ISSN-p 2306-7365
ISSN-e 2664-0686

UDC 801.8; IRSTI 16.41.21
https://doi.org/10.47526/2024-3/2664-0686.72

R.Zh. SAURBAYEV[1], A.K. ZHETPISBAY[2], F.T. YEREKHANOVA[3]✉
[1]Candidate of Philological Sciences, Professor
Toraighyrov University (Kazakhstan, Pavlodar), e-mail: rishat_1062@mail.ru
[2]Candidate of Philological Sciences, Associate Professor
Pavlodar Margulan Pedagogical University (Kazakhstan, Pavlodar), e-mail: a_kanzhygaly@mail.ru
[3]Candidate of Philological Sciences, Senior Lecturer of Central Asian Innovation University
(Kazakhstan, Shymkent), e-mail: siliconoasis702@gmail.com

PHILOLOGICAL STUDIES: COMPUTER ASPECTS OF STYLOMETRY AUTOMATION

Abstract. The article aims at considering stylometry automation of philological studies. The relevance of this article stems from the need to trace and critically analyze the results of numerous studies in the interdisciplinary field that have been actively developed in recent years to identify the author of the text using artificial intelligence methods (authorship attribution and profiling) as well as to provide theoretical foundations and a comprehensive stylometric methodology (i.e. based on the analysis of quantifiable linguistic features using statistical methods and machine learning algorithms) to identify the author of the text, based on the principles of explanation, objectivity, evidence, and open science. Within the framework of subject and activity identification idiolectology - is a developing scientific direction that focuses specifically on the systematic study of the phenomenon of idiolect in the identification of computer aspects using modern achievements of computational and corpus linguistics, and data science. The authors of the article claim that the task of computational and corpus linguistics is to provide the researcher with all the necessary material, to prepare the data for counting, and to offer a wide range of computational procedures that can be used to test hypotheses, together to confirm or refute ever subtler and profound philological observations.

Philological problems, in solution of which language information is used, usually have, a clear application, orientation, the language, and style of the text in such a situation are not the goal of the study, but a means of solving extra-linguistic problems. The solution of a specific philological problem (e.g., the problem of disputed authorship) is usually not limited to the rigid framework of a particular research methodology but is carried out using methods and facts in various fields of knowledge and practical activities.

Keywords: computational linguistics, stylometry, text structuring, computer aspects, stylometry automation, artificial intelligence, attribution of authorship.

Бізге дұрыс сілтеме жасаңыз:
Saurbayev R.Zh., Zhetpisbay A.K., Yerekhanova F.T. Philological Studies: Computer Aspects of Stylometry Automation // *Ясауи университетінің хабаршысы*. – 2024. – №3 (133). – Б. 47–61. https://doi.org/10.47526/2024-3/2664-0686.72
Cite us correctly:
Saurbayev R.Zh., Zhetpisbay A.K., Yerekhanova F.T. Philological Studies: Computer Aspects of Stylometry Automation // *Iasaui universitetinin habarshysy*. – 2024. – №3 (133). – B. 47–61. https://doi.org/10.47526/2024-3/2664-0686.72

Date of receipt of the article 10.12.2023 / Date of acceptance 27.09.2024

ISSN-p 2306-7365
ISSN-e 2664-0686

Р.Ж. Саурбаев[1], Ә.Қ. Жетпісбай[2], Ф.Т. Ереханова[3]
[1]филология ғылымдарының кандидаты, профессор
Торайғыров университеті (Қазақстан, Павлодар қ.), e-mail: rishat_1062@mail.ru
[2]филология ғылымдарының кандидаты, доцент, Ә. Марғұлан атындағы Павлодар педагогикалық университеті (Қазақстан, Павлодар қ.), e-mail: a_kanzhygaly@mail.ru
[3]филология ғылымдарының кандидаты, Орталық Азия Инновациялық университетінің аға оқытушысы (Қазақстан, Шымкент қ.), e-mail: siliconoasis702@gmail.com

Филологиялық зерттеу: стилометрияны автоматтандырудың компьютерлік аспектілері

Аңдатпа. Мақаланың мақсаты – филологиялық зерттеулерде стилометрияны автоматтандыруды қарастыру. Бұл мақаланың өзектілігі жасанды интеллект әдістерін (авторлық атрибуция және профильдеу) қолдана отырып, мәтін авторын анықтау үшін соңғы жылдары белсенді дамып келе жатқан пәнаралық саладағы көптеген зерттеулердің нәтижелерін бақылау және сыни тұрғыдан талдау қажеттілігімен, сондай-ақ теориялық негіздер мен күрделі стилометриялық әдістемені ұсынумен байланысты (яғни, статистикалық әдістер мен машиналық оқыту алгоритмдерін қолдана отырып, сандық лингвистикалық ерекшеліктерді талдау негізінде) түсіндіру, объективтілік, дәлелдеу және ашық ғылым принциптеріне негізделген мәтін авторын анықтау.

Пәндік және белсенділікті сәйкестендіру шеңберінде идиолектология - компьютерлік және корпустық лингвистиканың, сондай-ақ деректер ғылымының заманауи жетістіктерін пайдалана отырып, компьютерлік аспектілерді анықтау кезінде идиолект құбылысын жүйелі түрде зерттеуге бағытталған және дамып келе жатқан ғылыми бағыт. Мақала авторлары компьютерлік және корпустық лингвистиканың міндеті зерттеушіге барлық қажетті материалдарды ұсыну, санау үшін деректерді дайындау және терең филологиялық бақылауларды растау немесе жоққа шығару үшін біріктірілген гипотезаларды тексеру үшін пайдаланылуы мүмкін есептеу процедураларының кең ауқымын ұсыну деп санайды.

Тілдік ақпаратты шешуде қолданылатын филологиялық есептер, әдетте, мұндай жағдайда мәтіннің нақты қолданылуы, бағыты, тілі мен стилі зерттеудің мақсаты емес, экстралингвистикалық есептерді шешудің құралы болып табылады. Сондай-ақ, белгілі бір филологиялық мәселені шешу (мысалы, даулы авторлық мәселе), әдетте, белгілі бір зерттеу әдіснамасының қатаң шеңберімен шектелмей, әртүрлі білім мен практикалық қызмет салаларындағы әдістер мен фактілерді қолдану арқылы жүзеге асырылады.

