Legal Issues in the Digital Age. 2022. Vol.3. No 2.
Вопросы права в цифровую эпоху. 2022. Т. 3. № 2.

Research article
УДК 343.98.065
DOI:10.17323/2713-2749.2022.2.90.115

Automation of Forensic Authorship Attribution: Problems and Prospects

Tatiana Vladimirovna Romanova[1], Anna Yurievna Khomenko[2]

[1] Department of Humanities, National Research University Higher School of Economics, 25/12 Bolshaya Pechyorskaya Str., Nizhny Novgorod 603155, Russia, tvromanova@hse.ru, ORCID: 0000-0002-1833-2711

[2] Department of Humanities, National Research University Higher School of Economics, 25/12 Bolshaya Pechyorskaya Str., Nizhny Novgorod 603155, Russia, akhomenko@hse.ru, ORCID: 0000-0003-3564-6293

Abstract

The article deals with validation of an integrative attribution algorithm based on the analysis of the author's idiostyle using methods of interpretative linguistics with objectification of the available data with the help of mathematical statistics. The algorithm addresses the identification problem of the attribution. The choice of parameters describing the individual style of an author assumes that the text is a product of an authentic language personality described by psycholinguistic (Yu.N. Karaulov), sociolinguistic and forensic linguistic (S.M. Vul, M. Coulthard, R. Shuy) methods. To validate a hypothesis that the identification problem of attribution is best resolved by the integrative methodology, we have created the KhoRom application which brings together the aforementioned approaches to the analysis of language personality: http://khorom-attribution.ru/#/. It can be used to compare two language personality models and determine to what extent they are similar using the following metrics: Pearson correlation coefficient, linear regression determination coefficient and Student's t-criterion. Importantly, this application also describes the interpreted model of language personality to inform the user on the importance of values of each parameter. The system has a wealth of features, with the user able to choose parameters, view parameter implementation in the document and edit the final list of parameter implementations (in case of malfunction, the application performance can be corrected manually). The created application is only a part of the attribution algorithm. The data produced by mathematical statistics need to be analyzed by expert judgment through the use of methodological recommendations developed for the algorithm. The effectiveness of this methodology has been proved by its validation on texts of various length and genres, with a number of documents pertaining to fiction, journalism, official and colloquial styles being analyzed. For texts of all discourses except colloquial, the developed algorithm has demonstrated a high level of accuracy (F-score of 0.8 to 1). For better applicability of the algorithm to colloquial texts, the authors have developed a number of improvements pending implementation.

Keywords

attribution, language personality, automated text processing, linguistic model, mathematical model, attributive software, forensic authorship attribution.

Acknowledgments: The study was performed with financial support of the Russian Fundamental Fund of Researches as part of research project No. 19-312-90022.

For citation: Romanova T.V., Khomenko A.Yu. (2022) Automation of Forensic Authorship Attribution: Problems and Prospects. *Legal Issues in the Digital Age*, vol. 3, no. 2, pp. 90–115. DOI:10.17323/2713-2749.2022.2.90.115

1. Background

At present stage of progress in science a problem of automation of social processes has been discussed by specialists in all fields including forensic experts. "Forensic investigation means a procedural activity involving studies and opinions to be given by experts on issues which require specific knowledge in the area of science, technology, arts or crafts and which courts, judges, investigative authorities, inquiry officers, investigators or public prosecutors deal with in order to ascertain the circumstances to be proved as part of a specific case"[1]. A forensic investigation can be both criminal and noncriminal. While automated analytical tools have become customary for most criminal investigations (trace examinations, forensic genetics etc), software support is not yet available to all investigations of this kind in Russia. Thus, forensic authorship attribution is an inquiry associated with criminal investigations (classified as such by the Russian Ministry of Justice)[2], its purpose

[1] Federal Law No. 73-FZ "On State Forensic Investigations in the Russian Federation" dated 31 May 2001. Rossiyskaya Gazeta, No. 256 of 31.12.2001. Available at: URL: https://base.garant.ru/12123142/ (accessed: 03.05.2020)

[2] Order No. 237 "On Approving the List of Forensic Inquiry Types to Be Performed at Federal Offices of Forensic Services under the Ministry of Justice, and the List of Practitioners Authorized to Perform Investigations at Federal Offices of Forensic Services under the Ministry of Justice" of 27 December 2012 (as amended of 13 September 2018). Available at: URL: www.pravo.gov.ru (accessed: 03.05.2020)

being to attribute a text to a specific author (group of authors) or obtain information on individual authors. However, the extent of automation of this kind of inquiry is currently quite low. This is probably due to the fact that courts will often dismiss the requests for investigation of this kind.[3]

2. Problems and prospects of developing algorithms for automated forensic authorship attribution

2.1. Principles of authorship attribution in and outside Russia

In modern linguistics, automated analytical methods for textual attribution for purely research purposes are progressing worldwide. They are implemented as software products both in and outside Russia, the most popular still being models and algorithms based on n-gram speech recognition [Bacciu A., Morgia M., 2019]; [Litvinova T., Sboev A., Panicheva E.B., 2018: 167–169]; [Custódio J., Paraboni I., 2018]; [Murauer B., Tschuggnall M., Specht G., 2018]; [Muttenthaler L., Lucas G., Amann J., 2019], part-of-speech attribution of units [Litvinova T., Sboev A., Panicheva E.B., 2018: 177], variable length patterns [Custódio J., Paraboni I., 2018] and using cluster analysis [Panicheva P. et al., 2018], traditional [Gomzin A. et al., 2018] and modified [Korobov M., 2015: 320–332] Python libraries, vector transformation algorithms [Bacciu A., Morgia M., 2019] etc. There have been successful attempts to use linguistic models as such to determine who authored a text (based on the vector approach to analysis). As regards Russian software products, the following are worth mentioning.

M.A. Marusenko software based on the theory of image recognition. This approach to attribution of language personality could be seen in his studies [Marusenko M.A., 1990, 2003] and E.S. Rodionova [Rodionova 2008 a,b] focused on the analysis of deep text structures are best reflects the peculiarities of a person's cognitive processes. Such an approach will doubtlessly produce decent results due to the model being more complete and deductive and better reflecting the subject of study. Nevertheless, the model is extremely difficult to use and understand for anyone who doesn't have the theoretical knowledge of image recognition and mathematical statistics. The use of this model is still further complicated by the absence of

[3] The Court on Intellectual Property Rights of the Russian Federation, ruling of 4 December 2020 on case No. SIP-676/2019; The Court on Intellectual Property Rights, ruling of 29 November 2019, case No. SIP-695/2019. Appellate ruling of 26 December 2018. No 203–APU 18–25 etc. Available at: URL: https://base.garant.ru/75013773 (accessed: 03.05.2020)

a generally accessible user interface while repetition of all mathematical transformations described therein is very lengthy.

V.N. Zakharov software (Atributsia) based on the analysis of grammar and syntax [Zakharov V.N. et al., 2000] that allows to parse literary text using multiple linguistic features. The software consists of two parts: the grammatical analysis module and the syntactic analysis module. They enable to partially automate and formalize the parsing process across 69 parameters [Sidorov Ju. B. et al., 1999: 66]. However, this software requires the involvement of an expert philologist to check the correctness of part-of-speech attribution etc. V.N. Zakharov and his colleagues analyzed the works of Fyodor Dostoevsky and non-attributed texts of still disputed authorship. As a result of the experiments, this group of researchers has managed to identify certain anonymous texts as those authored by Dostoevsky and thus make them part of the classical author's literary heritage.

