ANNEX 3

The authorial invariant in Russian literary texts. Its application: who was the real author of the "Quiet Don"?

By *V. P. Fomenko* and *T. G. Fomenko**

COMMENTARY BY A. T. FOMENKO

(Moscow, the Moscow State University, Department of Mathematics and Mechanics)

The readers are invited to turn their attention to the results of the research conducted by my parents Valentina Polikarpovna Fomenko and Timofei Grigoryevich Fomenko in 1974-1981 as cited below. The complete body of their work was first published in [METH1]:3 in 1996. Its brief version was published in 1983 as part of the compilation entitled ***The Methods of Quantitative Analysis as Applied to the Text of Narrative Sources***, Moscow, 1983, The USSR Academy of Sciences Institute of Soviet History, pages 86-109.

The main result of the present work is the discovery of the "authorial invariant" for literary texts in Russian. It allows for distinguishing between various authors and proves useful in solving plagiarism issues. The result stems from a certain general idea – the statistical analysis of volume functions for narrative texts. The volume functions were introduced in [f19]; several new empirico-statistical models of information analysis for narrative texts were also suggested in the same work. These ideas were developed in [f20] as well.

The present work has seemingly got little to do with the research concerning the basics of the ancient chronology. However, this material demonstrates just how empirico-statistical methods can be used for the solution of problems, which also go beyond the scope of chronology and pertain to neighbouring paradigms, such as determining the authorship of a written document. And since our analysis of written history is based primarily on empirico-statistical methods, we decided to familiarize the reader with this research – especially considering that the issue of authorship determination in modern and ancient literature is a most poignant one, and all new methods in this field may be of use.

(End of commentary)

1. INTRODUCTION. A BRIEF EXCURSUS INTO THE HISTORY OF THE PROBLEM

One often sees the issue of attributing literary work arise in literature, history and linguistics alike. Was a given work really written by a single author – Plato's dialogues, for instance? Are Shakespeare's plays all brainchildren of a single genius, or could several authors have written them, perhaps? Who stands behind the name of "Shakespeare"? This problem becomes especially vital when a suspicion of plagiarism arises.

Let us merely mention several approaches to the solution of such problems.

* T. G. Fomenko is a candidate of technical sciences and a specialist in the field of ore dressing; he wrote a number of books on the subject of beneficiation and floatation, and has been Department Head of the Ukrainian Ore Dressing Research Centre in Lugansk, Ukraine. V. P. Fomenko is a specialist in the field of the Russian language and literature.

The work of V. Fuchs, for instance ([f1]), tackles the issue of the authorship of several ancient texts based on the statistical analysis of various grammatical structures pertinent to their language.

A great deal of research was dedicated to the discovery of various quantitative characteristics, which allow to distinguish between different literary genres – poetry, drama, journalism and so on ([f2]).

An account of the attempt of using exact mathematical methods for solving the problem of plagiarism is given in [f10], for instance.

The problem of discovering authorial invariants was dealt with in a great amount of scientific literature. Thus, for instance, the regulating function word usage frequency in the language of various authors was studied, (the Russian equivalents of the preposition "in" and the particle "not" in particular, qv in [f4]). However, experimentation demonstrates that the use of the linguistic ranges of function word regulation does not allow for the discovery of steady authorial invariants per se. This was pointed out by the Academician A. A. Markov as early as 1916 ([f5]); he states that a large amount of samples of this kind must "fluctuate around a single value, conforming to the general rules of the language", which naturally makes it more difficult to discriminate between different authors.

A useful approach was demonstrated in several works by V. Fuchs, where each author is characterized by such phenomena as the average amount of syllables employed, or the average amount of words in a sentence. This method allows to represent the text of an author as a point on a plane if two parameters are used, or a point in multi-dimensional space, should the amount of parameters grow.

Interesting research is also conducted by a number of Russian philologists, qv in [f6]-[f9], for instance.

One has to point out a common distinctive between the methods of these researchers and their colleagues not mentioned presently is that they are usually directed at studying the individual quantitative parameters of the texts in question, which the scientists would compare to each other in order to find these "salient traits" – ones which would allow to finally distinguish between different authors. However, the key issue here is which of these traits are to be considered significative, and which are to be disregarded; all such distinctions are very prone to being afflicted by subjectivism. This is where the primary hindrances for the application of statistical methods to the problems of this range are concealed.

2. THE DEFINITION OF AN AUTHORIAL INVARIANT

Under the authorial invariant we understand the quantitative characteristic of literary texts (a certain parameter), which would:

a) unambiguously characterize the works of a single author or a small group of "similar authors" by its behaviour.

b) be significantly different for the works of other author groups.

It is desirable that the amount of various "groups" of this kind should be large enough, and that each group would contain a small number of authors with similar literary styles.

However, the multitude of grammatical structures that takes part in the formation of literary texts complicates the search for such invariants to a great extent. The merest experiments involving calculus demonstrate the discovery of numerical characteristics, which would make the distinction between various authors feasible to be a most sinuous issue indeed. The matter is that conscious factors play as important a role in the writing of a book as their subconscious counterparts. For instance, the frequent use of rare and foreign words by an author can naturally be a certain gauge of his style or erudition; however, this is something an author can easily control on the conscious level, since the use of such words in the authorial narrative is something that the author in question will inevitably be aware of. As a result, this quantitative characteristic is of no utility as an authorial invariant, and there are actual calculations that prove it. This characteristic can be controlled by the author and therefore "fluctuates"; it can vary from one work to another.

We can thus see just how recondite a subject the quantitative assessment of a given author's distinctive traits may prove. Let us try and formulate the necessary characteristics that an authorial invariant should possess.

The quantitative characteristic that interests us must satisfy to the following natural conditions:

1) It should be of an overall character, integral to a writer's style and hard to control consciously. In other words, it has to be an "unconscious parameter" rooted deep enough to escape the author's attention altogether. Even if the author did reflect upon it, controlling it for a long time would be an absolute impossibility, and so the author would soon be forced back into his previous stable and *typical* condition.

2) The parameter that we're after must correspond to a certain "regular value", which is to remain roughly the same for all the works of a given author – its deviation from the average should be minimal throughout all of his works. It is this very quality that makes the parameter an invariant.

3) Finally, the invariant should allow for a confident distinction between various groups of writers. In other words, a sufficient amount of authorial groups should exist, whose invariant values would differ from each other significantly.

The third condition is very important. It is possible that a certain parameter will fluctuate minimally throughout the entire textual output volume of every single writer under study, but also assume the *same* value when calculated for *different* authors. In other words, it does not allow us to distinguish between various writers. Only the combination of all conditions as listed above allows us to make the claim of having found the authorial invariant.