Кілт сөздер: компьютерлік лингвистика, стилометрия, мәтінді құрылымдау, компьютерлік аспектілер, стилометрияны автоматтандыру, жасанды интеллект, авторлық атрибуция.

Р.Ж. Саурбаев[1], А.К. Жетписбай[2], Ф.Т. Ереханова[3]
[1]кандидат филологических наук, профессор, Павлодарский государственный университет имени Торайгырова (Казахстан, г. Павлодар), e-mail: rishat_1062@mail.ru
[2]кандидат филологических наук, доцент, Павлодарский педагогический университет имени А. Маргулана (Казахстан, г. Павлодар), e-mail: a_kanzhygaly@mail.ru
[3]кандидат филологических наук, старший преподаватель Центрально-Азиатского инновационного университета (Казахстан, г. Шымкент), e-mail: siliconoasis702@gmail.com

Филологические исследования: компьютерные аспекты автоматизации стилометрии

Аннотация. Целью статьи является рассмотрение автоматизации стилометрии в филологических исследованиях. Актуальность данной статьи обусловлена необходимостью

ISSN-p 2306-7365
ISSN-e 2664-0686

проследить и критически проанализировать результаты многочисленных исследований в междисциплинарной области, которая активно развивается в последние годы для идентификации автора текста с использованием методов искусственного интеллекта (атрибуция авторства и профилирование), а также предоставить теоретические основы и комплексную стилометрическую методологию. (т.е. на основе анализа поддающихся количественной оценке лингвистических особенностей с использованием статистических методов и алгоритмов машинного обучения) для идентификации автора текста, основываясь на принципах объяснения, объективности, доказательности и открытой науки.

В рамках предметной и деятельностной идентификации идиолектология - это развивающееся научное направление, которое конкретно фокусируется на системном изучении феномена идиолекта при идентификации компьютерных аспектов с использованием современных достижений компьютерной и корпусной лингвистики, а также науки о данных. Авторы статьи утверждают, что задача компьютерной и корпусной лингвистики состоит в том, чтобы предоставить исследователю весь необходимый материал, подготовить данные для подсчета и предложить широкий спектр вычислительных процедур, которые могут быть использованы для проверки гипотез, вместе взятых, чтобы подтвердить или опровергнуть все более тонкие и глубокие филологические наблюдения.

Филологические задачи, при решении которых используется языковая информация, обычно имеют четкое применение, ориентацию, язык и стиль текста в такой ситуации являются не целью исследования, а средством решения экстралингвистических задач. Решение конкретной филологической проблемы (например, проблемы оспариваемого авторства) обычно не ограничивается жесткими рамками конкретной исследовательской методологии, а осуществляется с использованием методов и фактов из различных областей знаний и практической деятельности.

Ключевые слова: компьютерная лингвистика, стилометрия, структурирование текста, компьютерные аспекты, автоматизация стилометрии, искуственный интеллект, атрибуция авторства.

Introduction

The third scientific revolution that we are living through is radically altering social reality. The invention of the steam engine was linked to the first, and the usage of electricity to the second. And lastly, the digital stage of the third revolution starts at the start of the 1970s [1, p. 12].

Philological analysis always requires a holistic integrated approach. This includes an analysis within the text, an analysis of the conditions of the creation of a text, and an examination of the relationship of that text to other texts. Philologists' conclusions are often subjectively based as they are on common sense, conjecture, intuition, and worldly wisdom and rarely lay claim to unconditional scientific rigor and finality of conclusion.

When it comes to the notion of the text, V.I. Bortnikov assumes that "Text is the highest unit in the system of language. However, if we consider the difference between language and speech, we should understand any text as a speech unit. The problem here, which still has no distinct answer, is whether there is a corresponding language unit or not" [2, p. 6].

Moreover, there is an article related to stylometric characteristics of short texts done by E.V. Popov and N.C. Lagutina. Three layers of stylometric metrics for posts from the social network Twitter are examined in that paper. The authors explained the aspects of their computation when creating a vector representing the numerical attributes of texts. Experiments were carried out to determine authorship and classify corpora of short texts by subject area based on the comparison of vectors. The outcomes of the experiment demonstrated that different levels of features have distinct contributions to the categorization quality [3].

ISSN-p 2306-7365
ISSN-e 2664-0686

In stylometry, as in traditional philology, the text remains a first empirical fact, but stylometric interest focuses primarily on the quantitative organization of the text, and the researcher's position is characterized by objectivity and methodological rigor [4]. In a computer environment, these cognitive principles are implemented with extreme rigor and acquire the character of a categorical imperative. The vision of the text here becomes "material", procedural, and technological; the researcher communicates with the text through the computer programs, as with a directly "tangible" material entity constructed according to certain linguo-polygraphic laws. According to K.K. Sarekenova, A.M. Melis, and S.R. Toibekova, the term "Computational Linguistics", mainly refers to language services computer programs, and language data used in modeling processing, gathering the necessary computer technology, and other language research related to computer-intervention [5, pp. 167–168].

The complexity of the traditional philological approach is lost in computer stylometry and the infinite richness of associations that arises in communicating with the text and its environment is lost. At the same time, almost unlimited possibilities are gained for uniform and rapid processing of large amounts of data from printed texts. The effectiveness of computer work with texts is especially great when complex methods of multidimensional text analysis are used: distributive-statistical method, algorithms of linguistic decoding methods of quantitative typology and taxonomy of texts, algorithms of statistical diagnostics.

Regarding "corpus linguistics", some works by domestic researchers come to mind. J. Jubataeva, in particular, says that "corpus linguistics is one of the new directions in teaching a foreign language. In the last decade, the field of corpus linguistics began to enter scientific knowledge to a significant extent. Especially in linguistic studies, it was found that corpora are very valuable material from the point of view of practical use, i.e., when preparing various grammars with dictionaries" [6]. According to Z.A. Makhanova et al, "A corpus is a set of texts selected and processed according to certain rules, used as a basis for language study. They are used for statistical analysis and testing of statistical hypotheses, confirmation of linguistic rules in this language" [7, p. 36].

