-группы	5	2	👁
27	Местоимения "ты, вы"-группы	5	4	👁
28	Сложные слов			
29	Мод			
30	!			
31	Наличие/отсутстви			

Местоимения "я, мы"-группы ×

Пример	Исключить
Я помню чудное мгновенье: Передо мной явилась ты, Как мимолетное виденье, Как гений чистой красоты.	☐
В томленьях грусти безнадежной, В тревогах шумной суеты, Звучал мне долго голос нежный И снились милые черты.	☑
Бурь порыв мятежный Рассеял прежние мечты, И я забыл твой голос нежный, Твои небесные черты.	☐
В глуши, во мраке заточенья Тянулись тихо дни мои Без божества, без вдохновенья, Без слез, без жизни, без любви.	☐

Figure 3: User module in ‘KhoRom’: the resource function of correcting inaccuracies

3.5 The results of the algorithm work

The above-described integrative attribution model is capable of solving the identification problem for the material of various length and genre. To prove these postulates, the software has been tested on texts of different discourses and volumes:

- a collection of famous authors fiction texts: 10 texts by S. Dovlatov and V. Astafiev (the average text length is about 20,000 words). Accuracy, precision and recall is 100%, F-score 1;
- a collection of modern Internet fiction texts from (‘Kniga Fanficov’, URL: https://ficbook.net/): texts of 3 female and 4 male authors; the total of 190 texts (the average text length is 1,500 to 40,000 words). Accuracy – 83%, precision – 67% and recall –100%, F-score 0,8;
- a collection of online journalism texts (‘The Village’, URL: https://www.the-village.ru/): texts of 3 female and 3 male authors; the total of 600 texts (the average text length is 500 to 1,500 words). Accuracy, precision and recall is 100%, F-score 1;
- a collection of e-comments texts (entertainment portal ‘YaPlakal’, URL: https://www.yaplakal.com/): texts of 3 female and 3 male authors; the total of 600 texts (the average text length is 50 to 100 words). Accuracy – 40%, precision – 0 and recall – 0;
- a collection of Russian business e-correspondence texts: 2 female and 2 male authors; the total of 218 texts (the average text length is 50 to 500 words). Accuracy – 80%, precision – 67% and recall – 100%, F-score 0,8.

The algorithm solves the problem of dividing material into two clusters «the hypothesis H_0 is proved – the hypothesis H_0 is rejected», so the metrics above could be used for evaluation of the its work. The final decision about text attribution is made by the researcher and based on the analysis of statistics according to the scoring tables elaborated for each genre. To create the tables researcher compare the texts according to the principle author A = author B (texts by the same author) and according to the principle author A ≠ author B (texts by different authors). The statistics “behavior” is analyzed on this material and the scoring table (Table 1) is created:

Discourse type	Pearson correlation coefficient	Linear regression determination coefficient	Student's t-test (p-value)	The author of the compared texts is probably[3] one person	The authors of the compared texts are probably different people	Comment
Online journalism	reaches 1.00	reaches 1.00	usually around 0.95; not less than 0.93	+	−	For journalism, the p-value of the Student's t-test is a significantly less relevant metric than for other discourses. If in journalism the values of Pearson correlation coefficient and determination coefficient reach 1, we can speak of the same authorship even if p-value is not very high
	usually about 0.88 - 0.89	usually about 0.71, but it can reach and 0.77	can be either low (0.60) or quite high: 0.85	−	+	
	not very high: about 0.71	low value: about 0.50	can be very high: 0.98	−	+	

Table 1: The example of scoring table for statistics estimation

Another part of each text collection is examined by the researcher manually according to the same schemes: author A = author B; author A ≠ author B, with the aim to find True Positive (TP), False Positive (FP), False Negative (FN),True Negative (TN) results of the algorithm work. The outcome of the investigation is the table with the following view (Table 2):

