COMPUTATIONAL MATHEMATICS

MSC 60H30, 34K50, 34M99 **DOI: 10.14529/jcem260201**

CONTENT ANALYSIS OF S.L. SOBOLEV'S SCIENTIFIC WORKS BASED ON COMPUTATIONAL LINGUISTICS METHODS

I. B. Krasnov[1]**,** uzpgo@mail.ru,
E. A. Krasnova[1]**,** krasnovae@mail.ru
[1]South Ural State University, Chelyabinsk, Russian Federation

The article is devoted to identifying and systematizing the quantitative morphological and syntactic features of the scientific style of Academician S.L. Sobolev in the field of equations of mathematical physics. Based on the developed software complex SciText-Analyzer in the Python language using the Natasha library, a comprehensive analysis of a representative text corpus of the scientist's scientific works was carried out in comparison with the works of his contemporaries. A stable diachronic dynamic in the evolution of S.L. Sobolev's idiostyle was revealed, which is manifested in a consistent growth of the conceptual density of the text with simultaneous syntactic compression and compositional economy in the late period of his creativity.

Keywords: content analysis; mathematical discourse; stylometry; Natasha library; computational linguistics; equations of mathematical physics.

Introduction

Today, as digital methods actively enter the humanitarian and historical-scientific fields, computational linguistics and stylometry are becoming essential tools for studying the academic heritage of outstanding scientists [2, 3]. In this regard, the figure of the prominent Soviet mathematician, Academician Sergey Lvovich Sobolev (1908–1989), represents a special research interest [9]. His fundamental works fundamentally changed the ideas on the formulation and methods for solving boundary value problems for partial differential equations, determined the development of the theory of generalized functions, and significantly influenced modern functional analysis [10, 12].

S.L. Sobolev's scientific biography is distinguished not only by the scale of his discoveries, but also by a pronounced internal dynamic, which is traditionally divided into three key periods: early (problems of elasticity theory and wave propagation in the 1930s), Moscow (intensive development of functional analysis in the 1940s–1950s), and Siberian (research in the field of computational mathematics and the theory of cubature formulas after 1958) [7, 8, 11]. Each of these stages has its own substantial specificity, which was inevitably projected onto the linguistic structure of the texts.

S.L. Sobolev's scientific prose possesses a unique, intuitively recognizable balance between symbolic and verbal modes of presentation: the introduction of complex abstractions required a descriptive language in which the mathematical formula does not displace verbal reasoning, but organically interacts with it [9, 10]. However, until now, the

scientist's scientific writing itself was considered mainly either in a historical-biographical or in a highly specialized mathematical context. Quantitative text content analysis, based on natural language processing (NLP) algorithms, allows a transition from subjective evaluations to a rigorous description of the stable features of academic speech in verifiable and reproducible indicators [1, 13].

The main problem of the research is that, despite all the study of S.L. Sobolev's scientific heritage, the morphological and syntactic features of the Russian-language mathematical discourse of the mid and second half of the 20th century are still described insufficiently systematically. Until recently, there were no specialized software solutions oriented toward batch processing and profiling of texts of the Soviet mathematical school in a diachronic aspect. This work is aimed at overcoming this gap by formalizing and providing a linguostatistical description of the stylistic features of scientific texts on equations of mathematical physics using the tools developed by the authors.

Partially, the approaches to the automated processing of mathematical texts and the primary results of the authors' research were successfully approved within the framework of the IX All-Russian Scientific and Practical Conference "South Ural Youth School on Mathematical Modelin" [4], which served as the basis for creating the comprehensive analytical pipeline presented in this article.

1. Methods and Software Complex Architecture

To perform an objective quantitative analysis of the Soviet mathematical school texts, the authors apply a specialized software complex called SciText-Analyzer, the software implementation of which was carried out by I. B. Krasnov. The program is implemented as a standalone module in the Python language (the final version size is 16,576 bytes) and is based on modern components of the open-source Natasha NLP library, including the Segmenter, NewsEmbedding, and NewsMorphTagger modules [13].