According to J. Langlois, "Stylometry is based on the principle that each author or organization develops its linguistic style and expression. Therefore, documents from an individual and an organization can be analyzed to identify the most probable authorship" [8, p. 51].

Research methods and materials

There are two main approaches to this problem: stylistic, and idiographic, i.e. based on the analysis of individual features of the text, subjectively selected by the researcher for each identification situation. Another approach consists of the application of data science methods and stylometry: quantitative and nomothetic, i.e. based on the analysis of a variety of stylistic features of the text using statistical methods or machine learning algorithms. As the English scholar D. Wright aptly notes, the aforementioned approaches develop in parallel at best and compete with each other at worst [9].

The absolute advantage of the stylometric approach is a higher degree of objectivity of the results obtained with its help compared to the stylistic approach. Due to the emergence of new powerful machine learning algorithms, including Deep Learning, the focus of work on stylometric identification of the author of a text, mostly performed by computer scientists, is on improving the quality metrics of the identification models, but not on their interpretability.

Results and discussion

The level of development of problem analysis of scientific literature shows that the number of works in the field of stylometry, i.e. identification of the author of a text based on linguistic analysis, is rapidly increasing, with the most active development in recent years. Despite the great research interest, this problem is far from being solved, which is due, on the one hand, to the

ISSN-p 2306-7365
ISSN-e 2664-0686

insufficient level of development of the theoretical foundations of linguistic identification of the author of a text. This is due, on the one hand, to the insufficient level of development of the theoretical foundations of the linguistic identification of the author of a text, in particular, to the lack of comprehensive theoretical research in the field of the study of the phenomenon of individual speech and individual speech activity and its essential features, and, on the other hand, to the lack of a comprehensive approach to the problem of the linguistic identification of the author of a text that combines objectivity, evidence, and explanation as leading methodological principles.

In the field of computational stylometry, the following main research areas are distinguished.

1. Theoretical research: a study of quantitative patterns in symbolic sequences, the study of the conditions of the law of large numbers in the text, search for robust linguistic statistical methods, stable statistics, etc.

2. Primary processing of linguistic data: Construction of distributional series, calculation of statistics, statistical analysis, testing of statistical hypotheses, and construction of theoretical models based on experimental data.

3. Systematic and taxonomic objectives:

A. Processing of multidimensional data using standard algorithmic procedures: Factor, discriminant, cluster, and other methods of multidimensional analysis.

B. Processing of linguistic data using special linguistic methods: decoding algorithms, distributive statistical methods, methods of dating, attribution, diagnostics, and typology of texts.

4. Lexicographic processing of text data: Creation of frequency and alphabet-frequency dictionaries, concordance dictionaries, word indicators, morpheme dictionaries, rhymes, scribal dictionaries, minimal dictionaries, keyword dictionaries, association dictionaries, idiolect dictionaries, etc.

5. Information retrieval tasks:

A. Automatic text search.

B. Search for text units with certain qualitative and quantitative features to solve stylistic and grammatical problems.

6. Linguodidactic tasks:

A. Supporting the teaching of Russian through dictionaries including the development of minimum dictionaries, bilingual dictionaries, reverse dictionaries, etc.

B. Developing systems of programmed learning using information on genre differentiation of texts.

Stylometry is a scientific discipline that measures the stylistic features of texts to organize, diagnose, identify, parameterize, taxonomize, assign, and periodize them. Stylistic features of prose texts change over time for literature in different languages, so they can serve as indicators of the epoch of the creation of works.

Researchers working in the field of theory and practice of forensic authorial expertise point out the need to increase the objectivity of methods for linguistic identification of the author of a text [10], [11]. At the same time, the ability to interpret modeling results and apply identification models to the analysis of "small data" is extremely important for this field. However, the low general mastery of data analysis methods among linguists leads to the fact that several papers explaining the combination of qualitative and quantitative approaches in solving forensic authorship problems either do not use precise methods at all or use them incorrectly. The most consistent principles of "evidence-based linguistics" are embodied by philologists who use a stylometric approach to solve certain tasks - to identify the author of a given text. In the works [12], [13], [14], [15], [16] the time of creation of a literary work is considered. In the works of these authors, precise methods are used to solve specific philological problems, with special attention paid to the possibility of basic interpretation of the modeling results. As a rule, however, the range of texts studied is limited to works of fiction of considerable size.

ISSN-p 2306-7365
ISSN-e 2664-0686

It should also be noted that in most scientific works the tasks of individual and group identification of the author of the text are considered separately. It is obvious that these tasks are different applied aspects of a basic scientific problem - the problem of studying the speech activity of the individual embodied in the texts. This is a complex, hierarchical-dynamic system that reflects the features of a person, both as a unique individual and as a representative of a particular group.

All complex algorithmic procedures of automatic text processing are variants of diagnostic work related to ordering and systematizing texts, parts of texts, and units of texts. An important stage of this work is the assignment of text units and their division. The orientation to computer processing dictates the necessity to deal with the volumetric-compositional arrangement of the linguistic material in the text, which is explicitly expressed in the rules and norms of graphic representation and spatial arrangement of units and parts of the text. The linguo-poligraphic structure of the text reflects not only purely external features but also stylistic features of the text.

In the future, the computer-aided explication of the linguo-poligraphic organization of the text will be referred to as structuring, thus emphasizing the computer-aided, formal aspect of the process of text division. Structuring is based on three principles: formality, homogeneity, and linguistic-stylistic reliability.

The principle of formality means that the process of dividing the text is based on elements of graphic design of the text, taking into account the rules of Russian punctuation and spatial arrangement of the material on the pages of printed publications. The use of dictionary information is allowed only as an auxiliary tool, for example, when analyzing standard abbreviations, analyzing combinations equivalent to a word, and resolving the homonymy "hyphen – hyphenation".