Text pair		**TP**	**FN**	**FP**	**TN**
1	A. Yakovlev, Podstavnyye znakomstva… [Fake acquaintances…] – A. Yakovlev, Kak vstrechayut Novyy god… [How they celebrate the New Year…]	+	−	−	−
2	O. Karaseva, Gde deshevle zimovat'… [Where is it cheaper to spend the winter…] - O. Karaseva, Na chto zhivut zhurnalisty federal'nykh kanalov [What do federal channels journalists live on?]	+	−	−	−
3	A. Yakovlev, Luchshiye sovetskiye mozaiki v Moskve [The best Soviet mosaics in Moscow] - K. Rukov, Vyzhivut tol'ko spekulyanty… [Only speculators will survive…]	−	−	−	+
4	O. Karaseva, Kak seychas poyekhat' na dachu [The way to go out of the city now] - A. Dergacheva, Rabochiye snova opustoshayut zapasy bobrov na Yauze [The workers empty the stocks of beavers on the Yauza again]	−	−	−	+
etc.					

Table 2: Evaluation of the algorithm work

[3] The probabilistic nature of the conclusion is associated with the fact that in each specific case the final decision about the authorship is made by the researcher.

Then accuracy, precision, recall and F-score are calculated using standard formulas [14].

From a theoretical point of view, the models constructed with the help of the algorithm are clear and simple, easily interpreted, sufficiently complete and adequately simulating the original object.

From a practical side, the identification problem of authorship attribution could be solved with the elaborated software: the technique can be applied in different discourses and for various text lengths if the models are carefully parameterized and the statistics is correctly interpreted. During the work, it was found that:

1. t-test is the most informative indicator for fictional discourse (both for the discourse of famous authors and for network literature) and e-correspondence and significantly less relevant for journalistic texts;
2. to determine the author of a paper text the values of the correlation and determination coefficients must reach 1 (the need for such a high level is associated with the volume of the textual material and its specifics);
3. for Internet fiction, the stylometry pool (lengths, indices) is uninformative: according to experimental data, the values of stylometry parameters are very similar in all studied texts;
4. for short messages: e-correspondence, comments in the Internet, it is necessary to create a representative sample of at least 500 words. The 100-word limit deduced by S.M. Vul and still relevant for forensic authorship in Russia should be enlarged for the proposed method because of mathematical statistics usage in parameterized model. To improve algorithm work on this material additional parameters of the so-called digital handwriting are currently being developed: graphic hybridization, piglet Latin, language game with archaic affixes, the use of text elements written in capital letters, emoticons and other graphic symbols;

   The texts of different genres could also be examined using the integrative technique (accuracy – 80%, precision –100% and recall – 67%, F-score 0,8): journalistic text could be compared with e-correspondence, for example.

The results of algorithm evaluation could be compared with other attribution algorithms work, for example, with those based on machine learning or neural networks. Thus, the experimental result of well-known system for forensic attribution “Avtoroved” [41; 43] work on famous fiction and journalism is classification accuracy of 96.6%. “Avtoroved” uses support vector machine and logistic regression to solve the authorship problem. On the first iteration, Dovlatov’s and Astafiev’s texts are excluded. Then the texts of both authors are compared in pairs with other 20 authors. The program recognizes all the texts (from 14 ones) by V. Astafiev, except “Zatesi”, “Posledniy poklon” [The Last Tribute], “Lovlya peskarei v Gruzii” [The Catching of Gudgeons in Georgia] due to the fact that these texts have a wide variation in length. “Avtoroved” also correctly identifies all the texts (from 12 ones) by S. Dovlatov, except “Ariel”, “Zapiski nadziratelya” [A Prison Camp Guard's Story], “Solo na undervude” [Solo on Underwood], “Kompromis” [The Compromise]. To increase the accuracy of the algorithm these 7 texts are sampled into several shorter parts. After that for 5 previously unauthorized texts it is possible to identify the author. The remaining 2 texts allow to obtain the correct result in comparison with 16 authors out of 20. Thus, we could see high level of results produced be “Avtoroved”. Otherwise, we also could find out that the algorithm is sensitive to the length of the texts. “KhoRom” is less sensitive to text volume (the exception is extremely short texts). The difference in volumes, which can affect the result, can be neutralized through correct selection of parameters and relevant interpretation of the resulting statistics.