A.N. Timashev software (Attributor) based on letter triads [Timashev A.N., 2007]. That researcher has proposed to use three-letter combinations — triads — as a criteria to distinguish an author's style. This approach includes single-letter and twin-letter function words into the analysis as making up a "significant part of the frequently used prepositions, conjunctions, particles and interjections traditionally believed to be meaningful style defining features" [Batura T.V., 2012: 87]. The above methodology uses a text database of 103 Russian authors of 19–20th centuries. At the start, the software uses a machine learning method involving an expert linguist. To avoid the errors resulting from a comparison of statistically non-comparable objects, the text should be at least 6 pages long.

A.S. Romanov software (Avtoroved) based on the support vector machine in the form of the most frequently used trigrams and words [Romanov A.S., 2010]. The authorship problem is regarded as a classification problem to be solved using the support vector machine where the idiostyle is described with symbol trigrams and words most frequently used in Russian. The main findings were produced on a set of 215 prose texts by 50 Russian writers borrowed from M. Moshkov's e-library. For texts authored by 2/5/10 persons, the experiments showed the most informative authorship features to be those restricted to 300–700 most frequent trigrams and 500 most frequently used words. The methodology proved to be practically useful for analysis of short electronic messages (which is remarkable since dealing with short texts is extremely complicated) when the software nicknamed Avtoroved and the underlying methodology were tested at a military base. The findings showed that in case of two potential authors the authorship of

100-symbol long texts could be attributed with a maximum accuracy of 0.76 ± 0.11. A sub-problem to identify the author of a web forum message was solved with an accuracy of 0.89 ± 0.08. Thus, the said method works relatively well for short e-messages which offers high experimental potential in the context of modern electronic communications.

KAT software was produced by N.I. Lobachevsky State University, Nizhny Novgorod. This product uses a database of Russian classical texts (written by Leo Tolstoy, Nikolai Gogol, Ivan Turgenev), with models relying on an analysis of coefficients of correlation between different parts of speech (after B.N. Golovin) [Radbil T.V., Markina M.V., 2019]. The use of such coefficients is undoubtedly well-founded from a psycholinguistic and behavioral perspective offered by fundamental science since the part-of-speech association of vocabulary of an author's idiolect is clearly a distinctive feature of style. Importantly, the software uses not just a transversal coefficient of correlation between all parts of speech but conscious relationships between them.

Lingster 3.0 software by the Institute of Forensic Science under the Federal Security Service [Rubtsova I.I., Ermolayeva E.I., Bezrukova M. Yu. et al., 2007], TextAnalyst 2.0 by the Moscow Research Center [Ionova S.V., Ogorelkov I.V., 2020]; RusIdiolect database by the laboratory of corpus ideolectology, Voronezh State Pedagogical University [Litvinova T.A., Gromova A.V., 2020: 77– 88].

Due to specifics of the legal practice, the principles of forensic authorship attribution somewhat differ from those applicable to solution of research problems as such. This follows in the first place from the Russian law: Federal Law No. 73-FZ "On State Forensic Investigations in the Russian Federation" of 31 May 2001 ("Law No.73-FZ)[4] and all codes establishing procedural standards (for criminal, arbitration and civil procedures and administrative offenses)[5] provide for personal liability of experts in respect of an opinion to be given. "An expert's opinion is a written document

[4] Available at: URL: http://www.consultant.ru/document/cons_doc_LAW_31871/ (accessed: 12.06.2020)

[5] 1) Code of Criminal Procedure of Russian Federation dated 18.12.2001, Federal Law No 174-FZ(as amended on 25.03.2022 and including modifications in force from 19.05.2022). Available at: URL: http://www.consultant.ru/document/Cons_doc_law_34481/ (accessed: 24.05.2022)

2) Code of Arbitration Procedure of Russian Federation dated 24.07.2002, Federal Law No 95-FZ (as amended on 30.12.2021, as modified on 10.01.2022}. Available at: URL: http://www.consultant.ru/document/cons_doc_LAW_37800/ (accessed: 24.05.2022)

3) Code of Civil Procedure of Russian Federation dated 14.11.2002, Federal Law No 138-FZ (as amended on 16.04.2022). Available at: URL: http://www.consultant.ru/document/Cons_doc_LAW_39570/ (accessed: 24.05.2022)

reflecting the course and findings of investigations conducted by the expert [italics added. — T.R., A.Kh.]"[6]. While this liability cannot be shifted to the machine, the expert should critically analyze the findings produced by the software (if any) and issue a "well-founded and objective opinion"[7] "within the ambit of the respective qualifications, comprehensively and to the full extent"[8]. Any failure to comply with requirements of the law will incur not only moral liability before the civil society for the opinion being issued but also criminal liability before the state under Article 307 of the Criminal Code of Russia[9].

Since the expert's personal liability is established by law, this constitutes an obstacle preventing the use of fully automated technologies of attribution analysis in Russian legal practice. But this obstacle is not the only one. A specific feature of the national regulatory framework including the codes of criminal procedure, civil procedure, arbitration procedure and administrative offenses, and Federal Law No. 73-FZ (Article 8), is that the expert dealing with questions to be explored should strictly remain within the ambit of his competence as determined by the amount of his expertise: "The expert may <...> 4) provide an opinion within his competence [italics added. — T.R., A.Kh.,] including on issues relevant to the subject of expert investigation though not mentioned in the order on forensic investigation"[10]. The same idea is present in the codes of civil procedure[11], arbitration procedure[12] and administrative offenses[13].

4) Code of Administrative Offenses of Russian Federation dated 30.12.2001, Federal Law No 195-FZ (as amended on 16.04.2022 and modified on 17.05.2022, including amendments and modifications in force from 27.04.2022). Available at: URL: http://www.consultant.ru/document/cons_doc_law_34661/ (accessed: 24.05.2022)

[6] Federal Law No. 73-FZ "On State Forensic Investigations in Russia" dated 31 May 2001. Rossiyskaya Gazeta. No 256 of 31.12.2001. P.9. Available at: URL: https://base.garant.ru/12123142/ (accessed: 03.05.2020)

[7] Ibid. P.8. Available at: URL: https://base.garant.ru/12123142/ (accessed: 03.05.2020)

[8] Ibid. P.9. Available at: URL: https://base.garant.ru/12123142/ (accessed: 03.05.2020)

[9] Criminal Code of Russian Federation dated 13.06.1996, Federal Law No. 63-FZ. Available at: URL: http://www.consultant.ru/document/cons_doc_LAW_10699/ (accessed: 03.05.2020)

[10] Code of Criminal Procedure of Russian Federation dated 18.12.2001, No 174-FZ. Available at: URL: http://www.consultant.ru/document/cons_doc_LAW_34481/ (accessed: 03.05.2020)

[11] Code of Civil Procedure of Russian Federation dated 14.11.2002, No 138-FZ. Available at: URL: http://www.consultant.ru/document/cons_doc_LAW_39570/ (accessed: 03.05.2020)

[12] Code of Arbitration Procedure of Russian Federation dated 24.07.2002, No 95-FZ. Available at: URL: http://www.consultant.ru/document/cons_doc_LAW_37800/ (accessed: 03.05.2020)

[13] Code of Administrative Offenses of Russian Federation dated 30.12.2001, No 195-FZ. Available at: URL: http://www.consultant.ru/document/cons_doc_LAW_34661/ (accessed: 03.05.2020)

"An expert's professional competence (from Latin competo — achieve, fit, correspond) assumes a set of theoretical, methodological and practical knowledge of expert investigation of a particular kind and type"[14] . The experts performing forensic authorship attribution will normally have basic linguistic or philological education and subject-specific retraining on investigation of speech language activity products and/or (preferably) investigation of written speech for attribution of authorship (in accordance with the Ministry of Justice classification)[15]. This background does not assume expertise in the field of big data, probability theory, machine learning and neural networks, mathematical statistics, image recognition theory, vector theory etc., as disciplines required to master and understand the software relying on the best performing algorithms for automatic identification of authors of written documents. Hence, the Russian Federation law on forensic investigation fundamentally (via provisions enshrined in the codes of procedure, federal laws, departmental instructions and orders) restricts the use of purely computer technologies in authorship attribution investigations, so that experts cannot rely on software alone to draw a conclusion as, for example, in the case of genetic investigation. Naturally, experts cannot use the software based on the principles they don't understand for lack of special knowledge of statistics, mathematics, probability theory etc.