The principle of homogeneity of division assumes automatic attribution of text units to one of the structural and functional types of speech within a particular genre. For example, within the limits of fiction, subsets of units belonging to different types of written speech are automatically allocated and grouped, followed by splitting the subset of units of someone else's speech into three homogeneous aggregates: direct speech, mixed speech, and nested direct speech. The principle of homogeneity of articulation in its content is close to the contact-variable articulation of the text but differs from it by the typology of articulation since it is based on the general punctuation and graphic features of the design of someone else's and the author's speech in printed texts. In the genre of fiction, this principle applies to such units as word usage, sentences, and paragraphs; it is aimed at ensuring the reliability of linguo-statistical and stylometric studies.

The principle of linguistic-stylistic reliability implies the rejection of a guaranteed result as the result of the work of an error-free algorithm for dividing the text. If necessary, structuring errors are eliminated by a philologist-researcher in dialogue with a computer system. The linguistic and stylistic reliability of the division itself depends on the functional, stylistic, and genre affiliation of a particular text, as well as on the individual author's manner of writing.

The practical expediency of structuring the text is well traced by the example of creating a machine fund of lexical units. The currently applied rule of defining the context as a certain number of characters to the right and left of a keyword cannot be considered flawless and is used only because the text stored in the computer's memory for receiving quotation cards is not structured. In the presence of a structurally fragmented text, the concept of "context" has an informal meaning even in formal systems, and the context itself, at the request of the researcher, can be constructed from an arbitrary number of units of various levels: word usage, phrases, sentences, paragraphs.

The formal punctuation structuring method is based on a detailed analysis of the rules of Russian punctuation and the transformation of these rules into a formal logical scheme of the recognizing automaton - the algorithmic core of the DISSKOTE dialogue system.

The minimum requirement for the text entered into the computer is the isolation of capital letters, the use of which is one of the main, fundamental points in Russian spelling. In modern Russian, uppercase letters are used to highlight:

ISSN-p 2306-7365
ISSN-e 2664-0686

a) a new segment of the text – and this shows their important role as a signalized of text division;

b) various types of lexical units - in these cases, uppercase letters perform their semantic function.

The development of a formal punctuation method for structuring fiction has encountered several significant contradictions between the fuzziness of the punctuation system, on the one hand, and the rigor of the formalized recording of the algorithm, on the other. The vagueness of the use of punctuation marks, and the lack of mandatory regulation, and all these hindered the development of works on the formal structuring of texts. Especially since artistic texts that were published at different times, carry not only punctuation features characteristic of their era, author's preferences, and freedom in the use of certain signs, but also editorial editing of subsequent editions.

As an example, let us consider the question of using a dash as an additional sign after delimiting punctuation marks. Modern punctuation rules do not provide for the use of a dash at the beginning of a sentence – after the delimiting mark at the end of the previous sentence except for their use in dialogical speech; moreover, this variant was not previously commonly used. Nevertheless, in the works of writers of the XIX century, we find: *Above each roof is a high birdhouse pole; above each porch is a carved iron steeple. – Uneven window panes cast rainbow colors (I.S. Turgenev. Village)*

The reasons for such an author's use of dashes lie, of course, in the desire to express the border of the meaning of the utterance more clearly, to outline their stylistic coloring, a new theme, a change of facts, or a change of feelings, etc. Such cases, unlike, say, the use of dashes after the same delimiting signs in dialogical speech, should be attributed to features of individual writing style: *– A bottle of beer! – So it's not possible? Curious (A.M. Gorky. The life of an unnecessary person).*

These cases not only feed our reflections on stylistic author-individual preferences and genre-historical features but also require – taking into account the problem of text division – their systematization and formalization.

The selection of a sentence – the main unit of the text during structuring – is based on the analysis of the right and left environment of punctuation marks, potentially acting as delimiters of sentences: dots, question and exclamation marks, and ellipses. The analysis of the environment of such signs is required because in literary texts the signs “.”, “!”, “?” in addition to its traditional place at the end of a sentence, it is often found in the middle of a sentence, whereas for other types of texts, this is practically excluded.

The formal punctuation approach also dictates the need to introduce an operational definition of the sentence. A sentence is a sequence of symbolic chains and punctuation marks between them from one end sign to another. Here, the “symbolic chain” is a textual graphic representation of word usage, and the “end sign” is a composite punctuation–a spatial segment of the text that allows formally recognizing the situation “end of the sentence” in the line.

For example, the text segment <dot + dash + capital letter> is the “final sign” in the above fragment from the story of I.S. Turgenev's “Village”. The general scheme of deciding on the selection of the next sentence from the text looks – in the most general form – as follows: a) a symbol belonging to a set of potential delimiters of sentences is recognized; b) the analysis of the right or left environment of the selected symbol is performed; c) the sentence is highlighted only if the search ends at the terminal vertex "end sign"; otherwise, the procedure is repeated for the next delimiting punctuation mark.

The development of a machine algorithm for selecting sentences required a thorough analysis and systematization of the known rules of punctuation of sentences. The purpose of this study was to compile a list of all possible and occurring situations in real literary texts in which the symbols of the end of sentences (signs “.”, “!”, “?”) mean the end of the sentence, as well as those situations in

ISSN-p 2306-7365
ISSN-e 2664-0686

which this is not the case. As a result, three variants of situations have been identified that certainly fix the end of the sentence: A) ZAG-situation: a sign before the title or subtitle; B) PI -situation: a sign before the beginning of a new paragraph; C) FIN-situation: a sign inside a paragraph surrounded by other signs or symbols that form a decisive situation together with it.

The choice is obvious: the title cannot tear the sentence into parts; therefore, the ability to formally identify the headings in the text will allow you to accurately fix the end of the sentences preceding them.

Variant B is also quite transparent but requires some detail. The sentences consisting of direct speech and the author's introductory words preceding it are often placed on different lines of the printed text. In such cases, the author's introductory words end with a colon, and direct speech is made out with a paragraph indentation (PI), followed by a dash. For example: *She suddenly felt insulted and said coldly: "We need to break up for a while, otherwise we can get bored (A.P. Chekhov. Hopalong).*

The presence of PI, thus, signals the end of the sentence only if the new line does not begin with a dash. If PI is followed by a dash, and the previous line ends with a colon, then PI is the situation <colon+PI+dash+PB>, where PB is a capital letter, is not decisive. If there is no colon at the end of the line preceding PI, then the situation <PI+dash+PB> is decisive.