3. OUR APPROACH. SAMPLES AND STEPS. THE EVOLUTION OF A PARAMETER ALONG THE NARRATIVE

Let us assume that we have a certain amount of a single author's works at our disposal. We shall arrange them in chronological order for the sake of simplicity (the order in which they were written, in other words), and then refer to the resulting sequence as to the text of a given author. Therefore, the text of an author (in our definition) might consist of several different works – novels, novellas, short stories etc.

After that, we shall study separate fragments of the text in question – samples of the same volume, consisting of the same amount of words (rigidly set a priori). The obvious name we can give this block of text is *sample volume.*

These samples whose volumes are equal are to be taken from every text at equal intervals – that is to say, they should be separated from each other by an equal amount of words. This "distance", or the interval between the neighbouring samples, shall be referred to as a "step", see fig. d3.1.

The volume of samples and the step value can vary depending on our objectives.

Thus, if we move forward along the text of a single author, we can, for instance take a sample of 2000 words every 10 pages of standard text. The longer the text under study, the more samples we can make. The amount of samples shall be small for shorter works, which would complicate the analysis, making the results erratic.

Let us now assume that we chose a linguistic parameter of some sort, for instance, the use frequency of the preposition "in". One can study the evolution of this parameter along the entire text, which might consist of several separate works that we have arranged into a sequence. This shall require us to take consecutive samples and calculate the value of the linguistic parameter that interests us for each of them. As a result, a certain number shall be assigned to each sample. It shall change from sample to sample, generally speaking. We shall proceed with building a graph, with integers like 1, 2, 3, …, to stand for sample numbers on the horizontal axis, and the values of the linguistic characteristic placed along the vertical.

As a result, the evolution of the parameter in question along the entire text that we study shall be represented as some curved line. Therefore, each writer is represented by a line graph and not a point on a plane or in space (the way it is done in such works as [f1] and [f2], for instance). It is rather demonstrative in displaying the behaviour of the parameter under study along the volume of the given author's works. Such graphs turn out to be very convenient for search-

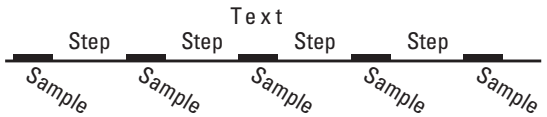

Fig. d3.1. Consecutive samples of equal volume taken from the entire bulk of the literary text under study over equal steps.

ing the authorial invariants. Indeed, the problem can be formulated again in the following manner:

One has to find a linguistic parameter, as well as the optimal sample volume, of such a nature that corresponding graphs would be almost horizontal for every single author (straight lines, or a minutely uneven ones).

In other words, the above implies that the numeric values of the invariant found wouldn't drift too far away from a single average value for an individual author. This phenomenon manifest in the zigzag's tendency to transform into a more or less even horizontal line shall be referred to as the stabilization of the linguistic parameter.

However, the mere observance of such stabilization does not yet suffice for declaring the parameter in question an authorial invariant. It is absolutely necessary for the stabilized graphs (almost horizontal lines) to differ from each other in height substantially – that is to say, they should be situated at different levels. Let us reiterate that these "horizontal lines" corresponding to different authors might be located nearly at the same level, in which case the values of the authorial invariants will be similar. We shall group the authors whose invariant values are close together. In order to make the authorial invariant really effective, it should separate all of the writers into several groups whose invariant values would differ from each other considerably.

Should the values of the authorial invariant for two texts under comparison prove similar, this by itself does not suffice to attribute them to the same author.

We are to understand that the existence of such conspicuous linguistic invariants isn't implied anywhere a priori. Their determination requires an experiment involving very extensive amounts of calculation. We have been conducting this experiment for several years on end; let us now proceed with relating our results.

4. THE EXPERIMENT IN ACTION. THE LIST OF PARAMETERS STUDIED

We have studied the following quantitative characteristics of texts in order to discover the "unconscious parameter", or the authorial invariant that the author can either control to a very small extent or not at all.

1) The length of sentences, or the average amount of words in a sentence calculated for every sample.

2) The length of words, or the average amount of syllables in a word calculated for every sample.

3) General frequency of function word usage (prepositions, conjunctives and particles), or the percentage of function words contained in every sample.

4) Noun usage frequency, or the percentage of nouns for every sample.

5) Verb usage frequency, or the percentage of verbs for every sample.

6) Adjective usage frequency (percentage).

7) Usage frequency of the preposition "in" (percentage, Russian equivalent).

8) Usage frequency of the particle "not" (percentage, Russian equivalent).

9) The amount of function words in a sentence (the average quantity of conjunctions, prepositions and particles contained in every sentence).

Some of the parameters listed above were studied before. However, parameter 3 that we propose (usage frequency of all function words) is a novel one to the best of our knowledge.

The parameters specified differ substantially in their character. Our parameter 3 is very prominent inasmuch as its integral quality is concerned, or the factor of "mass usage"; we count the summary percentage of all the function words, and there's a great abundance of those. The substantial amount of function words used in the Russian language makes this parameter all but impossible to control consciously. A writer can control the length of his sentences to a great extent; however, one finds it hard to imagine an author who could control his function word usage frequency.

Parameters 7 (usage frequency of the proposition "in") and 8 (usage frequency of the particle "not") refer to the distribution of separate function words, and are thus a lot less all-encompassing than the summary parameter 3. We have included them into our research in order to discover whether they can be stabilized at all, and whether they can serve as authorial invariants (and received a negative answer!).

Parameter 9, or the quantity of function words in a single sentence, is of an integral character; nevertheless, it is largely dependent on the length of the sentences and therefore the number of the latter contained in each sample. Calculations demonstrated

this latter value to be rather erratic and prone to fluctuating to a considerable degree without any stabilization whatsoever.

We have purposefully collected numerical characteristics of all possible kinds in our list in order to acquaint ourselves with the comparative behaviour of these parameters, selecting one of them which would in fact stabilize (or the authorial invariant), should the latter be possible to discover at all.

The research was based on the method of taking samples from the general bulk of text described above. Step value, or the interval between neighbouring samples, would equal 60 pages of standard text for large book.

Sample value would however vary. The size of the initial portion, deviated from the 1.000 word quota used by many authors before, equalling 2.000 words and then growing to 4.000, 8.000 and 16.000 words.

The experiment demonstrated that no further extension of volume was necessary, since the authorial invariant was discovered with 16.000 word samples.