The text processing procedure inside SciText-Analyzer is organized as a strict, sequential analytical pipeline consisting of the following key stages:

1. *OCR Text Normalization:* executed via the `normalize_text()` function. It automatically eliminates recognition artifacts, unifies various types of dashes and spaces, handles line-breaking hyphens, and removes page layout markers. This stage is methodologically essential to guarantee the homogeneity and comparability of data extracted from archival publications of the 1950s–1960s.
2. *Representative Fragment Extraction:* controlled by the `extract_by_occurrence()` function. Since the original scientific sources significantly vary in total volume and structure, this module extracts homogeneous text blocks of a fixed length based on semantic markers.
3. *Tokenization and Morphological Parsing:* the normalized text string is segmented into sentences and individual tokens, after which each wordform is assigned a corresponding morphological tag (part of speech). At this stage, formulaic elements are filtered using regular expressions (`MATHTOKEN_RE`) to exclude distortion of pure linguistic statistics.
4. *Linguistic Metric Evaluation:* batch calculation of stylometric coefficients is carried out on the basis of the obtained morphological and syntactic parsing.

To formalize the core properties of the academic style, the system evaluates three base stylometric coefficients:

$$K_n = \frac{N_{\text{NOUN}} + N_{\text{PROPN}}}{N_{\text{VERB}} + N_{\text{AUX}}} \tag{1}$$

where K_n is the nominalization coefficient representing the ratio of nominal parts of speech (including proper nouns) to verbal ones (including auxiliary verbs). This index captures the conceptual density and degree of abstraction of the discourse.

$$I_s = \frac{N_{\text{commas}}}{N_{\text{sentences}}} \tag{2}$$

where I_s is the syntactic complexity index calculated as the ratio of the total number of commas to the total number of sentences in the analyzed fragment. It measures the degree of structural branching and syntactic elaboration of the text.

$$I_{st} = \frac{N_{\text{declarative}}}{N_{\text{sentences}}} \times 100\% \tag{3}$$

where I_{st} is the academic strictness index showing the percentage of standard declarative constructions in the overall syntactic structure, which characterizes the emotional neutrality of the scientific statement.

Additionally, the SciText-Analyzer software complex monitors auxiliary properties such as the average sentence length and the share of mathematical formula tokens, which allows the formation of a multidimensional stylistic profile of the author. The primary approaches to constructing this analytical pipeline were previously presented by the authors in [4].

2. Experimental Study and Linguostatistical Analysis

The final computational stage of the computer-linguistic experiment was carried out not on the full texts of multi-volume monographs, but on representative and carefully verified fragments of a comparable volume, isolated from a pre-normalized text array. This approach is due to the need to strictly ensure the homogeneity of the studied subject area, as well as natural limitations associated with the uneven quality of optical character recognition (OCR) of archival publications from the 1950s–1960s.

The final linguostatistical analysis included three chronological fragments of S.L. Sobolev's fundamental works from different periods ("Equations of Mathematical Physics", 1954, early period, 33,787 characters; "Introduction to the Theory of Cubature Formulas", § 1, 1974, middle period, 59,998 characters; "Some Applications of Functional Analysis in Mathematical Physics", section 17, 1988, late period, 60,000 characters). As a strict control layer (a synchronous model of the academic discourse of contemporaries), representative texts of Academician M.A. Lavrentiev ("Variational Method in Boundary Value Problems for Systems of Elliptic Equations", 1962, 60,000 characters) and Professor I.G. Petrovsky ("Lectures on Partial Differential Equations", fragments of paragraphs, 1961, 42,245 characters) were used. The total volume of the formed computational corpus was 256,030 characters and 44,118 word tokens.

The main results of the automated calculation of linguistic indices by means of the developed SciText-Analyzer complex are presented in Table 1.