Less often in fiction and prose texts, there is such a variant when quotation marks are used after PI, and even less often - both dashes and quotation marks; therefore, in the end, PI–a more general situation turns out to be decisive, namely: <– (colon)+PI>, where "–" is the sign of logical negation.

Variants B is the most difficult, since the identification of the FIN situation is carried out inside the paragraph and, therefore, can be based only on the analysis of the relative position of signs and symbols around the potential center of the decisive situation. These centers, as we have already noted, are formed by signs ".", "?", "!". Let's denote the set of such punctuation marks by Z, then: Z={z1, z2, z3} and z1 = ".", z2 = "?", z3 = "!".

Any FIN situation is considered as a sequence of three simpler situations F, I, and N, i.e. FIN=(F+ I+N). Let's look at these situations in more detail.

1. The F-situation (the end of the sentence) is fixed in the presence of one of the elements of the set F=Z+T, where T= {empty, quotes, closing parenthesis}. Otherwise: F= {Z, Z", Z)}, which corresponds to any set of characters essential for the formal identification of the "end of the sentence" situation in the text.

2. The I-situation is determined by the characters that are used to fill in the interval between two sentences located inside the same paragraph: I= {i1, i2} and i1=space, i2=dash.

3. The N-situation is fixed in the presence of one of the elements of the set N=P+E, where P= {empty, quotes, opening parenthesis}, and E is the set of all uppercase letters. Otherwise: N= {E, "E, (E}, which corresponds to the character set used to graphically represent the beginning of any sentence in the middle of a paragraph.

Here are two examples explaining the introduced formalisms. *He used to come to our village; I happen to be his nephew. – What, brother, Vasya, – he will say, – come, brother, spend the night with me!" (I.S. Turgenev. Three meetings).*

Here the FIN situation is formed as follows: F="dot", I=i2-"dash", P="quotes", E="H". Hence, FIN=<dot=dash+quotes+H>.

Yes, to everything in addition, except for the shame, to bring the hated wife into the house! (Because you hate me, I know that!) (F.M. Dostoevsky. Idiot).

In this example, FIN=<! + space + (+N>.

Table 1 formalizes the rules for searching for the "end of sentence" situation, where situations F, I, and N are located vertically, and their possible meanings are located horizontally.

ISSN-p 2306-7365
ISSN-e 2664-0686

Table 1 – Rules for identifying the "end of sentence" situation in the middle of a paragraph

F-situation <end of the sentence>	Z	Z»	Z)
I-situation space <interval> dash	+ + + + +	+ + + + +	+ + - + -
S-situation <beginning of the sentence>	E «E (E	E «E (E	E «E (E)

Table 1 the records 18 possible meanings of FIN- situations. The "+" sign marks such sequences of characters that, actually occurring in literary texts, allow us to confidently recognize the "end of sentence" situation inside any paragraph of the text. Here, the "–" sign marks those situations that practically do not occur in real texts; they, however, also refer to decisive FIN situations.

The potential possibilities of graphic means of the end–beginning of a sentence are, as we see, combinatorially redundant: only 70% of these possibilities are used in the graphics of the Russian language. More generally, it should be said that the punctuation combinatorial redundancy of the graphic means of the Russian language, revealed during the development of computer structuring rules, is another example of the redundancy of the expressive means of the language as a whole.

Let us summarize some results. Considering the variants for fixing the "end of the sentence" situation in the text and having at our disposal only visual means of writing, i.e. letters of the alphabet and punctuation marks, we identified three possible cases: the "end of sentence" situation is fixed before the heading (ZAG-situation), before the beginning of the paragraph (PI-situation) and inside the paragraph (FIN-situation). Each of these cases can be formalized: it is only necessary to provide a search simultaneously in two adjacent lines of the analyzed text, first in an external cycle, and then – if a potential center of the decisive situation is identified – and an internal search.

Table 1 records 18 possible meanings of FIN situations. The "+" sign marks such sequences of characters that, actually occurring in literary texts, allow us to confidently recognize the "end of sentence" situation inside any paragraph of the text. The "–" sign marks those situations that practically do not occur in real texts; they, also refer to decisive FIN situations.

The potential possibilities of graphic means of the end–beginning of a sentence are, as we see, combinatorially redundant: only 70% of these possibilities are used in the graphics of the Russian language. More generally, it should be said that the punctuation combinatorial redundancy of the graphic means of the Russian language, revealed during the development of computer structuring rules, is another example of the redundancy of the expressive means of the language as a whole.

Let us summarize some results. Considering the variants for fixing the "end of the sentence" situation in the text and having at our disposal only visual means of writing, i.e. letters of the alphabet and punctuation marks, we identified three possible cases: the "end of sentence" situation is fixed before the heading (ZAG-situation), before the beginning of the paragraph (PI-situation) and inside the paragraph (FIN-situation). Each of these cases can be formalized: it is only necessary to provide a search simultaneously in two adjacent lines of the analyzed text, first in an external cycle, and then – if a potential center of the decisive situation is identified – and an internal search.

The analysis of the punctuation system of the Russian language and the techniques of individual punctuation, as the development of computer rules for the formal allocation of text units based on reference, features allowed us to clarify the traditional idea of the sentence and give it a new operational interpretation. Implementation of the formal punctuation method of text division in the DISSKOTE dialogue computer system and analysis of the structuring of prose texts by M.Y. Lermontov, A.P. Chekhov, M.A. Sholokhov, A.S. Serafimovich, and other writers confirmed

ISSN-p 2306-7365
ISSN-e 2664-0686

the correctness of the chosen approach to solving the problem of allocating text units and the high linguistic reliability of the structuring method.

The homogeneity of the studied objects is a significant requirement of linguostatistical and stylometric methods of text research. This principle becomes of central importance in the study of fiction, since artistry presupposes, first of all, diversity, imagery, and breadth of use of figurative and expressive linguistic means.

The originality of the style of a work of art is manifested each time in that particular "image of the author" or "image of the narrator", which is formed by the unity of specific lexical, syntactic, intonational, and other means of language. It is the speech "image of the author" that determines the entire tone of the narrative, and the peculiarity of the individual author's style of the work of art.