In the study of smaller textual volumes the step value would be smaller, and samples were taken more often. However, the experiment demonstrated that step values (unlike sample volume) don't affect the end result all that gravely.

The following principle was adopted as the stabilization criterion. Sample volume would grow until the discovery of the parameter whose deviation from the average values throughout the entire textual volume of all the authors under study would be significantly less than the fluctuation amplitude pertinent to the texts of different authors.

In other words, we would first calculate the deviation of the parameter from the average value, and then average these deviations for all authors in our search of the parameter whose end value would be considerably smaller than the difference between the maximal and minimal values of said parameter for all the authors under study.

5. THE LIST OF AUTHORS AND WORKS STUDIED

We were using the traditional periodic division of the Russian literary language ([f9]). The XIX century was chosen as the main historical period; we have selected 9 writers from this epoch who wrote in Russian and created large texts (see the list below).

However, in order to get a better impression of how the parameters in question evolved depending on the historical epoch, the temporal boundaries of the experiment were broadened with several XVIII and XX century writers added to the list. We got a list of 23 writers as a result (see below). Nearly all of the key works were processed for every writer. It turned out that the results obtained aren't really dependent on the volume of the works, provided the sample volumes are sufficient.

Let us cite the list of the literary works that we processed.

XVIII CENTURY WRITERS.

1) Choulkov, M. D. (1743-1792). – ***The Bonnie Cook,*** novel (written in 1770). Moscow, 1971.

2) Novikov, N. I. (1744-1818). – ***Zhivopisets*** ("The Painter", a magazine of satire. Published in 1772-1773). Moscow, 1971.

3) Fonvizin, D. I. (1745-1792). – ***Diaries of the First Journey*** (written in 1777-1778), ***The Tale of the Deaf and the Dumb***, novel (published in 1783), ***Kallisthenes,*** novel (published in 1786), ***A Friend of the Honest People, or the Archaically-Minded***, an epistolary œuvre (published in 1830), ***An Outspoken Confession of my Deeds and Intentions***, memoirs (published in 1830), Moscow, 1971.

4) Radishchev, A. N. (1749-1802) – ***The Journey from Petersburg to Moscow*** (published in 1790), Moscow, 1971.

5) Karamzin, N. M. (1766-1826) - ***History of the Russian State*** (written in 1816-1826), ***Poor Lisa***, novel (published in 1792), ***Isle Bernholm***, novel (published in 1794), ***Martha from Posad***, a novel (published in 1803), Moscow, 1971.

6) Krylov, I. A. (1769-1844) – ***Qa'ib***, novel (published in 1792), ***Eulogy*** (published in 1792), Moscow, 1971.

XIX CENTURY WRITERS.

7) Gogol, N. V. (1809-1852) – ***Evenings on a Farm near Dikanka, The Fair at Sorochintsy, The Eve of Ivan Kupala, The Night in May or the Drowned Maid, The Missing Letter, Christmas Eve, The Terrible Revenge, Ivan Ivanovich and his Aunt, The Enchanted***

Place, novels (published in 1831-1832), ***Mirgorod, The Countryside Squires, Taras Bulba, Viy, How Ivan Ivanovich Quarrelled with Ivan Nikiforovich, The Petersburg Tales: Nevsky Prospect, The Nose, The Tailor, The Overcoat, The Carriage, The Diary of a Madman*** and ***Rome*** (published in 1833-1842), ***Dead Souls*** (published in 1840), Moscow, 1959 and 1971.

8) Herzen, A. I. (1812-1870) – ***The Past and the Thoughts***, memoirs (published in 1852-1868), Moscow, 1969.

9) Goncharov, I. A. (1812-1891) – ***A Common Tale***, novel (published in 1847), ***Oblomov,*** novel (published in 1859), ***The Precipice*** (published in 1869), Moscow, 1959.

10) Tourgenev, I. S. (1818-1883) – ***Diary of a Hunter*** (written in 1855-1856), ***Roudin,*** novel (written in 1855-1856), ***The Nest of the Nobles,*** novel (written in 1859), ***The Eve,*** novel (written in 1860), ***Fathers and Children***, novel (written in 1862), Moscow, 1961.

11) Melnikov-Pechyorskiy, P. I. (1818-1883) ***The Krasilnikovs*** (travel diary, 1852), ***Grandfather Polikarp***, short story (written in 1857), ***Poyarkov,*** short story (written in 1857), ***The Days of Yore***, short story (written in 1857), ***In the Woods,*** novel (written in 1857-1875), Moscow, 1963.

12) Dostoyevsky, F. M. (1821-1881) – ***Crime and Punishment***, novel (written in 1866), ***The Brothers Karamazov,*** novel (written in 1879-1880), Moscow, 1970-1973).

13) Saltykov-Shchedrin, M. E. (1826-1889), ***Tale of a City*** (written in 1869-1870), ***The Golovlevs*** (written in 1875-1880), Moscow, 1975.

14) Leskov, N. S. (1831-1895) ***Lady Macbeth of the Mtsensk District,*** novel (written in 1864), ***The Warrioress***, novel (written in 1866), ***The Angel Imprinted***, novel (written in 1873), ***The Charmed Wayfarer***, novel (written in 1873), ***Will of Iron,*** short story (written in 1876), ***One Track Mind***, short story (written in 1879), ***Golovan who Feared not Death,*** short story (written in 1880), ***Southpaw,*** short story (written in 1881), ***The Toupee Artist***, short story (written in 1883), ***Sentry on Guard*** (written in 1889), ***A Winter's Day,*** short story (written in 1894), Moscow, 1973.

15) Tolstoy, L. N. (1828-1910), ***Childhood***, novel (written in 1852), ***Adolescence,*** novel (written in 1854), ***Youth,*** novel (written in 1856), ***The Raid***, short story (written in 1852), ***Squire's Morning,*** novel (written in 1856), ***The Cossacks,*** novel (written in 1863), ***War and Peace,*** novel (written in 1863-1869), ***Anna Karenina,*** novel (written in 1873-1877), ***The Resurrection,*** novel (written in 1899), Moscow, 1960-1964.

XX CENTURY WRITERS.