Table 1

Metrics for individual documents and research groups of the corpus

Document (Research fragment)	K_n	I_s	I_{st}, %	Average sentence length	Formula token share, %
Sobolev, (1954, early period)	3.3044	2.4877	84.9123	19.6035	16.5384
Sobolev, (1974, middle period)	3.5419	1.7083	94.5076	19.2822	12.6019
Sobolev, (1988,late period)	4.2303	1.6272	91.1661	18.2297	14.8478
Lavrentiev, (1962, control)	3.7545	2.5082	86.1413	29.2935	17.5417
Petrovsky, (1961, control)	3.9077	3.3299	94.5578	24.6667	12.2173
Average for the "Sobolev" group	3.6922	1.9411	90.1953	19.0385	14.6627
Average for the "Contemporaries" group	3.8311	2.9191	90.3496	26.9801	14.8795

The obtained empirical data allow us to adjust preliminary qualitative hypotheses and reveal deep regularities in the organization of mathematical prose from the mid-20th century. As Table 1 demonstrates, the average nominalization coefficient (K_n) in S.L. Sobolev's authorial corpus is stably high and confidently exceeds the standard target for scientific and technical texts ($K_n > 2.5$), capturing the predominance of nominal constructions over verbal ones and a high degree of "objectification" of thought. However, when comparing group averages, it is found that the overall nominalization rate for S.L. Sobolev is slightly lower than in the control group of his contemporaries (3.6922 versus 3.8311). This proves that the specificity of his individual style is determined not by an absolute isolated maximum of substantivity, but by a unique multidimensional configuration of features.

The main inter-author difference is localized at the syntactic level. S.L. Sobolev's texts demonstrate pronounced syntactic compactness and compositional economy. The average syntactic complexity index (I_s) in his corpus is 1.9411, while for his contemporaries this parameter reaches 2.9191, which indicates a much more developed and punctuationally complex structure of utterances for Lavrentiev and Petrovsky. This conclusion is also strictly confirmed by the average sentence length metric: 19.0385 words for Sobolev versus 26.9801 words in the control group. The mathematical discourse of his contemporaries turns out to be syntactically more branched, whereas S.L. Sobolev's style is focused on a concise, rigorous, and brief presentation of argumentative material while maintaining high logical discipline.

The captured diachronic dynamic of the evolution of S.L. Sobolev's idiostyle over more than thirty years of his scientific creativity is of particular scientific interest (see Table 1). The nominalization coefficient increases consistently and linearly — from 3.3044 in early works to an extreme value of 4.2303 in the late period. This indicates a continuous

strengthening of the abstractness of the presentation, a rapid growth in the density of the conceptual apparatus, and a concentration of terminological vocabulary toward the end of the scientist's life.

In parallel with the growth of abstractness, the reverse process occurs at the syntax level: the complexity index I_s monotonically decreases from 2.4877 to 1.6272, and the average sentence length drops from 19.60 to 18.23 words. Consequently, the evolution of the scientist's style went along the path of syntactic compression of the structure: late texts become as concise as possible, freeing themselves from redundant subordinate and introductory constructions, turning into a concentrated stream of mathematical concepts.

At the same time, the academic strictness index (I_{st}), which reflects the share of neutral declarative sentences, demonstrates a non-linear, dome-shaped character. It reaches its absolute maximum in the middle (Moscow) period of creativity — 94.5076%. This stage, which coincided with the rise of the domestic school of the theory of partial differential equations and functional spaces, can be considered the phase of highest compositional balance, academic impartiality, and refined rigor of presentation in Sobolev's discourse. In the late period, a slight decrease in I_{st} to 91.1661% is observed, which is associated with the appearance of elements of a popular science and memorial-essay character in the texts.

Additionally, the analyzed parameter of formula saturation (the share of specialized tokens containing numbers, operation signs, or variables) showed a striking convergence of group averages: 14.66% for Sobolev and 14.87% for his contemporaries. This allows us to state that the specificity and uniqueness of a mathematical text of the mid-century are determined not by the number of mathematical formulas themselves, but by the unique linguistic way they are integrated into the verbal fabric and syntax of an argumentative utterance.