Nevertheless, any comparison of proposed methodology with algorithms based on machine learning or neural networks is rather irrelevant because “KhoRom” is built under the principle, which differs from fully automatic ones. It is based on integrative model which needs interpretation by the researcher for making a final conclusion about the authorship.

4 Conclusion

The attribution algorithm based on integration of statistically objectified interpretative methods is rather effective. The main feature of the algorithm and created linguistic resource is the interpretability of the obtained mathematical models. The results could be understood even by the users with no professional knowledge, because the outcome models are intuitive, and the resource interface is simple.

The functionality of this resource is aimed to solve an identification problem of text attribution for «closed class» [16] with pair-wise comparison, but it is much wider than the initial capabilities. The resource can be used for solving diagnostic attributional problems (gender, age, etc. designation), and working under writers, journalists, etc. language personality description by forensic experts, philologists and cultural critics. Anyway, the model of a language personality will meet the principles of completeness, simplicity, adequacy, technically accurate and objective description of the original, it will be explanatory, communicative and interpretable.

Acknowledgements

The reported study was funded by RFBR, project number 19-312-90022.

References

[1] Bacciu A., Morgia M. La, Mei A., Nemmi E. N., Neri V., Stefa J. CrossDomain Authorship Attribution Combining Instance-Based and Profile-Based Features // Notebook for PAN at CLEF 2019. — Lugano, Switzerland, 2019. Access mode: http://ceur-ws.org/Vol2380/paper_220.pdf.

[2] Bhargava M., Mehndiratta P., Asawa K. Stylometric Analysis for Authorship Attribution on Twitter Author's // International Conference on Big Data Analytics, 2013. Access mode: https://www.researchgate.net/publication/299669552.

[3] Bessmertny I.A., Nugumanova A.B. The method of automatic thesaurus construction on the basis of statistical texts processing [Metod avtomaticheskogo postroenija tezaurusov na osnove statisticheskoj obrabotki tekstov na estestvennom jazyke]. — Bulletin of the Tomsk Polytechnic University [Izvestija tomskogo politehnicheskogo universiteta], 2012. — No. 5, pp. 125-130.

[4] Burrows J. 'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship. — Literary and Linguistic Computing, 2002. — Vol. 17, Issue 3, pp. 267–287. Access mode: https://doi.org/10.1093/llc/17.3.267.

[5] Campbell L. The Sophisties and Polilicus of Plato, 1867. — Oxford, Clarendon, 170 p.

[6] Coulthard M. Author identification, idiolect, and linguistic uniqueness. — Applied Linguistics, 2004. — No 24(4), pp. 431–447.

[7] Custódio J. E., Paraboni I. An Ensemble Approach to Cross-Domain Authorship Attribution // Experimental IR Meets Multilinguality, Multimodality, and Interaction: Notebook for PAN at CLEF 2018, 2018, pp. 201-212. Access mode: http://ceur-ws. org/Vol-2125/paper_76.pdf.

[8] Ionova S.V., Ogorelkov I.V. Personality speech diagnostics in author identification based on gender parameter: quantitative approach [Rechevaya diagnostika lichnosti po gendernomu priznaku v avtorovedenii: kvantitativnyy podkhod]. — Bulletin of Volgograd State University [Vestnik Volgogradskogo gosudarstvennogo universiteta]. Series 2, Linguistics, 2020. — T. 19, No. 1, pp. 115–127. Access mode: https://doi.org/10.15688/jvolsu2.2020.1.10.