Apart from the law, the use of automated technologies to identify the author of a text is restricted by virtue of the national scientific tradition related to a wide dissemination of the interpretative research paradigm in philology in general and in forensic linguistics in particular. Thus, forensic attribution methodologies proceed from the ideas proposed by S.M. Vul [Vul S.M., 2007] and further elaborated by A.Yu. Komissarov [Komissarov A.Yu., 2000]; E.I. Goroshko [Goroshko E.I., 2003: 221–226]; E.I. Galiashina and E. I. Ermolova [Galashina E.I., Ermolova E. I., 2005: 20–22]. They are based on the theory of distinctive style shaped by a certain social environment and cognitive processes unique for each person. The work under the title Comprehensive Methodology of Authorship Attribution [Rubtsova I.I., Ermolayeva E.I., Bezrukova A.I. et al., 2007] is currently one of the relevant institutional methodologies.

[14] Encyclopedia of Forensic Investigations. Moscow, 1999. P. 177.

[15] Order No. 237 "On Approving the List of Types of Forensic Investigations to be Performed at Federal Offices of Forensic Services under the Ministry of Justice, and the List of Practitioners Authorized to Perform Investigations at Federal Offices of Forensic Services under the Ministry of Justice" dated 27 December 2012 (as amended of 13 September 2018). Available at: URL: www.pravo.gov.ru (accessed: 03.05.2020)

The practice of automatic text attribution in Russia is currently borrowed from the West European and North American schools of thought where authorship identification has been traditionally — from L. Campbell [Campbell L., 1867] down to modern day [Koppel M., Schler G., 2003: 72–80]; [Wright D., 2007: 212–241] etc. — related to methodologies of computational stylometry. Meanwhile, these schools have a tradition similar to that existing in Russia, that is, the use of properly linguistic, qualitative text attribution techniques/methodologies [McMenamin G., 2002], with forensic authorship identification practices relying on the idiolect theory [Coulthard M., 2004: 447]. In the Western tradition, idiolect has always been perceived as a construct which represents "not merely what a speaker says at one time: it is everything that he could say in a given language" [Bloch B., 1948: 3–46]. For an English speaker, a major parameter defining the idiolect is the speaker's social status. The language style is linked to linguistic variability that follows from social context. A language style offers two types of choice: variation within or deviation from the established norm. A change within the limits of a norm assumes a choice of grammatically acceptable ("correct") forms (twenty-six/ twenty six/26) while a deviation from the norm assumes a choice that covers grammatically wrong or inacceptable ("incorrect") forms (I might go/I could go/I might could go/I might could did go). A norm can be described in terms of both linguistics and statistics. Linguistic norms assumed in the use and perception of a language are described in detail in dictionaries and grammar books. Statistical norms are those that reflect the linguistic norm in the form of a certain frequency distribution of each form within the population of particular native speakers [McMenamin G., 2002].

Courts in certain parts of the USA and the UK (once a permission in respect of a particular case is given) will accept attribution investigations of quantitative content [Juola P., 2006: 233-334] involving the use of a software. A number of examples could be cited: Court of Appeal, London, 1991: the Queen vs. Thomas McCrossen; Leicester Crown Court, 1992: the Queen vs. Frank Beck. However, the use of fully automated investigations for forensic attribution in the West is an exception rather than rule. In Russia, as was noted above, this practice is altogether absent. Overall, courts in Russia will not often order an investigation to attribute authorship of a text. Authorship attribution investigations are frequent in respect of music and art[16] and

[16] The Court on Intellectual Property Rights, ruling of 4 March 2019 on case No. A63-22578/2017; The Court on Intellectual Property Rights, ruling of 18 June 2019 on case No. A40-224162/2017; The Court on Intellectual Property Rights, ruling of 13 January 2020 on case No. A57-15203/2018, etc.

much less so in respect of texts[17]. In criminal investigations, text attribution is ordered more frequently[18]; however, given the complex matters to be explored and the probability of making wrong conclusions in the absence of knowledge necessary for their assessment, we believe this happens less often than required.

In the English-language forensic linguistics, the principal event of automatic text processing to identify authorship and other individual features of a language personality is apparently a series of PAN events of the Conference and Labs of the Evaluation Forum or Cross-Language Evaluation Forum[19] in which researchers from Russia — such as Tatiana Litvinova of Rus Profiling Lab [Litvinova T.A. et al., 2017: 1–7] — are also involved. It is worth noting, however, that Rus Profiling Lab is virtually the only organization in Russia engaged on a permanent, professional basis in developing open-source, publicly available automatic attribution algorithms for Russian-language texts including for forensic purposes. A.S. Romanov and his team from the Tomsk State University of Control Systems and Radioelectronics [Romanov A.S. et al., 2021: 1–16] are currently working on improvements for the already available Avtoroved software in the interest of high-security institutions.

Despite the strongly prominent tradition of interpretative linguistics at both Russian-language and English-language forensic attribution schools, the preference for qualitative methods owes itself not so much to persistence of traditions in this branch of linguistics as to the law which makes experts personally liable for their opinions (in and outside Russia) before the civil society and the state. Importantly, no validated and commonly recommended methodology of automatic (computer-assisted) attribution analysis based only on statistics retrieved from the text is now available on a full scale either in Russia or elsewhere. The reason is the complexity of texts to be analyzed which may largely differ in terms of length, functional style, metadata affecting their structure etc. At this stage, given a lack of

[17] Determination of 20 July 2020 on case No. SIP-250/2017 to suspend proceedings and conduct an investigation.

[18] Order of 05 September 2018 by R.R. Saifetdinov, investigator of criminal investigation unit No. 6, Sverdlovsk Oblast office, Ministry of Interior, under criminal case No. 11801650081000303; order of 15 June 2018 by E.A. Nikiforova, senior investigator of the investigation unit, Noyabrsk office, Ministry of Interior, under criminal case No. 11701711492002633; order of 22 February 2017 by F.V. Tyutnev, senior investigator of the investigation unit, Volga Federal District office, Ministry of Interior, under criminal case No. 11701000150103930 etc.

[19] Available at: https://pan.webis.de. (accessed: 10.05.2022)

a shared, generally accepted and commonly recommended automatic research algorithm for attribution of texts and the current legal provisions in Russia, experts cannot apply strictly statistical methods, unless they are supported by interpretative approaches.