16) Gorky, A. M. (1868-1936) – ***Makar Choudra,*** short story (written in 1892), ***Grandpa Arkhip and Lyonka,*** short story (written in 1894), ***Izerghil the Crone,*** short story (written in 1894-1895), ***Mistake,*** short story (written in 1895), ***One Night,*** short story (written in 1895), ***The Tyke,*** short story (written in 1896), ***The Comrades,*** short story (written in 1897), ***The Orlov Couple,*** short story (written in 1897), ***Formerly People,*** short story (written in 1897), ***Mallow,*** short story (written in 1897), ***For the Sake of Boredom,*** short story (written in 1897), ***Varenka Olesova,*** short story (written in 1898), ***Mates,*** short story (written in 1898), ***The Reader,*** short story (written in 1898), Moscow, 1939. Further also: ***Childhood,*** novel (written in 1912-1913), ***Exposed to the World***, novel (written in 1914-1915), ***My Universities,*** novel (written in 1923), ***The Artamonovs' Case*** (written in 1925). Moscow, 1967.

17) Bounin, I. A. (1870-1953) – ***Antonovskiye Apples***, short story (written in 1900), ***Village,*** novel (written in 1909-1910), ***The Dry Dale,*** novel (written in 1911), ***Zakhar Vorobyov,*** short story (written in 1911-1912), ***Brothers,*** short story (written in 1916), ***Gentleman from San Francisco*** (written in 1915), ***The Lord's Tree,*** short story (written in 1913), ***Natalie,*** short story (written in 1941), ***Good Monday,*** short story (written in 1944), Moscow, 1973.

18) Novikov-Priboy, A. S. (1877-1944) – ***In the Dark,*** short story (written in 1911), ***Slaughterhouse,*** short story (written in 1906), ***Some Joke that was,*** short story (written in 1913), ***The Tainted,*** short story (written in 1912), ***The Call of the Sea,*** novel (written in 1919), ***First Rank Captain,*** novel (written in 1936-1944), ***Tsushima,*** novel (written in 1905-1941), Moscow, 1963.

19) Fedin, K. A. (1892-1977) – ***The Cities and the Years,*** novel (written in 1924), ***Brothers,*** novel (written in 1928), Moscow, 1974.

20) Leonov, L. M. (1899-1994) – ***Russian Woods,*** novel (written in 1953), Moscow, 1974.

21) Shishkov, V. Y. (1873-1945) – ***Taiga,*** novel

(written in 1916), ***Lake Peinus,*** novel (written in 1931), ***Ugryum River*** (written in 1918-1932), Moscow, 1960.

22) Fadeyev, A. A. (1901-1956) – ***Rout,*** novel (written in 1926), ***Young Guard,*** novel (written in 1945).

23) Sholokhov, M. A. (1905-1984) – ***Collected Works in 8 Volumes,*** Moscow, 1962: early short stories – Volume 1, ***The Quiet Don,*** novel – Volumes 2-5, ***Wild Land Pioneered,*** novel – Volumes 6 and 7, short stories – volume 8.

6. THE CALCULATION EXPERIMENT

For each of these writers, we have processed all the works contained in the list in 1974-1977. Namely, the values of the nine linguistic parameters listed above were calculated for all the multiple volumes of text as listed above. As a result, frequency graphs for samples of 2.000, 4.000, 8.000 and 16.000 words in volume were built. All this tremendous body of work was performed manually, since we did not have electronic versions of all these books back then (we aren't certain of whether they actually exist today).

The principle of frequency graph construction was as follows. Along the horizontal axis we put serial numbers of each sample, and along the vertical – the values of linguistic parameters. This resulted in a line graph built for every writer. The parametric fluctuations, or their deviations from the average value, were calculated according to the formula

$$d = (N\ max - N\ min) / N\ avg$$

N. max, N. min and N. avg standing for the maximal, minimal and average value respectively.

7. THE RESULTS OF THE EXPERIMENT

It turned out that all the parameters listed above, except for parameter 3, either fail to stabilize altogether with the growth of the volume sample, or the range of their values for one author is comparable to the maximal value discrepancy for various authors. That is to say, in the latter case all the authors become "collated", and cannot be distinguished between numerically. It is understandable that such parameters could be of no use even for telling one group of authors from the other.

A typical example of the former situation (lack of stabilization with the growth of sample volume) is the evolution of parameter 1 – the amount of words in a sentence, qv in fig. d3.2. It is plainly obvious that even in case of 16.000 word samples the zigzags are chaotic and intermixed to a great extent; their fluctuation amplitude is also excessive.

A typical example of the latter situation (the collation of all writers) is the behaviour of parameter 2 – the amount of syllables in a word, qv in fig. d3.3. Although in case of 16.000 word samples the zigzags begin to assume homogeneity, all the trajectories be-

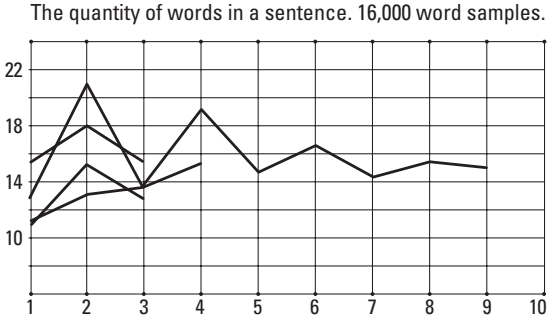

Fig. d3.2. The behaviour of the parameter: word quantity in a sentence for 16,000 word samples. One can instantly see this parameter to be unfit for our purposes due to the fact that it does not stabilize.

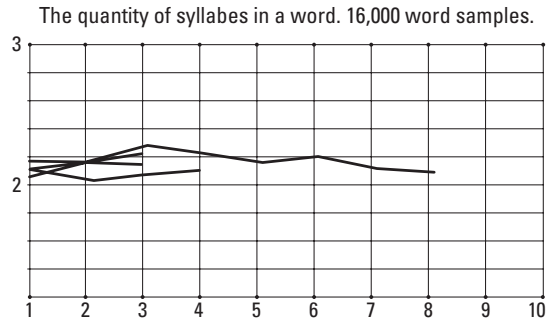

Fig. d3.3. The behaviour of the parameter: syllable quantity in a word. This parameter is obviously unfit for our purposes; it does in fact stabilize, but its values for different authors are virtually the same; thus, it does not allow us to distinguish between various authors.

come virtually coincident, or collated, which makes it impossible to discriminate between authors.

We see a similar picture in case of the parameters 4, 5, 6, 7, 8 and 9. For instance, the graphs of parameter 9 become intermixed and fail to stabilize. The behaviour of parameter 8 is similar to that of parameter 2 – although a large sample volume makes the graphs stabilize, they become too similar to each other and gravitate towards a single value, which is apparently dictated by the laws of the language itself, and not the individual characteristics of the writer.

This makes the utility of parameters 1, 2, 4, 5, 6, 7, 8 and 9 for the purpose of distinguishing between various authors very dubious indeed.