[9] Federal Law of May 31, 2001 N 73-FL «On State Forensic Science Activities in the Russian Federation» [Federal'nyj zakon ot 31 maya 2001 g. N 73-FZ «O gosudarstvennoj sudebno-ekspertnoj deyatel'nosti v Rossijskoj Federacii»]. — Russian newspaper [Rossiyskaya Gazeta]. — N 256 of December 31, 2001. Access mode: https://base.garant.ru/12123142/.

[10] Galyashina E.I. Osnovy sudebnogo rechevedeniya. [Basics of judicial speech]. — Moscow, 2003. 236 p.

[11] Golovin B. N. Yazyk i statistika [Language and statistics]. – Moscow. Education, 1970. 190 p.

[12] Gomzin A., Laguta, A., Stroev, V., Turdakov, D. (2018), Detection of author's educational level and age based on comments analysis // Paper presented at Dialogue 2018. – Moscow, Russia, 2018. Access mode: http://www.dialog-21.ru/media/4279/gomzin_turdakov.pdf.

[13] Goroshko E.I. Forensic classification examination: problems of establishing the gender of the document author [Sudebno-avtorovedcheskaya klassifikacionnaya ekspertiza: problemy ustanovleniya pola avtora dokumenta]. Theory and practice of forensics examination [Teoriya i praktika sudebnoj ekspertizy i kriminalistiki]. – Har'kov, Pravo, 2003. – No. 3, pp. 221-226.

[14] Goutte C., Gaussier E. A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation // Proceedings of the 27th European conference on Advances in Information Retrieval Research. – Springer-Verlag Berlin Heidelberg, 2005. DOI:10.1007/978-3-540-31865-1_25.

[15] Juola P., Sofko J., Brennan P. A Prototype for Authorship Attribution Studies. – Literary and Linguistic Computing, 2006. – No. 21, Issue 2, 1, pp. 169–178. Access mode: https://doi.org/10.1093/llc/fql0.

[16] Juola P. et al. Authorship attribution. – Found Trends Inf Retr 1(3), 2008, pp. 230 – 333.

[17] Karaulov Yu. N. The Russian Language and the Language Personality [Russkij yazyk i yazykovaya lichnost']. – Moscow, Nauka, 1987, pp. 200 – 264.

[18] Komissarov A.Yu. Forensic study of written language [Kriminalisticheskoe issledovanie pis'mennoj rechi: ucheb. Posobie]. – Moscow, Ministry of Internal Affairs of Russia, 2000. – 126 p.

[19] Koppel M., Schler J. Exploiting Stylistic Idiosyncrasies for Authorship Attribution. Proceedings of // IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis. – 2003, No. 69, pp. 72-80.

[20] Korobov M. (2015), Morphological analyzer and generator for Russian and Ukrainian languages. AIST – Springer, Cham, 2015. CCIS, Vol. 542, pp. 320–332. Access mode: https://doi.org/10.1007/978-3-319-26123-2_31.

[21] Labbe C., Labbe D. Inter-Textual Distance and Authorship Attribution. Corneille and Molièr. – Journal of Quantitative Linguistics, Taylor & Francis (Routledge). – 2001, No. 8 (3), pp.213-23.

[22] Linguistics of constructions [Lingvistika konstruktsiy]. Ed. E. V. Rakhilina. – Moscow, Publishing center Azbukovnik, 2010. – 584 p.

[23] Litvinova T., Sboev A., Panicheva P. (2018). Profiling the Age of Russian Bloggers // Proceedings of the 7th International Conference, AINL 2018. – St. Petersburg, Russia 2018, pp. 167–177.

[24] Litvinova T., Seredin P., & Litvinova O. Using part-of-speech sequences frequencies in a text to predict author personality: a corpus study. – Indian Journal of Science and Technology, 2015 – 8, 93.

[25] Litvinova T., Seredin P., Litvinova O., Zagorovskaya O. Gender identification in Russian written texts. – XLinguae, 2017. – No. 10 Issue 3, pp. 176-183. Access mode: 10.18355/XL.2017.10.03.14.