2.2. The prospects of forensic authorship attribution in Russia

Due to peculiarities of the Russian regulatory framework which provides for experts' personal liability before the state for the judgment they make, inadequate software implementation of automatic attribution algorithms with the resulting low accuracy for forensic purposes, and the strong tradition of interpretative linguistics, on the one hand, and imminent digitization of all spheres of social life, on the other hand, the only way forward for forensic attribution in Russia is, in our view, the integration of computer-assisted methodologies of quantitative text analysis with interpretative qualitative investigations performed by experts in a single software package. Obviously, there have been efforts to do that [Baranov A.N., 2001]; [Ionova S.V., Ogorelkov I.V., 2020: 115–127], and it is logical to move on.

The main purpose of this study is to develop an integrative text attribution methodology including formalization of language personality attribution models in order to make the algorithm adaptable to: a) computer-assisted implementation; b) wide range of linguists including forensic experts. The study is expected to result in an operational algorithm prototype for automatic/semi-automatic identification of authors of written texts.

2.3. Integrative attribution software

At the moment, the authors have tested a prototype methodology with the said parameters where the interpretative linguistic methods identify the information on the author's competences in the traditional sense (thesaurus and pragmaticon of a language personality, levels of mastering written speech competencies) while the stylystatistics allows to add objectivity to the findings of interpretative analysis. The KhoRom attribution resource prototype is available in the Internet[20].

The prototype solves the identification problem of attribution linguistics of the "sample comparison" type where one or more texts of unknown authorship and a sample text of known authorship are available. The method-

[20] Available at: URL: http://khorom-attribution.ru/#/ (accessed: 24.04.2022)

ology was tested on authorized texts to check its functional capability and ensure successful application as a forensic tool.

The proposed methodology implements the following algorithm. It will: automatically retrieve parameters describing the author's pragmaticon, thesaurus and lexicon; search for traditional stylometric data (text statistics data); assign a weight to each parameter; construct mathematical models of the compared texts; compare the mathematical models; perform expert analysis of statistical data. Importantly, this is not the authentic way to automatically attribute authorship but an integrative methodological concept bridging two approaches to objectify the interpretation with statistics followed by analysis of statistical data.

The formalization of multi-level structure of a language personality is based on the postulates of Yu. N. Karaulov's theory [Karaulov Yu. N., 2010] where a language personality is understood as a set of communicative skills (ability to produce oral speech and written texts, level of verbal communication culture, ability to achieve the purpose of communication etc.) acquired by the individual in a certain social environment during the period of development. In fact, the formalization process follows the principles of semantic syntax [Paducheva E.B., 1974] and Russian grammar rules[21].

The structure of language personality is regarded as a combination of three levels: verbal semantic, linguo-cognitive and motivational [Karaulov Yu.N., 2010].

A language personality is understood as a result of development in a certain social environment based on autobiographic, sociolinguistic and juridical linguistic approaches [Vinogradov V.V., 1961]; [Coulthard M., 2004: 431]; [Shuy R., 2005]; [Vul S.M., 2007].

Based on empirical study of 10 text fragments totaling 116 thousand words we have identified a number of language personality parameters that are invariably important as components of individual style, original authentic language, explicit feature of the author's language personality and at the same time are automatically retrievable from the text with minimum pre-processing required. For computer-assisted retrieval, all formal rules were programmed and incorporated into the KhoRom linguistic resource: http://khorom-attribution.ru/#/.

As a result of empirical study, the search parameters such as attribution of words to different parts of speech (number of content words, ratio

[21] The Russian Grammar. Available at: URL: htpp://rusgram.narod.ru/index.html (accessed: 16.11.2020)

of different parts of speech — legibility index, objectness coefficient etc.), average word lengths, presence/absence of compound hyphenated words, modal particles, interjections, presence/absence of "-to" modal postfix, preferable intensifiers were programmed at the verbal semantic level. The formalized search of units at this level is carried out in accordance with the text's morphological profile, that is, by tagging each word as a part of speech and all grammatical categories associated with the given part of speech. For instance, a search of elements with "-to" modal postfix will follow this algorithm:

1) + Prnt-to

2) — SPRO, nom / gen / dat / acc/ ins / loc / voc / gen2 / acc2 / loc2, sin / pl

3) — APRO, nom / gen / dat / acc/ ins / loc / voc / gen2 / acc2 / loc2, sin / pl[22].

Thus, the diagram can be read as follows: the search is for any part of speech with "-to" modal postfix (except pronouns and adjective pronouns) in any case of singular or plural.

Intensifiers are understood as words used to identify the extent of semantic category of intensity. These are mostly adverbs whose range is limited albeit great (in the modern discourse — ochen, silno, adski [very, strongly, damned]). But the category of intensity is not limited to exclusively adverbial content, for example: Kakaya krasota! [What a beauty!]. In this case, it is the pronoun kakaya that serves as an intensifier. Thus, a code of rules was developed as part of the study to search for structures with intensifiers; the list of intensifiers includes both adverbs with certain grammatical limitations (structures where the adverb does not express the category of intensity: for instance, it makes part of a compound nominal predicate, such as in On chuvstvuyet sebya khorosho [He feels good] and certain adjectives and pronouns in relevant grammatical structures such as: A "nastoyaschy", nom / acc, sin / pl + N: nastoyaschy bardak [real mess].

Regarding the search for parameters of the verbal semantic level, a total of 107 authentic rules were developed to identify 11 different structures in the text. The search for chosen parameters at this level, that of idiolect in accordance with the concept, is easy to formalize since the verbal semantic level has "more formal language features a priori believed to be stable

[22] Hereinafter the designations corresponding to part-of-speech tagging of the Russian National Corpus are used. Available at: URL: https://ruscorpora.ru/new/corpora-morph.html (accessed: 24.05.2022); «/» — or, «+» — presence of several elements in the structure; A — adjective, N — noun;

though the issue of their stability has not been specifically explored" [Litvinova T.A., 2019: 2].

To represent a fragment of personal thesaurus, we have chosen parameters such as key lexemes, frequently used word trigrams and bigrams, and explicators of axiological text dominants of the friend-foe dichotomy.

The key lexemes are identified using the logarithmic plausibility algorithm as the text of interest is compared to a large reference database (Opencorpora was used, URL: http://opencorpora.org, accessed 08.02.2020, 1,540,034 words as of the access date). As a result, a list of key words with numerical explication of the measure of logarithmic plausibility (loglikelihood score or LL) is generated for each text. The final list has only the words with LL value higher than 50.

A search for word bigrams and trigrams is based on the absolute frequency of finding words next to each other and is implemented using the functions of the chosen programming language. The most frequent word combinations for the texts in question are identified after the above preprocessing. The calculation also takes into account whether a given word is not in the list of stop words, words spelled in Cyrillic and those longer than 2 symbols. As a result of comparing two texts, a list of the most frequent word combinations is generated for each.

In analyzing key lexemes and most frequent word combinations, those with proper names are deleted from the resulting lists since these lexemes identify the thematic association of text rather than features of the author's idiostyle.

In this study, explicators of axiological text dominants of friend-foe groups are understood as the dispersion of pronouns of the I-we and you-they groups — that is, all classes of pronouns in direct and indirect cases are calculated across relevant groups [Stepanenko A.A., 2017: 17–25].

The thesaurus level is the hardest to formalize. While it is possible to create physical explication of the author's thesaurus [Bessmertny I.A., Nugumanova A.B., 2012: 125–130], it is still very difficult to identify how its lexemes "form up an orderly, fairly strict hierarchical system which reflects to some (indirect) extent the world's structure" [Karaulov Yu. N., 2010: 52]. This level is represented by the least number of parameters (three standard stylometric algorithms and one authentic rule) since the idea is not simply to formalize certain language personality elements for computer representation but also to make the resulting model interpretable.