The combination of two principles of depicting reality – from the outside and the inside, from the position of the author and the characters – is realized in writing by a complex interweaving and spatial interaction of two types of written speech: the authors and someone else's. There is no doubt that the author's speech is the basis of the speech's "image of the author", therefore, if we consider the problem of homogeneity of units of a literary text, not in general, but concerning any particular feature, it is obvious - this is the opposition "the author's speech/someone else's speech".

Lexical, syntactic, stylistic differences, and often sharp differences, between the author's and someone else's speech lead, for example, to the fact that direct speech is difficult to analyze using existing language models. At the same time, a comparison of direct speech in different works and its comparison with a speech in live communication can give very informative results when studying the influence of the individual author's manner of writing on the form of content.

So, the stylistic study of fiction should be preceded by the differentiation and stratification of speech material into two categories – the author's speech and someone else's speech. Someone else's speech includes direct speech and other types of non-author narration. Having at our disposal only formal means – punctuation marks and symbols of the graphic system of the language, we attempted to find such visual means that would allow us to distinguish text units automatically by types of written speech. Our stratification algorithm is based on a strict description of the means of registration of direct speech and its spatial arrangement on the pages of printed publications. The main units of the text, in this case, are a paragraph and a sentence, as for word usage, they are part of sentences and can participate in research as units of the lexical level of the author's or non-author's narrative.

In the stratification algorithm, the main structural element of the text is considered to be a paragraph: all sentences of a paragraph, isolated from the text one by one by the structuring program, are analyzed for the presence of extraneous speech in them; the decision to assign one or another paragraph to one of the types of written speech is made after the analysis of all sentences of that paragraph is completed.

The construction of a text with foreign speech can be done in two ways:

1) Another person's speech is reproduced in the form of separate replicas of the participants in the conversation; in the text, these replicas begin with a new line;

2) Another person's speech is included in the author's narrative paragraph; in these cases, another person's speech is usually enclosed in quotation marks.

The construction of the layering algorithm for the first variant does not present any particular difficulties: The situation at the beginning of the line and the absence of a colon at the end of the preceding line are necessary and sufficient conditions for establishing the beginning of a paragraph of direct speech. The same is true, of course, in the case of <PI+quotes>.

The second variant uses quotation marks, and if these quotation marks did not serve as a means of formalizing borrowings, the stratification algorithms would be trivial here as well.

Cf. *Petushkov started up, and the soldier stretched out, wished him "good health" and handed him a large envelope sealed with a government seal* (I.S. Turgenev. Petushkov).

ISSN-p 2306-7365
ISSN-e 2664-0686

This sentence, of course, should not be assigned to sentences with direct speech; therefore, we must algorithmically distinguish such cases and assign them to the author's speech.

There is only one way to solve this problem: analyzing and formalizing the rules for the graphic design of direct speech by quotation marks. Another, at first sight, more promising solution – clarification and formalization of the peculiarities of the design of sentences with borrowings – cannot be algorithmized. There is only one reason for this – the absence of restrictions on the use of quotations and other borrowings in the text, which makes the task hopeless in searching for characteristic punctuation features.

To describe the rules for recognizing direct speech in quotation marks, we will use the formalisms introduced earlier and some new ones: By S we will denote some punctuation situation in the middle of a sentence, by the sign V we denote some punctuation-sign situation in the middle of a sentence, with the sign V we will denote the logical operation "or", and with the special word nil – the absence of quotation marks in N- or F-situations. Then the rules for fixing direct speech framed with quotation marks can be reduced to four cases presented in Table 2.

Variant 1 implements the case when direct speech is torn into two parts by the author's words. Thus, the S-situation should contain two dashes separated by some text:

"Maybe, I thought, that's why you loved me: joys are forgotten, but sorrows are never..." (M.Y. Lermontov. The hero of our time).

Table 2 – Rules for detecting direct speech (when using quotation marks)

F-situation <end of the sentence>	<Z»>V<«Z>		nil	
S -situation <beginning of the sentence>	<«E>	nil	<«E>	
M- situation <middle of the sentence>	<–…–>	<:«E>	<», –>	<Z», –>
Variant	1	2	3	4

Variant 2 implements the case when direct speech is at the end of a sentence and follows the words of the author:

The headman first nimbly jumped off his horse, bowed to the master from the waist, and said: "Hello, Father Arkady Pavlovich." (I.S. Turgenev. Burmistr).

Variants 3 and 4 implement cases when direct speech is at the beginning of a sentence, and the author's words follow it. Cf. variant 4:

"Hurry, hurry to the city for a doctor!" – shouted Vladimir (A.S. Pushkin. Dubrovsky).

The same rules for identifying direct speech are also applicable in more complex cases, namely when direct speech consists of not one, but several sentences. While determining the type of the paragraph, those sentences that are situated inside quotation marks are considered as a whole, and punctuation marks ".", "?" and "!" are not taken into account. Cf. variants 1 and 3:

"Maybe we'll never see each other again, – he told me. – Before parting, I would like to explain myself to you". (A.S. Pushkin. Shot).

"Spit on him, a woman. There would be a neck, but there would be a yoke", one advised with undisguised regret (M.A. Sholokhov. A Quiet Don).

The formal rules developed and presented above for identifying direct speech cover the three most common variants of constructing sentences with the author's output words, namely when the author's words:

1) precede direct speech;
2) follow direct speech;
3) are included in direct speech, dividing it into parts.

ISSN-p 2306-7365
ISSN-e 2664-0686

There is another, rarer variant when the author's introductory words include direct speech. However, these variants are also formalized by some complications of the decisive rules:

To my question: "Is the old caretaker alive?" no one could give me a satisfactory answer (A.S. Pushkin. Stationmaster).

And only when he whispered: "Mom! Mom!" – it seemed to be easier for him... (A.P. Chekhov. Steppe).

These examples are quite enough to make sure of the following: the constructions we are considering contain the author's introductory words, which are to the left of direct speech and are separated from it by a colon. Therefore, the search procedure for sentences with the author's introductory words (variants 5) looks like this:

(S1=<: «E>) & ((S2=< «, >) V (S3=<Z»>)),

where the & sign indicates the logical operation "and", and situation S1 must precede the appearance of situation S2 or situation S3 in the text. Therefore, for the latter variants, the rule of formalized detection of direct speech has the form (F=nil) & (N=nil) & (S1) & (S2 V S3).