8.
FUNCTION WORD USAGE FREQUENCY TURNS OUT TO BE THE AUTHORIAL INVARIANT

A most notable exception is parameter 3 – usage frequency of all function words in general – prepositions, conjunctives and particles. The evolution of this parameter in accordance with the growth of sample volume can be seen in figs. d3.4, d3.5, d3.6 and d3.7.

The list of Russian function words as given by the authors comprises 55 words. It may be incomplete, but allows for the differentiation between the authors.

Important experimental fact.

1) Sample volume of 16,000 words made the function word percentage for each author in our list (with the exception of a single writer whose case shall be analyzed below) roughly the same for each of his

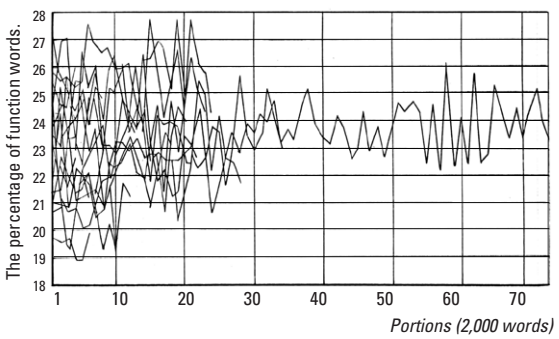

Fig. d3.4. The behaviour of the parameter: function word usage for 2,000 word samples. The line graphs are chaotic.

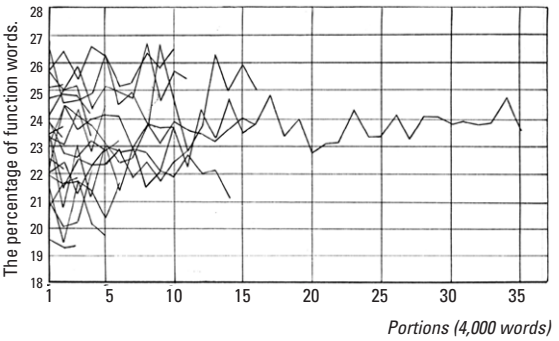

Fig. d3.5. The behaviour of the parameter: function word usage for 4,000 word samples. The line graphs remain chaotic, yet demonstrate a tendency to stabilize.

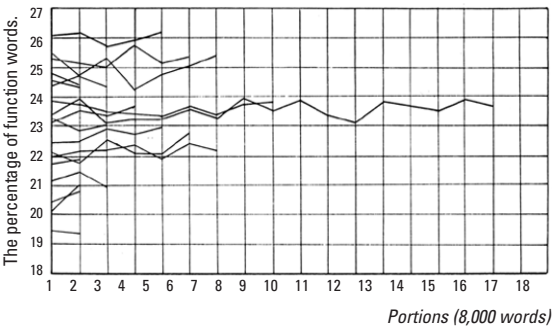

Fig. d3.6. The behaviour of the parameter: function word usage for 8,000 word samples. The line graphs still "intertwine", but demonstrate a growing tendency to stabilize.

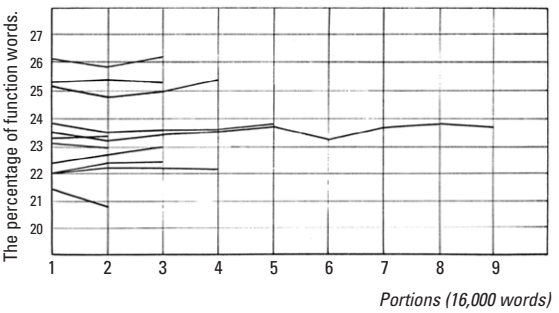

Fig. d3.7. The behaviour of the parameter: function word usage for 16,000 word samples. The line graphs became even, which implies parameter stabilization. The values of the parameter are substantially different for various authors, which makes the parameter fit for our purposes. It is thus the authorial invariant and allows us to distinguish between certain authors.

works, that is, the frequency graph is almost horizontal. This stabilization takes place in case of 22 writers out of 23 studied, see fig. d3.7.

2) The difference between the maximal and minimal value of parameter 3 (with the minimum and maximum taken for each writer under study) is a lot greater than its fluctuation amplitude as given for the works of other authors. The parameter's fluctuation amplitude for various authors is great enough – 19% to 27.5%, qv in fig. d3.7. Hence we see that parameter 3 is useful enough for differentiating between many authors.

Therefore we shall be referring to parameter 3 as to the authorial invariant. It may serve for the attribution of unknown works as well as the discovery of plagiarism, albeit with a certain amount of care, since we have discovered writers whose authorial invariants are very close to each other, for instance D. I. Fonvizin and L. N. Tolstoy, qv below. Also, one needs large volumes of text in order to arrive at any confident conclusions.

The main inference here is the rather seminal assertion concerning the existence of an authorial invariant applicable to Russian literary texts. It would be of great interest to continue with the experimentation in order to discover other authorial invariants.

Let us point out that such conclusions can only be made after large-scale computational experimentation. Only upon having received empiric proof that this or the other parameter really stabilizes within the framework of œuvres written by a single author one can consider the parameter in question an invariant. The list of authors processed also needs to be large enough – several dozen of them at least. Constructing any theories of any kind is a rather pointless activity if they are based on the comparison of texts belonging to just one or two authors, as we see it.

It is interesting that the authorial invariant that we have discovered is virtually independent from the epoch: the list that can be seen above represents the authors of three centuries – from the XVIII to the XX.

9. QUANTITATIVE EXAMPLES

Since we have discovered the 16,000 word sample graphs to be of the greatest interest to us, we shall be regarding just this case in our study.

Let us cite a value table of the following parameters for the works of I. S. Tourgenev and L. N. Tolstoy:

3 – the amount of all function words used (percentage),
1 – the amount of words in a sentence,
2 – the amount of syllables in a word,
9 – the amount of function words in a sentence,
7 – usage frequency of the preposition "in" (percentage),
8 – usage frequency of the particle "not" (percentage).