A language personality's pragmaticon (a set of strategies and tactics, as well as means of their implementation that serve to achieve a speaker's communicative purposes during communication) is formalized by the following set of parameters: parenthetic words and constructions expressing the subjective modality; purposive, intensifying and comparative locutions representing to what extent the author has mastered the written speech competencies and associated communicative strategies and tactics; syntactic clusters which give an idea, in particular, on the author's preferences regarding functional and stylistic association of the text; comparative, subordinate, one-member verb sentences expressing the functional type of narration; presence/absence and types of address as a contact establishing element. A total of 10 standard stylometric (searching for text statistics) algorithms and 32 unique rules were used.

It is not the pragmaticon units themselves ("communicative environment: domains, situations, roles" [Karaulov Yu. N., 2010: 61]) but indirect representatives of these units, components of the syntactic level that are assigned for the said level in the model. Therefore, in particular, the developed algorithm is not implementable without as an expert's judgment. That is, the author's competencies and aptitudes should be reproduced at the pragmatic level from the resulting statistical/syntactic information through interpretation. Let's take Sergei Dovlatov's collected stories "Nashi" to illustrate this process. Using the KhoRom software, we can extract 171 parenthetical constructions, a vast majority of which are conjunctive parenthetical constructions (krome togo, bolee togo, znachit etc. [except, moreover, hence] that create anaphoric linkages in the text. Thus, Dovlatov implements a competency of producing a coherent text, "aptitude of associating intentions, motives, planned meanings with the ways of their objectivation in the text". The identified value of parameters also allows to assert that the emotional charge of the speech ("aptitude of using stylistic means of this or another sublanguage") is largely produced by constructions different from parenthetical elements. The imagery becomes a major technique to create emotion in the text as proved by a comparison of syntactic complicators: the text has much more comparative than purposive phrases, their relative frequency of occurrence being 2669.85 against 715.14.

To analyze the syntactic structures, we introduced the rules based on POS tagging and on the types of syntactic relations found in the sentence [Paducheva E.B., 1974] and grammatical constructions implemented by its components. For instance, to identify parenthetical words, the formalized rule (search algorithm) will look as follows:

a vocabulary of all possible parenthetical words in Russian is created for computer-assisted representation;

a grammatical punctuation rule is assigned to identify parenthetical constructions rather than those homonymous to them:

1) __, Prnt,__

2) <start of sentence> Prnt,

where Prnt is any part of speech; __ — some part of the sentence while <start of sentence > marks the beginning of the sentence.

A search for one-member verb sentences — for example, definite personal ones — follows this algorithm:

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

+ V, sg / pl, imper

3) — N / SPRO, nom, sg / pl

4) — NUM, nomn _+ N в gen/ gen2, pl

5) — many/few/several/some/considerable _ + N in gen/ gen2, pl.

The rule to search for purposive constructions is based on the semantic slot concept [Paducheva E.B., 1974: 44] and the grammar of prepositional constructions with double prepositions. Compound prepositions such as s tselyu/iz rascheta [for the purpose of/with a view to] will require an infinitive (as semantic slot condition) to have a purposive phrase, so the formalized rule to search for such constructions will look as follows: s tselyu/iz rascheta + INF where INF designates an infinitive.

Once all word structure-related parameters are retrieved, the ipm (instance per million) calculation is carried out. For syntactic parameters, the number of each parameter is divided by the number of sentences in the text. Designing a rule for automatic search of structures of the verbal semantic and motivational levels (those chosen for this study) is relatively simple. The resulting accuracy is high, with F-measure for all parameters varying from 0.89 to 1.

The output delivered by the algorithm are values of the Pearson correlation coefficient, linear regression (where determination coefficient should be assessed), Student's t-criterion for models of both compared texts, as well as the metrics of each parameter of the two texts to prove or refute H_0 hypothesis that both were authored by the same person.

Importantly, this module is not the final step in the developed methodology. As was said before, the text statistics need to be interpreted. Whereas a correlation coefficient of more than 65 percent is believed to be significant for the traditional mathematical statistics, it should be more than 86 percent for a software before we can assume the models are similar

[Radbil T.B., Markina M.V., 2019]. It is on purpose that the software does not generate the result in the form two compared texts are authored by the same person/two compared texts are authored by different persons since under the developed methodology the final attribution decision is to be made by an expert based, in particular, on statistical data (using checklist tables that was created on the basis of research findings, see Table 1) and his own investigative experience.

To construct such tables, the authors used text collections (see paragraph 3 of this paper for description), with 40 percent of texts in each analyzed through the use of the KhoRom resource in accordance with the patterns Author A = Author B (both texts were authored by the same person) and Author A ≠ Author B (texts were authored by different persons) in an equal or almost equal proportion (20 percent to 20 percent) to observe the statistical "behavior" in different instances. Based on the findings, checklist tables were constructed for each genre (non-genre prose fiction, web fiction, web journalism, entertainment journalism, corporate correspondence).

The methodology's performance was assessed from two perspectives: on the one hand, the resulting models of language personalities were considered from the viewpoint of theoretical assessment [Bloomfield L., 1926: 153–164]; [Hjelmslev L., 2005]; [Losev A.F., 2004]; [Apresyan Yu. D., 1966]; [Shtoff V., 1966]; [Revzin I.I., 1977]; [Belousov K.I., 2010: 94-97] etc., along with a set of criteria for indentifying the type of linguistic models (speech activity models, research models, meta-models etc.).

Thus, it could be asserted from a theoretical perspective that an integrative attribution model which includes parameters of three language levels quantitatively objectified and qualitatively assessed by an expert provides a relatively complete, comprehensive and at the same time objective imitation of the original. The point is that the resulting pool of parameters can reflect the information sufficient and necessary for author identification (completeness); the model structure extensively reproduces the author's original, individual style by incorporating the features of all three levels of the language personality (comprehensive imitation) while being devoid of the expert's personal assessments and judgments (objectivity).

All this allows the developed model to successfully solve practical problems of closed set identification (for a limited number of authors) through a pair-wise comparison of written texts of different lengths and genres.

[23] This probabilistic conclusion is due to the fact that under the developed methodology the authorship is to be attributed by the researcher.

Table 1. **Example of a checklist to assess the attribution model output**

Discourse type	**Pearson correlation coefficient**	**Linear regression determination coefficient**	**Student's t-criterion (p-value)**	**Compared texts are likely[24] to be authored by the same person**	**Compared texts are unlikely to be authored by the same person**	**Comments**
Web journalism	at 1.00	at 1.00	normally about 0.95; at least 0.93	+	—	P-value of Student's t-criterion is much less relevant for web journalism than for other discourses. If CC and DC values for web journalism reach 1, one can assume the compared texts were authored by the same person even if p-value of Student's t-criterion is not too high. On the other hand, p-value of Student's t-criterion may seem high but if the values of other metrics are low or not very high, one should adopt a comprehensive approach and analyze all information.
Web journalism	normally about 0.88 — 0.89	normally about 0.71 but can reach 0.77	can be both low (0.60) and relatively high (0.85)	—	+	
Web journalism	not very high at about 0.71	low: at about 0.50	can be very high: 0.98	—	+	