We have thus formulated and formalized the rules that allow us to create algorithms for stratifying prose texts by types of written language. Again, the main units of the text are a paragraph and a sentence: each sentence refers to the variant of written speech to which the whole paragraph belongs.

The main variants of written speech isolated by the stratification procedure are the author's speech, direct speech, "mixed" speech, and "nested" direct speech. The last three are excerpts from another author's speech:

a) direct speech is dialogic speech, without inclusions of the author's speech;

b) "mixed" speech includes direct speech within a paragraph of the author's speech;

c) "nested" direct speech includes direct speech within a paragraph of direct speech. Example of "nested" speech:

"Don't get excited, though. I somehow entered into a conversation with her at the well by chance; her third word was: "Who is this gentleman who has..." (M.Y. Lermontov. Princess Mary).

All three variants of someone else's speech characterize, special stylistic techniques that each author owns to one or another level and uses in his works, techniques that should be studied both particularly and in their entity, integrating into an individual author and genre-stylistic set of methods, techniques, means of writing.

Table 3 – Stylistic features of texts

Features	**Number of paragraphs**	**Number of sentences**
	Текст ½	Текст 1/2
Author's speech .... Someone else's speech direct .... mixed ..... embedded	21/22 18/42 9/34 6/7 3/1	51/31 117/117 39/92 34/22 44/3
Total::	39/64	168/158

A small fragment of a comparative analysis of two stories (Table 3) demonstrates some stylistic features of the author's manner of writing by A.S. Serafimovich (the story "Politkom" – text 1) and S.P. Podyachev (the story "Understood" – text);

2) Structuring and layering texts are performed automatically by the DISSKOTE system.

The volume of these two stories is almost the same (7 pages of text each), while other features can, without a doubt, serve as a starting point for deep stylistic research.

ISSN-p 2306-7365
ISSN-e 2664-0686

Conclusion

The study of the interaction of different speech types within a single whole and the use of various concepts, methods, and algorithms for this purpose constitutes one of text linguistics' central and most urgent problems.

The objective of computational linguistics in this inexhaustible creative search is to equip the researcher with all the necessary material to prepare counting data and offer a wide range of computational procedures that can collectively test hypotheses, to confirm or refute increasingly subtle and profound philological observations. This vision of goals and objectives underlies the development and creation of the LINDA computer system – an expert philological framework, multifunctional in its purpose and "friendly" to its users.

So, we have completed the discussion of methods and specific variants for the formation of homogeneous counting units of the text. This seemingly auxiliary, preparatory stage should be planned and carried out with special care and scientific justification, because no, even the subtlest, linguistic-stylistic procedures guarantee the accuracy and reliability of the results when processing incorrectly formed (set, selected) source counting units of the text. There is also no doubt that the method of structuring and stratification of texts, developed for the formation of homogeneous sets of texts of fiction, should, firstly, be improved taking into account the practical experience of its use and, secondly, expand to extend it to other genre-stylistic variants of written speech, for example, poetic texts and dramatic works. This will require the development of subtler algorithms, taking into account the need to solve the "hyphen–hyphenation" problem, the analysis of standard abbreviations, the author's remarks, and various genre-stylistic inclusions.

The authors' contribution consists in choosing a topic, conducting experimental work, and interpreting the obtained results. In particular, the authors have developed criteria for stylometric identification of the author of the text, using modern computer-specialized software; the main trends in the development of the subject area "Stylometric identification of the author of the text" have been identified; the methodological and theoretical-methodological problems of stylometric identification of the author of the text are formulated and analyzed in detail.

BIBLIOGRAPHY

1. Социальные науки и образование в условиях становления электронноцифровой цивилизации / Научно-практическая конференция. – М.; СПб.: Нестор-История, 2020. – 152 с.
2. Бортников В.И. Лингвистический анализ текста: учебно-методическое пособие / под общ. ред. О.В. Обвинцевой. – Екатеринбург: Изд-во Урал. ун-та, 2020. – 112 с.
3. Попов Е.В., Лагутина Н.С. Определение стилометрических характеристик коротких текстов и их применение в задачах классификации // Сборник научных статей. – Ярославль: Ярославский государственный университет им. П.Г. Демидова, 2020. – №12. – С. 254–261.
4. Мартыненко Г.Я. Стилометрия: возникновение в становление в контексте междисциплинарного взаимодействия // Структурная и прикладная лингвистика: межвуз. сб. / под ред. А.С. Герда и И.С. Николаева. – СПб.: Изд-во С.-Петерб. ун-та, 2015. – Вып. 11. – С. 9–28.
5. Сарекенова Қ.Қ., Меліс А.М., Тойбекова С.Р. Филология мамандығы бойынша білім алушыларды компьютерлік технология бағытында оқыту – заман талабы // Л.Н. Гумилев атындағы Еуразия ұлттық университетінің Хабаршысы. Филология сериясы. – 2018. – №2 (123). – Б. 166–172.
6. Жубантаева Ж. Корпустық лингвистика. [Electronic Resource]. URL: https://www.academia.edu/39122192 (қаралған күні: 15.10.2023)
7. Маханова З.А., Қожабекова П.А., Сейтжаппар М.А., Сабит Н.Е. Қазақ тілінің автоматтандырылған маркерлік корпусын әзірлеу // ҚазҰТЗУ хабаршысы. – 2021. – №1. – Б. 36–39.