The practical significance of the work is determined by the fact that the created algorithmic pipeline and the SciText-Analyzer software complex can be directly scaled for automated profiling, stylistic comparison, and attribution of other Russian-language academic texts of a physical-mathematical and natural-science profile. The research clearly demonstrates the powerful analytical potential of digital humanities (Digital Humanities) in studying the history of science and the evolution of the language of scientific knowledge.

As prospects for further work, the authors highlight:

- significant expansion of the computational corpus by attracting the works of a wider circle of representatives of the Soviet mathematical school;
- deepening syntactic parsing for a detailed description of the types of logical-grammatical connections in proof structures;
- expanding the system of metrics with semantic and terminological indicators for constructing semantic-thematic clusters of scientific discourse.

References

1. Baranov A.N. [*Introduction to Applied Linguistics*], Moscow, Editorial URSS, 2001. – 360 p. (in Russian)
2. Belousov K.I. [*Theory and Practice of Corpus Linguistics*], Moscow, Aspekt Press, 2020. – 215 p. (in Russian)
3. Zakharov V.P. [*Corpus Linguistics*], St. Petersburg, SPbSU Publishing House, 2005. – 48 p. (in Russian)

4. Krasnov I.B., Krasnova E.A. [Statistical Content Analysis of Scientific Text Based on Python Algorithms] // *South Ural Youth School on Mathematical Modeling: Proceedings of the IX All-Russian Scientific and Practical Conference, June 15–16, 2026* / Ed. by E. V. Bychkov. – Chelyabinsk: Publishing Center of SUSU, 2026. – P. 74–77. (in Russian)
5. Lavrentiev M.A. [*Variational Method in Boundary Value Problems for Systems of Elliptic Equations*], Moscow, USSR Academy of Sciences Publishing House, 1962. – 136 p. (in Russian)
6. Petrovsky I.G. [*Lectures on Partial Differential Equations*], Moscow, GIFML, 1961. – 400 p. (in Russian)
7. Smirnov V.I. Course of Higher Mathematics. Vol. 4. *Equations of Mathematical Physics*, Moscow, Nauka, 1974. – 336 p. (in Russian)
8. Sobolev S.L. [*Introduction to the Theory of Cubature Formulas*], Moscow, Nauka, 1974. – 808 p. (in Russian)
9. Sobolev S.L. Selected Works. Volume I. *Equations of Mathematical Physics. Computational Mathematics and Cubature Formulas*, Novosibirsk, Publishing House of the Institute of Mathematics, 2003. – 576 p. (in Russian)
10. Sobolev S.L. [*Some Applications of Functional Analysis in Mathematical Physics*], Moscow, Nauka, 1988. – 333 p. (in Russian)
11. Sobolev S.L. [*Equations of Mathematical Physics*], Moscow, Gostekhizdat, 1954. – 444 p. (in Russian)
12. Sobolev Institute of Mathematics SB RAS [*Electronic resource*] // About S.L. Sobolev. – URL: http://nsc.ru_Sobolev_SL.htm (accessed: 28.05.2026). (in Russian)
13. Natasha: Python library for NLP on Russian language [*Electronic resource*]. – URL: https://github.com (accessed: 28.05.2026).