[26] Lutoslawski W. The original and growth of Plato's logic. – London, 1897. – 613 p.

[27] Martynenko G. Ya. (2015). Stylometry: emergence and formation in the context of interdisciplinary interaction [Stilemetriya: vozniknovenie i stanovlenie v kontekste mezhdisciplinarnogo vzaimodejstviya]. Structural and applied linguistics [Strukturnaya i prikladnaya lingvistika], No. 11 // Intercollegiate Compendium. – St. Petersburg, St. Petersburg University, pp. 9 – 28.

[28] Marusenko M. A. (1990). Attribution of anonymous and pseudonymous literary works by pattern recognition methods [Atribuciya anonimnyh i psevdonimnyh literaturnyh proizvedenij metodami raspoznavaniya obrazov]. – Leningrad, Leningrad State University. – 164 p.

[29] McMenamin G.R. Forensic Linguistics: advances in forensic stylistics. – 2002. – 331 p.

[30] Mendenhall T. The characteristic curves of composition. – Science, 1887. – No. 9, pp. 237-249.

[31] Merriam T. An Application of Authorship Attribution by Intertextual Distance in English. – Corpus, 2003. – No. 2, pp. 142–168.

[32] Morozov N. A. (1916). Linguistic Specters: a means for distinguishing of plagiarism and originalal works for famous authots [Lingvisticheskie spektry: sredstvo dlya otlicheniya plagiatov ot istin. proizvedeniy togo ili dr. izvestnogo avt.]. Petrograd, Type of Imp. Acad. Sciences, 42 p. Access mode: http://www.textology.ru/library/book.aspx?bookId=1&textId=3.

[33] Mosteller F., Wallace D. L. Applied Bayesian and Classical Inference: The Case of the Federalist Papers. – Addison-Wesley, Reading, MA, 1984.

[34] Murauer B., Tschuggnall M., Specht G. Dynamic Parameter Search for Cross-Domain Authorship Attribution // Notebook for PAN at CLEF 2018. – 2018. Access mode: http://ceur-ws.org/Vol-2125/paper_84.pdf.

[35] Muttenthaler L., Lucas G., Amann J. (2019). Authorship Attribution in Fan-Fictional Texts given variable length Character and Word N-Grams // Notebook for PAN at CLEF 2019. – 2019. Access mode: http://ceur-ws.org/Vol-2380/paper_49.pdf.

[36] Order of December 27, 2012 N 237 «On approval of the List of types of forensic examinations performed in federal budgetary forensic institutions of the Ministry of Justice of Russia, and the List of expert specialties for which the right to independently conduct forensic examinations in federal budgetary judicial expert institutions of the Ministry of Justice of Russia» [Prikaz ot 27 dekabrya 2012 goda N 237 «Ob utverzhdenii Perechnya rodov (vidov) sudebnyh ekspertiz, vypolnyaemyh v federal'nyh byudzhetnyh sudebno-ekspertnyh uchrezhdeniyah Minyusta Rossii, i Perechnya ekspertnyh special'nostej, po kotorym predstavlyaetsya pravo samostoyatel'nogo proizvodstva sudebnyh ekspertiz v federal'nyh byudzhetnyh sudebno-ekspertnyh uchrezhdeniyah Minyusta Rossii»] (as amended on September 13, 2018). Access mode: www.pravo.gov.ru.

[37] Paducheva E.V. (1974). About syntax semantics [O semantike sintaksisa]. – Moscow, Nauka. – 291 p.

[38] Panicheva P., Mirzagitova A., Ledovaya Y. (2018). Semantic feature aggregation for gender identification in Russian Facebook // AINL 2017. CCISю – Springer, Cham, 2017. – Vol. 789, pp. 3–15. Access mode: https://doi.org/10.1007/978-3-319-71746-3_1.