PARAMETERS	3	1	2	9	7	8
	22.01	11.26	2.17	2.44	2.36	2.19
TOURGENEV	22.36	15.58	2.16	3.49	2.05	1.87
	22.38	13.35	2.21	3.04	-	-
Average value	22.24	13.40	2.17	2.98	2.20	2.04
Deviation	0.016	0.322	0.023	0.35	0.14	0.16

PARAMETERS	3	1	2	9	7	8
	23.67	13.13	2.11	3.09	2.10	2.05
	23.34	20.75	2.15	4.79	2.56	1.72
	23.45	14.27	2.28	3.35	2.38	1.67
	23.58	18.93	2.16	4.62	2.46	1.87
TOLSTOY	23.78	14.86	2.15	3.64	2.74	1.88
	23.35	16.33	2.19	3.80	2.71	1.93
	23.77	14.23	2.11	3.47	2.15	2.17
	23.82	15.24	2.11	5.75	2.19	2.07
	23,77	14.97	2.20	3.42	2.49	1.75
Average value	23.62	15.95	2.16	3.81	2.36	1.92
Deviation	0.020	0.477	0.08	0.45	0.27	0.26

One can plainly see that the parameters with the smallest deviation values are the third and the second, namely, 0.016 and 0.023 for Tourgenev and 0.020 and 0.08 for Tolstoy. However, parameter 2 cannot serve as authorial invariant since its values for most authors in our list are all rather close to each other – 2.17 for Tourgenev and 2.16 for Tolstoy, for instance. Therefore, from the point of view of parameter 2, all the writers "merge into one", which doesn't allow us to distinguish between them.

Parameter 3 – function word usage frequency – isn't merely an invariant; it allows to discriminate between a sufficient amount of authors. For instance, it equals 22.24 for Tourgenev and 23.62 for Tolstoy. The difference equals 1.38, which is greater than the

value of the parameter's fluctuations in the works of Tourgenev and Tolstoy.

Parameter 3 may assume values from 19.4 to 27.5 per cent, which means that the range of its meanings is broad enough as compared to the fluctuations of the parameter inside the texts of separate authors.

Let us now cite the table of parameters 3, 7 and 8 as measured for Gogol, Herzen, Dostoyevsky, Leonov and Fadeyev.

PARAMETERS	3	7	8
	23.82	2.25	2.10
	23.54	2.29	1.86
GOGOL	23.61	2.61	1.82
	23.62	2.75	1.90
	23.85	2.10	2.50
Average value	23.65	2.45	1.95
Deviation	0.013	0.027	0.35

PARAMETERS	3	7	8
	22.42	2.87	2.03
HERZEN	22.87	3.10	2.04
	22.98	2.64	1.92
Average value	22.71	2.91	2.01
Deviation	0.024	0.16	0.06

PARAMETERS	3	7	8
	25.26	2.23	1.70
DOSTOYEVSKY	25.43	2.48	2.21
	25.29	2.13	2.14
Average value	25.32	2.38	2.02
Deviation	0.007	0.15	0.25

PARAMETERS	3	7	8
LEONOV	23.11	2.97	1.81
	23.04	2.58	2.00
Average value	23.06	2.83	1.90
Deviation	0.003	0.14	0.10

PARAMETERS	3	7	8
FADEYEV	23.40	2.54	1.78
	23.43	2.72	1.99
Average value	23.40	2.62	1.89
Deviation	0.002	0.07	0.11

Let us cite the table of parameters 3, 1, 2 and 9 for Goncharov and Leskov.

PARAMETERS	3	1	2	9
	25.13	11.67	2.09	2.92
	24.88	13.16	2.03	3.31
GONCHAROV	24.98	13.72	2.06	3.68
	25.47	15.05	2.10	3.58
Average value	25.06	13.41	2.06	3.37
Deviation	0.019	0.25	0.03	0.26

PARAMETERS	3	1	2	9
	26.08	15.65	2.05	3.99
LESKOV	25.83	18.11	2.16	4.69
	22.98	2.64	1.92	
Average value	26.18	15.40	2.11	4.02
Deviation	0.010	0.16	0.05	0.163

The values of parameter 3 are characterized by high stability for Gorky: 22.02, 22.21, 22.20, 22.17 etc. The average value is 22.15, the deviation equalling 0.009.

A propos, the values of all the parameters listed above were calculated to three places of decimals. The values in the table are rounded off to two decimals. Three decimals were only used for the deviations from the average value of parameter 3.

Since parameter 3, or the percentage of all function words used, demonstrates amazing stability and distinctive capacity, it would be interesting to trace its fluctuations using samples of different volume specifically.

Let us site the table that demonstrates the dependency of the deviation value from the average with differing sample volume.

WRITERS	Function word percentage	The deviation of the parameter from the average value with samples of the following volume (in words):			
		2,000	4,000	8,000	16,000
Radishchev	22.30	0.054	0.018	-	-
Karamzin	19.44	0.051	0.014	0.003	-
Krylov	23.67	0.040	0.013	-	-
Gogol	23.65	0.169	0.066	0.019	0.013
Herzen	22.71	0.165	0.109	0.025	0.024
Goncharov	25.06	0.229	0.116	0.046	0.019
Tourgenev	22.24	0.126	0.069	0.040	0.016
Melnikov-Pecherskiy	24.49	0.240	0.062	0.005	-
Dostoyevsky	25.32	0.203	0.098	0.030	0.007
Saltykov-Schedrin	24.56	0.173	0.042	0.016	-

WRITERS	Function word percentage	The deviation of the parameter from the average value with samples of the following volume (in words):			
		2,000	4,000	8,000	16,000
Leskov	26.01	0.132	0.057	0.017	0.010
Tolstoy	23.62	0.199	0.103	0.036	0.020
Gorky	22.15	0.201	0.109	0.020	0.009
Bounin	24.64	0.143	0.027	0.013	-
Novikov-Priboy	21.10	0.129	0.090	0.049	-
Fedin	21.20	0.151	0.064	0.028	0.019
Leonov	23.08	0.147	0.049	0.014	0.003
Shishkov	20.60	0.152	0.115	0.019	-
Fadeyev	23.40	0.184	0.111	0.018	0.002

As one sees from the table, the stabilization of parameter 3 sometimes takes place with samples smaller than 16,000 words.

This is particularly true for the XVIII century authors – for Karamzin, the stabilization of the authorial invariant takes place at volumes of 8,000 words, and the same is true for Fonvizin. This may indicate a greater stylistic rigidity of the XVIII century authors as compared to their colleagues in the XIX and XX century.

This early stabilization that we discovered demonstrates that in certain cases the authorial invariant (percentage of function words) can also be used for the analysis of texts whose volume isn't all that large. However, extensive research requires 16,000-word samples, since it is only in the latter case that the stabilization of parameter 3 takes place simultaneously for all the authors under study.

After the discovery of the authorial invariant for the 22 writers listed above, the range of works processed during the experiment was widened, with similar calculations performed for the works of five other authors: A. N. Ostrovskiy, A. K. Tolstoy, V. A. Zhukovskiy, A. S. Pushkin and A. P. Chekhov. The works selected were all in prose, and all of a large volume. The extended experiment proved the high stability of parameter 3 with the use of 16,000-word samples, as well as its capacity for discerning between various groups of authors.

