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
Аннотация

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.Turadkov [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 stylistmetric features, even in a large number of them (M. Bhargava, P. Mehdinratta, 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 au-
Linguistic Modeling as a Basis for Creating Authorship Attribution Software

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 ‘us/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, prae / 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**
   1.1. **Preprocessing**
   1.1.1. Sentence-splitting
   1.1.2. Tokenization
   1.1.3. Morphological parsing
   2. **Processing**
   2.1. **Stylometry block**
   2.1.1. Calculation of basic metrics (number of words, sentences)
   2.1.2. Search for traditional stylometric textual data (n-grams, indices)
   2.2. **Cognitive block**
   2.2.1. Search for parameters by preset rules
   2.2.2. Assigning weight to each parameter
3. **Building mathematical models of the texts being compared: attribute presentation as a sequence of numerical features**
   3.1. 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.
4. **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, Vue, 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].
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).

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.
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 H0 is proved – the hypothesis H0 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:
Table 1: The example of scoring table for statistics estimation

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 19-31-27001.

References