2.4. Validation of the attribution algorithm

The developed algorithm was tested and validated using the following text collections:

collection of prose fiction (10 texts in total) including texts by Sergey Dovlatov ("Nashi" [Our Folks], "Chemodan" [Suitcase], "Inostranka" [Foreigner], "Zapovednik" [Wildlife Sanctuary], "Zona: Zapisky Nadziratelya" [A Prison Camp Guard's Story], and Victor Astafiev ("Oberton" [Overtone], “Posledniy poklon” [The Last Tribute], “Zvezdopad” [Shooting Star Shower], “Tak Khochetsya Zhit” [A Lust for Life]. The algorithm performed to 100 percent in terms of accuracy, precision and recall, with F-measure at 1[24];

collection of web fiction (Kniga Fanfikov web portal, 190 texts in total (https://ficbook.net/) including texts by 3 female and 4 male authors. The algorithm performed to 83 percent in terms of accuracy, precision and recall, with F-measure at 0.8;

collection of web journalism (The Village[25] newspaper, 600 texts in total) including texts by 3 female and 3 male authors. The algorithm performed to 100 percent in terms of accuracy, precision and recall, with F-measure at 1;

collection of entertainment journalism (Ya Plakal web portal, 600 texts in total) including texts by 3 female and 3 male authors. The algorithm performed to 40 percent in terms of accuracy, 0 percent in terms of precision and recall, with F-measure at 0;

collection of corporate Russian-language correspondence (218 texts in total) including texts by 2 female and 2 male authors. The algorithm performed to 83 percent in terms of accuracy, 67 percent in terms of precision and 100 percent in terms of recall, with F-measure at 0.8.

The authors explored a part of each text collection (about 60 percent) using the KhoRom tool in accordance with the patterns Author A = Author B and Author A ≠ Author B in an equal or almost equal proportion to search for true positive (TP), false positive (FP), false negative (FN) and true negative (TN) results of the algorithm's performance. The findings were presented in tables of the following form (Table 2):

Thus, where for the paired texts by A. Yakovlev “Podstavnye znakomstva” — A. Yakovlev “Kak vstrechayut Novy God v platzkarte, samolyote y na trasse” the KhoRom algorithm delivers the following statistics: Pearson correlation coefficient 1; linear regression determination coefficient 1; Student's t-criterion: p-value 0.94, an expert using a checklist table (Table 1) will conclude that “the compared texts were probably authored by the same person”. This conclusion is true to the reality which means that the TP (true positive) column should be selected in Table 2.

As a result of analysis, conclusions were drawn and the following results obtained: the methodology could be used for attributing texts of different dis-

[24] Hereinafter the values of the metrics are specified in connection with interpretation of statistical data through the use of methodological recommendations and checklist tables developed for analytical purposes.

[25] Blocked in Russia.

Table 2. **Calculation of estimates to determine the algorithm's performance**

Text pairs		TP	FN	FP	TN
1	A. Yakovlev: "*Podstavniye Znakomstva*" [Fake acquaintances] — A. Yakovlev "*Kak vstrechayut Novy God v platzkarte, samolyote y na trasse*" [Celebrating the New Year on the train, plane and road] (texts of the same genre by the same male author)	+	—	—	—
2	O. Karasyova: "*Gde deshevle zimovat — na Bali ily Shri-Lanke*" [The cheapest place to stay in winter: Bali versus Sri-Lanka] — O. Karasyova: "*Na chto zhivut zhurnalisty federalnykh kanalov*" [How the journalists of the federal channels make their living] (texts of the same genre by the same female author)	+	—	—	—
3	A. Yakovlev: "*Luchshye sovetskiye mozaiky v Moskve*" [The best Soviet-time mosaics in Moscow] — K. Rukov: "*Vyzhivut tolko spekulyanty: kak russky treider zarabotal million na obvale amerikanskoy birzhy*" [Only speculators will survive: how a Russian trader made a million on a U.S. stock market crash] (texts of the same genre (subject is disregarded) by different male authors)	—	—	—	+
4	O. Karasyova: "*Kak seitchas poyekhat na dachu*" [Going to one's country house right now] — A. Dergachyova: "*Rabochiye snova opustoshayut zapasy bobrov na Yauze*" [Workers destroy beavers' cache in the Yauza River again] (texts of the same genre (subject is disregarded) by different female authors)	—	—	—	+
etc.					

courses, given correct parameterization of models and correct interpretation of statistics for each text. In the course of the study, it was established that:

Student's t-statistics is the most informative for prose fiction discourse (both for established and pulp fiction authors);

stylo-statistics sets are non-informative for modern fiction texts since, as evidenced by experimental data, values of stylo-statistical parameters are closely related for all texts under study;

to identify the author of a journalistic text (in order to acknowledge H_0 hypothesis as true) the values of correlation and determination coefficients should reach 1 (the need for these values to be that high is explained by the length and specific features of such texts). Importantly, it should be admitted that t-statistics — being the most informative for prose fiction texts — is much less relevant to the journalistic discourse. As regards gender differentiation of texts, it is noteworthy that "female" journalistic texts correlate more with other "female" texts which is equally true for "male" texts; the largest correlation differences are observed in individual styles of language personalities of different genders;

short text messages — corporate correspondence, Internet comments — require a representative sample of texts totaling at least 500 words. A limitation of 100 words suggested by C.M. Vul in his time and persisting in forensic authorship attribution to this day [Rubtsova I.I., Yermloayeva E.I., Bezrukova M.Yu. et al., 2007] as a length required to identify an author should be increased when statistical data is added to the analysis. For better handling of such texts, more parameters are currently being developed to construct idiostyle models as representations of language personality of the author since they are linked with the so-called digital handwriting style:

graphical liturative;

graphical hybridization;

playing upon archaic affixes;

using capitalized text elements;

emoticons and other graphical symbols expressing emotion of speech;

texts of different genres can also be validly examined using the developed integrative methodology (for instance, an electronic message can be compared with a feature article): the algorithms performs to 83, 67 and 100 percent in terms of accuracy, precision and recall, respectively, with F-measure at 0.8.

The methodology maximizes the value of idiostyle models rather than output data of an automatic algorithm. These models created as representation of authors' language personalities are understandable, simple, easily interpretable by experts, on the one hand, and provide a sufficiently complete and adequate imitation of the original, on the other hand.

The functionality of the algorithm in question and developed web resource is much wider than the capabilities originally built therein. The methodology can be used not only to solve identification problems of attribution linguistics but also to explore language personalities of writers, journalists, politicians etc. in diagnosing the language personality of specific individuals to address psycholinguistic and psychological problems, explore the generalized language personality of a given social group, subculture etc. to solve sociolinguistic and social science problems. Importantly, when the developed methodology is applied to any of the above cases, the model of a language personality will correspond to the theoretical principles of completeness, simplicity, adequacy, technically accurate and objective description of the original; it will be explanatory, communicative and interpretable.

3. Conclusions

Thus, it should be asserted that the integrative methodology combining the approaches of interpretative and cognitive linguistics with traditional stylometry is undoubtedly effective. The integrative approach seems to be the most appropriate basis for development of forensic investigation in Russia for a number of reasons: peculiarities of the regulatory framework in Russia; strong national tradition of interpretative linguistics; inadequacy of all known fully automatic methods of text attribution for forensic purposes (in terms of accuracy).

Importantly, under the proposed approach experts are not expected to do the interpretative part of the analysis themselves since the identification criteria can be assigned automatically while the process can be automated without prior manual text pre-processing and without using syntactic parsers. This feature is useful for developing a software prototype applicable, in particular, to problems of forensic linguistics as experts in authorship attribution do not always possess the required knowledge of corpus linguistics, statistics etc. The integration of all analytical modules in one software interface will allow to partially or probably fully automate the attribution analysis.