Thus, the complete list of writers for which parameter 3 serves as a stable and distinctive authorial invariant was extended to include 27 authors instead of 22.

10. THE POSSIBLE USES OF THE AUTHORIAL INVARIANT. ITS POTENTIAL FOR THE DISCOVERY OF PLAGIARISMS

One of the possible uses of the authorial invariant that we discovered is the identification of plagiarisms, the ascertainment of possible authorship etc. One could suggest using the following method as natural: if the difference between the values of parameter 3 (function word percentage) is greater than one, there are reasons to attribute the texts under comparison to different authors. The greater the difference of the invariant value, the more we are entitled to suspect this.

On the other hand (and the same is true for the problem of paternity tests), similar invariant values aren't reason enough to attribute the works in question to the same author. As we already pointed out, there are writers whose invariant values are close to each other – such as Fadeyev and Leonov, whose invariant values equal 23.08 and 23.40, respectively.

Apart from that, one has to act with the utmost caution if one applies this method of authorial identification to texts of small volume. The complications that arise here can be illustrated by the example of large and small works of A. P. Chekhov. Parameter 3 (function word percentage) was calculated for all of his oeuvres that came out as the 1960-1964 collected works publication. We have discovered that parameter 3 behaves in the following manner:

	SHORT STORIES					LARGE TEXTS		
VOLUME NUMBER	I	II	III	IV	V	VI	VII	VIII
Function word percentage	22.6	22.5	23.4	22.7	23.4	25.4	25.5	25.4

The difference between the parameter 3 values for Chekhov's early short stories collected in Volumes I-V, and the larger works of his late period (Volumes VI-VIII) is rather ostensible, qv in fig. d3.8. It isn't that his earlier works employ less function words – the key factor is that they're scattered about to a greater extent than in the ensuing large works. Chekhov's voluminous (late) works are characterized by a highly stable authorial invariant, as well as all the remaining 26 authors of *large* texts from our list. Chekhov is no ex-

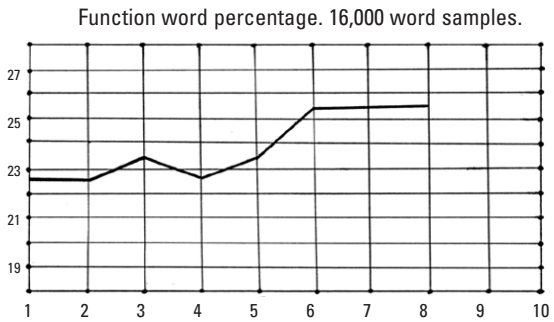

Fig. d3.8. The percentage of function words may demonstrate instability when applied to small-volume texts. It is demonstrated by the example of A. P. Chekhov.

ception – parameter 3 "serves" all of his *large* works perfectly well.

Let us conclude with relating another interesting circumstance. It turns out that the percentage of function words corresponds to a more stable value in case of prose (with sample volumes equalling 8,000 and 16,000 words), and a less stable one in case of poetry. This issue deserves to be considered separately, and we shall not linger on it here.

The discovery of the authorial invariant in literary Russian language makes the hypothesis of the existence of similar authorial invariants in other languages. They may naturally correspond to other factors than the percentage of function words used. Authorial invariants in Greek and Latin would be of the utmost interest, if we are to consider the use of similar methods for authorship identification in case of ancient texts.

11.
THE STATISTICAL ANALYSIS OF THE WORKS OF M. A. SHOLOKHOV.
The authorial invariant of "The Quiet Don" is drastically different from the authorial invariant of all the other works written by M. A. Sholokhov

Attentive readers must have already noticed that one of the writers wasn't considered in out list of 28. This writer is Mikhail Aleksandrovich Sholokhov, and we're about to conduct a research of his works. All the conclusions we arrive at are based on the analysis of his collected works that came out as an 8-volume edition in Moscow, 1962.

We must immediately point out that we by no means claim to have made any finite conclusion, publishing the results of our research in hope that they might prove useful for the researchers of Sholokhov's works.

It is widely known that M. A. Sholokhov attained a rather prominent position in Russian and world literature, and his Nobel Prize of 1965 testifies to his international acclaim as well.

Nevertheless, it is for a couple of decades now that a number of specialists in Russia as well as abroad have been expressing doubts about whether M. A. Sholokhov is really the author of the ***Quiet Don,*** or whether the work in question may have been written by the Cossack writer Fyodor Dmitrievich Kryukov who was a soldier in the White Army of Don and died of typhoid fever in 1920.

We already stated that we do not intend to support either party in this discussion, and merely want to relate the statistical results of our research.

Let us briefly relate the subject of the argument.

It is common knowledge that during the First World War as well as the Russian Civil War F. Kryukov had written a lot about the Don Cossacks. After his death (according to the author known to us under the alias D., for instance, whose research entitled ***The Stirrup of the Quiet Don*** ([f11]) came out in 1974), Kryukov's manuscript of the ***Quiet Don*** was found by Sholokhov, who is supposed to have made some alterations and replaced Kryukov's Cossack nationalism by more pro-Soviet sentiments, subsequently publishing the novel under his own name ([f21]).

"D." proceeds to claim that both the language and the style of Kryukov's texts demonstrate an astonishing similarity to those of the ***Quiet Don.*** He is of the opinion that about 95% of the I and the II books of the *Quiet Don* and 68-70% of books III and IV were written by Kryukov, and Sholokhov could only have been a co-author. One cannot ignore the fact that Kryukov was specifically a Cossack writer, and thus was well familiar with the life and history of the Cossacks.

In his preface to the book by "D.", A. Solzhenitsyn wrote that "from the day it came out in 1928, the ***Quiet Dawn*** spawned a great many mysteries which cannot be explained until the present day. The readers were confronted by a case that has no precedent in world

literature. A 23-year old debutant creates a work utilizing material that far exceeds his experience and level of education (4 forms). The young provision commissary (subsequently a navvy in Moscow and then as a clerk in a housing office at Krasnaya Presnya) published an œuvre that could only be prepared as a result of numerous conversations with representatives of many strata of the pre-revolutionary Don society and was all the more fascinating that it demonstrated inside knowledge of the life and the psychology of the strata mentioned above".

The postulations of "D" were sharply criticized by Yermolayev ([f15] and [f16]). On the other hand, the conclusions of "D." were supported by A. Solzhenitsyn and R. Medvedev.