ISSN-p 2306-7365
ISSN-e 2664-0686

8. Langlois J. When Linguistics meets computer science: Stylometry and professional discourse // Original Research Journal. Training Language and Culture. More than Meets the Eye: A Closer Look at Professional Discourse. – 2021. – Issue 2. – №5. – P. 51–61.
9. Wright D., May A. Identifying idiolect in forensic authorship attribution: an n-gram textbite approach. Language and Law // Linguagem e Direito. – 2014. – №1(1). – P. 37–69.
10. Galyashina E.I. Forensic linguistics in Russia: the current situation and new challenges // Theory and practice of forensic expertise. – 2018. – Vol. 13. – №4. – P. 28–37.
11. Nikishin V.D. Criteria of Extremist Speech Acts: Forensic Linguistic Diagnostic Complexes // European Journal of Social & Behavioural Sciences. – 2021. – №30(2). – P. 3394–3408. DOI:10.15405/ejsbs.296.
12. Чернявская В.Е. Дискурсивный анализ и корпусные методы: необходимое доказательное звено? Объяснительные возможности качественного и количественного подходов // Вопросы когнитивной лингвистики. – 2018. – №2 (55). – С. 31–37. DOI: 10.20916/1812-3228-2018-2-31-37
13. Burrows J. “Delta”: a measure of stylistic difference and a guide to likely authorship // Literary and Linguistic Computing. – 2002. – Vol. 17 (3). – P. 267–287.
14. Demsar J. Statistical comparisons of classifiers over multiple data sets // Journal of Machine Learning Research. – 2006. – №7. – P. 1–30.
15. Romero-Barranco J., Rodríguez-Abruñeiras P. Current trends in Corpus Linguistics and textual variation // Research in Corpus Linguistics. – 2021. – №9(2). – P. i-xiii. https://doi.org/10.32714/ricl.09.02.01
16. Desagulier G. Corpus linguistics and statistics with R. Introduction to quantitative methods in linguistics (quantitative methods in the humanities and social sciences). – Springer International Publishing Springer, 2017. – 353 p.

REFERENCES

1. Socialnye nauki i obrazovanie v usloviah stanovlenia elektronnocifrovoi civilizacii [Social sciences and education in the context of the formation of an electronic digital civilization] / Nauchno-prakticheskaia konferencia. – M.; SPb.: Nestor-Istoria, 2020. – 152 s. [In Russian]
2. Bortnikov V.I. Lingvisticheskiy analiz teksta [Linguistic analysis of the text]: uchebno-metodicheskoe posobie / pod obsh. red. O.V. Obvincevoi. – Ekaterinburg: Izd-vo Ural. un-ta, 2020. – 112 s. [In Russian]
3. Popov E.V., Lagutina N.S. Opredelenie stilometricheskih harakteristik korotkih tekstov i ih primenenie v zadachah klassifikacii [Determination of stoichiometric characteristics of short texts and their application in classification tasks] // Sbornik nauchnyh statei. – Iaroslavl: Iaroslavskiy gosudarstvennyi universitet im. P.G. Demidova, 2020. – №12. – С. 254–261. [In Russian]
4. Martynenko G.Ia. Stilometria: vozniknovenie v stanovlenie v kontekste mejdiciplinarnogo vzaimodeistvia [Stylometry: emergence into formation in the context of interdisciplinary interaction] // Strukturnaia i prikladnaia lingvistika: mejvuz. sb. / pod red. A.S. Gerda i I.S. Nikolaeva. – SPb.: Izd-vo S.-Peterb. un-ta, 2015. – Vyp. 11. – S. 9–28. [In Russian]
5. Sarekenova Q.Q., Melis A.M., Toibekova S.R. Filologia mamandygy boiynsha bilim alushylardy kompiuterlik tehnologia bagytynda oqytu – zaman talaby [Training of students in the specialty philology in the direction of computer technology is a modern requirement] // L.N. Gumilev atyndagy Eurazia ulttyq universitetinin Habarshysy. Filologia seriasy. – 2018. – №2 (123). – B. 166–172. [in Kazakh]
6. Jubantaeva J. Korpustyq lingvistika [Corpus linguistics]. [Electronic Resource]. URL: https://www.academia.edu/39122192 (date of access: 15.10.2023) [in Kazakh]
7. Mahanova Z.A., Qojabekova P.A., Seitjappar M.A., Sabit N.E. Qazaq tilinin avtomattandyrylgan markerlik korpusyn azirleu [Development of an automated marker body of the Kazakh language] // QazUTZU habarshysy. – 2021. – №1. – B. 36–39. [in Kazakh]
8. Langlois J. When Linguistics meets computer science: Stylometry and professional discourse // Original Research Journal. Training Language and Culture. More than Meets the Eye: A Closer Look at Professional Discourse. – 2021. – Issue 2. – №5. – P. 51–61.

ISSN-p 2306-7365
ISSN-e 2664-0686

9. Wright D., May A. Identifying idiolect in forensic authorship attribution: an n-gram textbite approach. Language and Law // Linguagem e Direito. – 2014. – №1(1). – P. 37–69.
10. Galyashina E.I. Forensic linguistics in Russia: the current situation and new challenges // Theory and practice of forensic expertise. – 2018. – Vol. 13. – №4. – P. 28–37.
11. Nikishin V.D. Criteria of Extremist Speech Acts: Forensic Linguistic Diagnostic Complexes // European Journal of Social & Behavioural Sciences. – 2021. – №30(2). – P. 3394–3408. DOI:10.15405/ejsbs.296.
12. Cherniavskaia V.E. Diskursivnyi analiz i korpusnye metody: neobhodimoe dokazatelnoe zveno? Obiasnitelnye vozmojnosti kachestvennogo i kolichestvennogo podhodov [Discursive analysis and corpus methods: a necessary evidentiary link? Explanatory possibilities of qualitative and quantitative approaches] // Voprosy kognitivnoi lingvistiki. – 2018. – №2 (55). – S. 31–37. DOI: 10.20916/1812-3228-2018-2-31-37 [In Russian]
13. Burrows J. "Delta": a measure of stylistic difference and a guide to likely authorship // Literary and Linguistic Computing. – 2002. – Vol. 17 (3). – P. 267–287.
14. Demsar J. Statistical comparisons of classifiers over multiple data sets // Journal of Machine Learning Research. – 2006. – №7. – P. 1–30.
15. Romero-Barranco J., Rodríguez-Abruñeiras P. Current trends in Corpus Linguistics and textual variation // Research in Corpus Linguistics. – 2021. – №9(2). – P. i-xiii. https://doi.org/10.32714/ricl.09.02.01
16. Desagulier G. Corpus linguistics and statistics with R. Introduction to quantitative methods in linguistics (quantitative methods in the humanities and social sciences). – Springer International Publishing Springer, 2017. – 353 p.