References

1. Apresyan Yu.D. (1966) *Ideas and methods of modern structural linguistics*. Moscow: Nauka, 302 p. (in Russ.)

2. Bacciu A., Morgia M. et al. (2019) Cross-domain authorship attribution combining instance-based and profile-based features. Notebook for PAN at CLEF 2019. Available at: http://ceur-ws.org/Vol2380/paper_220.pdf (accessed: 05.07.2020)

3. Baranov A.N. (2001) Introduction to Applied Linguistics. Manual. Moscow: Editorial URSS, 360 p. (in Russ.)

4. Batura T.V. (2012) Formal Ways of text authorship identification. *Vestnik Novosibirskogo gosudarstvennogo universiteta. Informatcionnye tehnologii*=Journal of Novosibirsk State University. Information Technology, vol. 2, no. 4, pp. 81–94 (in Russ.)

5. Belousov K.I. (2010) Linguistic Models and Language Reality Modeling Issues. *Vestnik Orenburgskogo gosudarstvennogo universiteta*=Journal of Orenburg State University, no. 11, pp. 94–97 (in Russ.)

6. Bessmertny I.A., Nugumanova A.B. (2012) Automatic Thesaurus Building Method Based on Statistical Processing of Texts in the Natural Language. *Izvestia Tomskogo gosudarstvennogo politekhnicheskogo universiteta*=Proceedings of Tomsk State Polytechnical University, no. 5, pp. 125–130 (in Russ.)

7. Bloch B. (1948) A set of postulates for phonemic analysis. *Language*, vol. 24, no. 1, pp. 3–46.

8. Bloomfield L. (1926) A set of postulates for the science of language. *Language*, vol. 2, no. 2, pp. 153–164.

9. Campbell L. (1867) *The Sophisties and Polilicus of Plato*. Oxford: Clarendon Press, 170 p.

10. Coulthard M. (2004) Author identification, idiolect, and linguistic uniqueness. *Applied Linguistics*, vol. 24, no. 4, pp. 431–447.

11. Custódio J., Paraboni I. (2018) EACH-USP Ensemble Cross-Domain Authorship Attribution. Notebook for PAN at CLEF 2018. Available at: http://ceur-ws. org/Vol-2125/paper_76.pdf (accessed: 05.07.2020)

12. Encyclopedia of Forensic Science (1999) T.B. Averyanova (ed.). Moscow: Prospekt, 442 p. (in Russ.)

13. Galyashina E.I., Yermolova E.I. (2005) Linguo-forensic tools for authorship attribution of written and oral texts. Papers of the International Research Conference. Moscow, pp. 20–22 (in Russ.)

14. Gomzin A. et al. (2018) Detection of author's educational level and age based on comments analysis. Paper presented at Dialogue, Moscow, 30 May–2 June 2018. Available at: URL: http://www.dialog-21.ru/media/4279/gomzin_turdakov.pdf (2018) (accessed: 05.07.2020) (in Russ.)

15. Goroshko E.I. (2003) Forensic authorship attribution: gender identification of the author of a document. Theory and practice of forensic investigation and science. *Pravo*, no. 3, pp. 221–226 (in Russ.)

16. Hjelmslev L. (2005) *Prolegomena to a theory of language*. Moscow: Editorial URSS, 243 p. (in Russ.)

17. Juola P. (2006) Authorship Attribution. *Foundations and Trends in Information Retrieval*, vol. 1, no. 3, pp. 233–334.

18. Ionova S.V., Ogorelkov I.V. (2020) Gender-based Individual Speech Diagnostics in Authorship Attribution: Quantitative Approach. *Vestnik*

Volgogradskogo gosudarstvennogo universiteta. Linguistika=Journal of Volgograd State University. Linguistics, vol. 19, no. 1, pp. 115–127. DOI:https://doi.org/10.15688/jvolsu2.2020.1.10 (in Russ.)

20. Karaulov Yu. N. (1987) *Russian Language and Language Personality*. Moscow: Nauka, 264 p. (in Russ.)

21. Khmelyov D.V. (2002) Linguo-analyzer. E-resource. Available at: URL: http://www.rusf.ru/books/analysis/ (accessed: 16.11.2017) (in Russ.)

22. Khomenko A., Baranova Yu., Romanov A., Zadvornov K. (2021) The Linguistic modeling as a basis for creating authorship attribution software. Computational linguistics and intellectual technologies. Proceedings of the International Conference "Dialogue 2021" Moscow. Available at: URL: http://www.dialog-21.ru/media/5315/khomenkoaplusetal048.pdf (accessed: 23.06.2021) (in Russ.)

23. Komissarov A.Yu. (2000) Forensic Investigation of Written Speech: Manual. Moscow: Forensic Agency of the Interior Ministry of the Russian Federation, 126 p. (in Russ.)

24. Koppel M., Schler J. (2003) Exploiting Stylistic Idiosyncrasies for Authorship Attribution. Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis, vol. 69, pp. 72–80.

25. Korobov M. (2015) Morphological analyzer and generator for Russian and Ukrainian languages. In: Khachay M.Y., Konstantinova N.A. (eds.). AIST 2015. CCIS, vol. 542, pp. 320–332. Available at: https://doi.org/10.1007/978-3-319-26123-2_31 (accessed: 05.07.2020) (in Russ.)

26. Leonard R., Ford J., Christensen T. (2017) Forensic linguistics: applying the science of linguistics to the issues of the law. *Hofstra Law Review*, vol. 45, pp. 881–897.

27. Linguistics of Constructions (2010) E.V. Rakhilina (ed.). Moscow: Azbukovnik Publishing, 584 p. (in Russ.)

28. Litvinova T.A. (2019) Idiolect as Object of Corpus Idiolectology: Towards a New Field in Linguistics. *Vestnik Novgorodskogo gosudarstvennogo universiteta imeni Yaroslava Mudrogo*=Bulletin of the Yaroslav Mudriy Novgorod State University, no. 7, pp. 1–5 (in Russ.)

29. Litvinova T., Rangel F. et al. (2017) Overview of the Rus Profiling PAN at FIRE Track on Cross-genre Gender Identification in Russian. Working notes of FIRE 2017. Forum for Information Retrieval Evaluation. Bangalore, pp. 1–7. Available at: URL: http://ceur-ws.org/Vol-2036/T1-1.pdf (accessed: 05.07.2019) (in Russ.)

30. Litvinova T.A., Gromova A.V. (2020) The Use of Computer Technologies for Forensic Authorship Attribution: Issues and Prospects. *Vestnik Volgogradskogo gosudarstvennogo universiteta. Lingvistika*=Journal of Volgograd State University. Linguistics, vol. 19, no. 1, pp. 77–88. DOI: https://doi.org/10.15688/jvolsu2.2020.1.7 (in Russ.)

31. Litvinova T., Sboev A., Panicheva P. (2018) Profiling the age of Russian bloggers. Proceedings of the 7th International Conference, AINL 2018. Saint Petersburg, pp. 167–177 (in Russ.)

32. Losev A.F. (2004) Introduction to the General Theory of Linguistic Models. Moscow: Editorial URSS, 293 p. (in Russ.)

33. Marusenko M.A. (1990) The use of image recognition methods for attribution of anonymous and pseudonymous literary texts. Leningrad: University, 1990. 164 p. (in Russ.)