[39] Pimonova E., Durandin O., Malafeev A. (2020). Doc2vec or better interpretability? A method study for authorship attribution // Paper presented at Dialogue 2020. – Moscow, Russia, 2020. Access mode: DOI: 10.28995/2075-7182-2020-19-606-614.

[40] Radbil' T. B., Markina M. V. Probabilistic-Statistical Models in Conducting Authoring Expertise of Russian Texts [Veroyatnostno-statisticheskie modeli v proizvodstve avtorovedcheskoj ekspertizy russkoyazychnyh tekstov]. – Political linguistics [Politicheskaya lingvistika], 2019. – Vol. 2 (74), pp. 156-166.

[41] Romanov A., Kurtukova A., Shelupanov A., Fedotova A., Goncharov V. Authorship Identification of a Russian-Language Text Using Support Vector Machine and Deep Neural Networks. MDPI Future Internet. 2021, no.1. Access mode: https://www.mdpi.com/1999-5903/13/1/3/htm..

[42] Rodionova E.S. (2008). Linguistic methods of attribution and dating of literary works (To the problem of «Moliere - Cornel») [Lingvisticheskie metody atribucii i datirovki literaturnyx proizvedenie (K probleme «Mol'er - Kornel'»)]. The thesis abstract for the degree of Candidate of Philology [Avtoreferat dissertacii na soiskanie stepeni kandidata filologicheskix nauk]. Access mode: http://epir.ru/pragmat!/projects/corneille/files/autoreferat.pdf.

[43] Romanov A.S., Kurtukova A.V.; Sobolev A.A., Shelupanov A.A., Fedotova A.M. Determining the Age of the Author of the Text Based on Deep Neural Network Models. MDPI Information. 2020, no. 12. Access mode: https://www.mdpi.com/2078-2489/11/12/589/htm.

[44] Rubtsova I.I., Ermolaeva E.I., Bezrukova A.I., Ogorelkov I.V., Zakharov M.P. (2007). Integrated methodology for the production forensic authorship examinations: Methodological recommendations [Kompleksnaya metodika proizvodstva avtorovedcheskih ekspertiz: Metodicheskie rekomendacii]. – Moscow, Ministry of Internal Affairs of Russia, 2007. – 192 p.

[45] Russian grammar [Russkaja grammatika]. (2005). – Moscow: The V.V. Vinogradov Russian Language Institute of the Russian Academy of Sciences, 2005. Access mode: http://rusgram.narod.ru/index.html.

[46] Shuy R. W. Creating language crimes: How law enforcement uses (and misuses) language. – New York: Oxford University Press, 2005. – 194 p.

[47] Solnyshkins M., Guryanov I., Gafiyatova E., Varlamova E. (2018).. Readability metrics: the case of Russian educational texts // Proceedings of ADVED 2018- 4th International Conference on Advances in Education and Social Sciences. – Istanbul, Turkey, 2018.

[48] Stepanenko A.A. (2017). Gender attribution in social network communication: the statistical analysis of pronouns frequency [Gendernaya atribuciya tekstov komp'yuternoj kommunikacii: statisticheskij analiz ispol'zovaniya mestoimenij], Tomsk State University Journal [Vestnik Tomskogo gosudarstvennogo universiteta], 2017 – No. 415, pp. 17–25. Access mode: DOI: 10.17223/15617793/415/3.

[49] Vul S.M. (2007). Forensic Attribution Identification Examination: Methodological Basics: Methodological Manual [Sudebno-avtorovedcheskaya identifikacionnaya ekspertiza: metodicheskie osnovy: Metodicheskoe posobie]. – Kharkov, Kharkov Scientific Research Institute of Forensic Expertise, 2007. – 64 p.

[50] Wright D. Using word n-grams to identify authors and idiolects: A corpus approach to a forensic linguistic problem. – International Journal of Corpus Linguistics, 2017. – No. 22(2), pp. 212–241. Access mode: https://core.ac.uk/download/pdf/84587040.pdf.