Ilia B. Krasnov, Bachelor of Mathematics, Department of Equations of Mathematical Physics, South Ural State University (Chelyabinsk, Russian Federation), uzpgo@mail.ru

Ekaterina A. Krasnova, PhD(Math), Associate Professor, Department of Mathematical and Computer Modelling, South Ural State University (Chelyabinsk, Russian Federation), krasnovaea@susu.ru

Received June 1, 2026

УДК 517.9 **DOI: 10.14529/jcem260201**

КОНТЕНТ-АНАЛИЗ НАУЧНЫХ РАБОТ С.Л. СОБОЛЕВА НА ОСНОВЕ МЕТОДОВ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ

И. Б. Краснов[1]***, Е. А. Краснова***[1]
[1]Южно-Уральский государственный университет, Челябинск, Российская Федерация

Статья посвящена выявлению и систематизации количественных морфологических и синтаксических особенностей научного стиля академика С.Л. Соболева в области уравнений математической физики. На основе разработанного программного комплекса SciText-Analyzer на языке Python с применением библиотеки Natasha проведен комплексный анализ репрезентативного текстового корпуса научных трудов ученого в сопоставлении с работами его современников. Выявлена устойчивая диахроническая динамика эволюции идиостиля С.Л. Соболева, выражающаяся в последовательном росте понятийной плотности текста при одновременном синтаксическом уплотнении и композиционной экономии в поздний период творчества.

Keywords: контент-анализ; математический дискурс; стилометрия; библиотека Natasha; компьютерная лингвистика; уравнения математической физики.

Литература

1. Баранов, А.Н. Введение в прикладную лингвистику / А.Н. Баранов. – М.: Эдиториал УРСС, 2001. – 360 с.
2. Белоусов, К.И. Теория и практика корпусной лингвистики / К.И. Белоусов. – М.: Аспект Пресс, 2020. – 215 с.
3. Захаров, В.П. Корпусная лингвистика / В.П. Захаров. – СПб: Изд-во СПбГУ, 2005. – 48 с.
4. Краснов, И.Б. Статистический контент-анализ научного текста на основе алгоритмов Python / И.Б. Краснов, Е.А. Краснова // Южно-Уральская молодежная школа по математическому моделированию : сборник трудов IX Всероссийской научно-практической конференции, Челябинск, 15–16 июня 2026 г. / под ред. Е.В. Бычкова. – Челябинск : Издательский центр ЮУрГУ, 2026. – С. 74–77.
5. Лаврентьев, М.А. Вариационный метод в краевых задачах для систем уравнений эллиптического типа / М.А. Лаврентьев. – М.: Изд-во АН СССР, 1962. – 136 с.
6. Петровский, И.Г. Лекции об уравнениях с частными производными // И.Г. Петровский. – М.: ГИФМЛ, 1961. – 400 с.
7. Смирнов, В.И. Курс высшей математики. Т. 4. Уравнения математической физики / В.И. Смирнов. – М.: Наука, 1974. – 336 с.
8. Соболев, С.Л. Введение в теорию кубатурных формул / С.Л. Соболев. – М.: Наука, 1974. – 808 с.
9. Соболев, С.Л. Избранные труды. Том I. Уравнения математической физики. Вычислительная математика и кубатурные формулы // С.Л. Соболев. – Новосибирск: Изд-во Ин-та математики, 2003. – 576 с.

10. Соболев, С.Л. Некоторые применения функционального анализа в математической физике / С.Л. Соболев. – М.: Наука, 1988. – 333 с.
11. Соболев, С.Л. Уравнения математической физики / С.Л. Соболев. – М.: Гостехиздат, 1954. – 444 с.
12. Институт математики им. С.Л. Соболева СО РАН [Электронный ресурс] // О С.Л. Соболеве. – URL: http://nsc.ru_Sobolev_SL.htm (дата обращения: 28.05.2026).
13. Natasha: Python library for NLP on Russian language [Электронный ресурс]. – URL: https://github.com (дата обращения: 28.05.2026).

Краснов Илья Борисович, бакалавр математики, кафедра уравнений математической физики, Южно-Уральский государственный университет (г. Челябинск, Российская Федерация), izrpgo@mail.ru

Краснова Екатерина Александровна, кандидат физико-математических наук, доцент, доцент кафедры математического и компьютерного моделирования, Южно-Уральский государственный университет (г. Челябинск, Российская Федерация), krasnovaea@susu.ru

Поступила в редакцию 1 июня 2026 г.