34. Marusenko M.A. (2003) Attribution of anonymous and pseudonymous texts as a standard image recognition problem. *Istoriographiya y istochnikovedeniye otechestvennoy istorii*=Historiography and Research of Sources of National History, no. 3, pp. 18–22 (in Russ.)

35. McMenamin G. (2002) *Forensic linguistics: advances in forensic stylistics*. London: Routledge, 361 p.

36. Murauer B., Tschuggnall M., Specht G. (2018) Dynamic Parameter Search for Cross-Domain Authorship Attribution. Notebook for PAN at CLEF 2018. Available at: http://ceur-ws.org/Vol-2125/paper_84.pdf (accessed: 05.07.2020)

37. 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. Available at: http://ceur-ws.org/Vol-2380/paper_49.pdf (accessed: 05.07.2020)

38. Paducheva E.B. (1974) *On semantics of syntax*. Moscow: Nauka, 291 p. (in Russ.)

39. Radbil T.B., Markina M.V. (2019) Probability Statistical Models in Attribution of Texts by Russian Language Authors. *Politicheskaya Lingvistika*=Political Linguistics, no. 2, pp. 156–166 (in Russ.)

40. Revzin I.I. (1977) *Modern Structural Linguistics: Issues and Methods*. Moscow: Nauka, 263 p. (in Russ.)

41. Rodionova E.S. (2008a) Linguistic Methods of Attribution and Dating of Literary Texts: towards Corneille-Moliere Problem. Candidate of Philological Sciences Summary. Saint Petersburg, 25 p. (in Russ.)

42. Rodionova E.S. (2008b) Methods of literary text attribution. In: Structural and applied linguistics: inter-university collection. A.S. Gerda (ed.). Saint Petersburg: University, 2008, pp. 118–127 (in Russ.)

43. Rogov A.A. et al. (2019) Software support for solving text attribution problems. *Programmnaya Inzheneriya*=Programming Engineering, no. 5, pp. 234–240 (in Russ.)

44. Romanov A.S. (2010) Methodology and Software Package for Identification of Authors of Unknown Texts. Candidate of Engineering Sciences Summary. Tomsk, 26 p. (in Russ.)

45. Romanov A.S., Kurtukova A., Fedotova A. et al. (2021) Authorship Identification of a Russian-Language Text Using Support Vector Machine and Deep Neural Networks. *Future Internet*, vol. 13, issue 3, pp. 1–16.

46. Rubtsova I.I., Yermolayeva E.I., Bezrukova A.I. et al. (2007) Comprehensive methodology of authorship attribution: methodological recommendations. Moscow: Forensic Agency of the Ministry of Interior, 192 p. (in Russ.)

47. Russian Grammar Rules: Collected works (2005) N. Yu. Shvedov (ed.). Moscow: Nauka, 665 p. Available at: URL: http://rusgram.narod.ru/index.html. (in Russ.)

48. Shevelyov O. G. (2007) Methods of automatic classification of texts in the natural language: manual. Tomsk: TML-Press, 144 p. (in Russ.)

49. Shtoff V. (1966) *Modeling and philosophy.* Moscow: Nauka, 304 p. (in Russ.)

50. Shuy R. (2005) *Creating Language Crimes: How Law Enforcement Uses (and Misuses) Language*. N. Y.: Oxford University Press, 194 p.

51. Sidorov Yu.B. et al. (1999) Computer-assisted system for linguistic analysis of literary texts. In: Saint Petersburg Assembly of Young Researchers and Specialists. Abstracts of reports. Saint Petersburg: University Press, p. 66. (in Russ.)

52. Stamatatos E. (2017) Authorship attribution using text distortion. Proceedings of 15th Conference of the European Chapter of the Association for Computational Linguistics, Long Papers, pp. 1138–1149.

53. Stepanenko A.A. (2017) Gender attribution of computer network communication texts. *Vestnik Tomskogo gosudarstvennogo universiteta*=Journal of Tomsk State University, no. 5, pp. 17–25. DOI: 10.17223/15617793/415/3 (in Russ.)

54. Timashev A.N. (2007) Atributor: version 1.01: software description. Available at: URL: http://www.textology.ru/atr_resum.html (accessed: 01.02.2016) (in Russ.)

55. Vinogradov V.V. (1961) *The Authorship Problem and Theory of Styles*. Moscow: Goslitizdat, 614 p. (in Russ.)

56. Vul S.M. (2007) Forensic Authorship Identification: Methodological Basis. Guidebook. Kharkov: KhNIISE Press, 64 p. (in Russ.)

57. Wright D. (2017) Implementing word n-grams to identify authors and idiolects: a corpus approach to a forensic linguistic problem. *International Journal of Corpus Linguistics*, vol. 22, no. 2, pp. 212–241.

58. Zakharov V.N. et al. (2000) System Support Programme for Attribution of Articles Authored by F.M. Dostoevsky. *Trudy Petrozavodskogo gosudarstvennogo universiteta*=Research Works of the Petrozavodsk

State University. Applied Mathematics and Information Technology Series, issue 9, pp. 113–122 (in Russ.)

59. Zakharov V.N., Khokhlova M.V. (2008) The Statistical Method for Identification of Collocations. Language Engineering in Search of Meanings. Collection of reports to the conference workshop “Web-Based Linguistic Information Technologies”. 11th All-Russia Joint Conference “Internet and Modern Society”. Saint Petersburg: University, 2008, pp. 40–54 (in Russ.)

Information about the authors:

T.B. Romanova — Professor, Doctor of Sciences (Philology).

A.Yu. Khomenko — Senior Lecturer, Candidate of Sciences (Philology), expert.

The article was submitted to the editorial office 17.01.2022; approved after reviewing 21.03.2022; accepted for publication 15.04.2022.

Legal Issues in the Digital Age. 2022. Vol. 3. No. 2.
Вопросы права в цифровую эпоху. 2022. Т. 3. № 2.

Comment

Review
УДК: 347
DOI:10.17323/2713-2749.2022.2.116.140

Key Issues in the Intellectual Property Court's Presidium Rulings

Natalia Igorevna Kapyrina[1], Maria Alexandrovna Kolzdorf[2]

[1] MGIMO University, 76 Prospekt Vernadskogo, Moscow 119454, Russian Federation, n.kapy rina@ my.mgimo.ru, ORCID: 0000-0003-1276-1600, Researcher ID: AAQ-3784-2021

[2] National Research University Higher School of Economics, 20 Myasnitskaya Str., Moscow 101000, Russian Federation, Researcher ID: AAI-1625-2019, mkolzdorf@hse.ru, ORCID:0000-0003-3227-3348, Researcher ID: AAI-1625-2019

Abstract

The comment reviews key positions in the rulings of the Presidium of the Russian Intellectual Property Court (IPC) issued in December 2021 and January 2022. This Chamber hears cassation appeals against the decisions of the IPC first instance and deals primarily, but not only, with matters of registration and validity of industrial property rights. Therefore, this review predominantly covers substantive requirements for patent and trademark protection, as well as procedural issues both in the administrative adjudicating mechanism at the Patent office (Rospatent) and at the IPC itself. The current review encompasses a variety of topics related to trademark law: signs that are contrary to the public interest, signs conflicting with an earlier trademark or an appellation of origin, signs using a geographical name, deceptive signs, the comparison of signs, trademark revocation for lack of use, unfair competition, procedural challenges, etc. The review further considers one patent case, in which the IPC Presidium resolved the issue of establishing priority date for a divisional application for a utility model derived from an application initially filed for an invention.