By the way, according to the authors of [f18], in May, 1990 N. A. Struve, the publisher of ***The Stirrup of the Quiet Don,*** discovered the identity of "D." – it turned out to be I. N. Medvedeva-Tomashevskaya, a prominent literary critic ([f18], page 7).

In 1991 A. G. Makarov and S. E. Makarova published their work entitled ***The Melancholy Thistle. Towards the Sources of the Quiet Don*** ([f18]). In their analysis of the novel's language as well as its historical and chronological contents, A. G. Makarov and S. E. Makarova come to the conclusion that Sholokhov processed a work of a different author and published it under his own name, after their comparison of the novel's text with the surviving written materials of other authors. Also see their book entitled ***Around the Quiet Don*** published in 2000 ([f23]).

It has to be pointed out that Sholokhov was accused of plagiarism as early as 1928, when the first two books of the ***Quiet Don*** were published.

The issue of Kryukov's authorship was also raised by the relatives of Kryukov; however, their claims weren't satisfied due to the lack of direct evidence.

However, rumours of any sort can hardly be regarded as evidence unless they are backed up by a solid body of research. All the claims and statements uttered in this respect made two Swedish and two Norwegian researchers analyse Sholokhov's texts with the aid of a computer ([f10], [f13] and [f14]). See more details in [f10], published in 1984 (Russian translation published in 1989).

The analysis of various frequency characteristics (statement length, word length etc) led them to the conclusion that all parts of the Quiet Don can be attributed to Sholokhov.

However, above we demonstrate that such parameters as well as the ones related to them either fail to stabilize altogether, or aren't sensitive enough for the discovery of authorship. This is easy to see from a comparison of sentence and word length performed with the entire bulk of all Sholokhov's works published as a series of 8 volumes in 1962.

Sholokhov's works	Words per sentence, average	Syllabes per word, average
Vol. I - **Short stories**	9.17	2.18
Vol. II - **The Quiet Don**	8.73	2.27
Vol. III - **The Quiet Don**	9.85	2.32
Vol. IV - **The Quiet Don**	9.30	2.31
Vol. V - **The Quiet Don**	9.66	2.21
Vol. VI - **Wild Land Conquered**	8.77	2.19
Vol. VII - **Wild Land Conquered**	10.70	2.15
Vol. VIII - **Short stories and novellas**	10.30	2.28

We can see that if the average amount of words per sentence fluctuates here, the average amount of syllables per word remains more or less constant. Therefore, if we were to judge by the behaviour of the syllable-per-word value, we could come to the conclusion favouring Sholokhov if we wanted to. However, such conclusion would by all means be premature, since we know that none of these parameters happen to be the authorial invariant.

It has to be said that the researchers in question (see [f10]) had neither discovered our invariant, nor come up with methods whose effectiveness would stem from a study of many other authors.

It is natural that we should become interested in the subject – the primary motivation wasn't so much curiosity as the wish to try out the method that we discovered, which was conceived with similar objectives in mind.

Having acquainted ourselves with the works written on this subject that we had at our disposal, we learnt that the researchers would often compare various frequency characteristics of Sholokhov's works to those of other writers – Kryukov, for instance, without going beyond the works of two authors (Sholokhov and Kryukov, for instance). This com-

parison would then serve as basis for a conclusion of some sort, in Kryukov's favour or in favour of other claimants.

However, as far as we know, previous experts did not bother to find out whether the frequency characteristics they used were in fact authorial invariants, which is a sine qua non in the study of such problems as the authorship issue. One would need to discover an authorial invariant first, processing several dozen authors of all sorts the way we did. The first stage inevitably involves a large-scale statistical experiment involving a great amount of material. It is only afterwards, after the discovery of a stabilizing and differentiating invariant, should this prove feasible at all, that one can attempt to apply it to the problem of the ***Quiet Don,*** for instance.

In other words, one first needs to "forge the tools of the research" (in an extensive calculation experiment involving many authors representing a great number of literary fields), and only then attempt to use them practically.

This is the way we chose. First we had to discover the stabilizing and differentiating invariant; it proved to be the percentage of function words used by a given author. Then we applied it to the study of Sholokhov's texts.

We found the result perfectly flabbergasting.

Function words in his works are distributed so unevenly that one has to present Sholokhov as two authors – Sholokhov I and the alleged Sholokhov II.

The exact result is given in fig. d3.5 and the table below.

Sholokhov's works	Function words (%)
Short stories	22.46
The Quiet Don, books I and II, parts 1-5 and the beginning of part 6 in book III	19.55
The Quiet Don, the second part of book III and the entire book IV (i.e. part 6 continued and parts 7-8)	22.69
Wild Land Conquered, books I and II	23.07
Late short stories and novellas	24.37
Essays, articles, causeries and speeches	23.35

See a more detailed table at the end of the present article.

This enables the formulation of the following three important conclusions:

1) The works we can attribute to Sholokhov I are as follows:

a) his early short stories;

b) the last section of part 6 and the final parts 7 and 8 of the ***Quiet Don,*** as well as:

c) all the works that followed – ***Wild Land Conquered*** and late short stories and novellas.

2) The alleged Sholokhov II can be credited as the author of parts 1, 2, 3, 4 and 5 of the ***Quiet Don,*** as well as the beginning of part 6.

3) Part 6 occupies an intermediate position between the works of Sholokhov I and the alleged Sholokhov II. Its first section (about 100 pages) can be confidently attributed to the alleged Sholokhov II, whereas the ensuing pages of the 6th part were definitely written by Sholokhov I.

The table and fig. d3.9 make it perfectly obvious that the style of Sholokhov's early short stories (1924-1927) is virtually indistinguishable from the style of the final parts 7 and 8 of ***The Quiet Don*** as well as everything written after that inasmuch as the percentage of function words is concerned.

If this value equals 19.55% in average for parts 1-5 and the beginning of part 6 of ***The Quiet Don,*** it becomes 23.03% for all the other works of Sholokhov, either written later or earlier.

The difference of roughly 3.48% between the authorial invariant values for Sholokhov I and the alleged Sholokhov II (see fig. d3.9) is so great that one cannot afford to disregard it. These texts are highly unlikely to be attributable to a single author.

Our conclusion.

Statistical results obtained in the course of authorial invariant analysis confirm the hypothesis that parts 1, 2, 3, 4 and 5, as well as a large section of part 6, of the novel ***The Quiet Don*** were not written by M. A. Sholokhov.

However, we may encounter counter-argumentation – for instance, claims that Sholokhov had changed his writing style dramatically when he was creating parts 1-5 of ***The Quiet Don.*** His authorial invariant had possessed a given value initially which would then change along with his "style change" which coincided with the creation of the first five

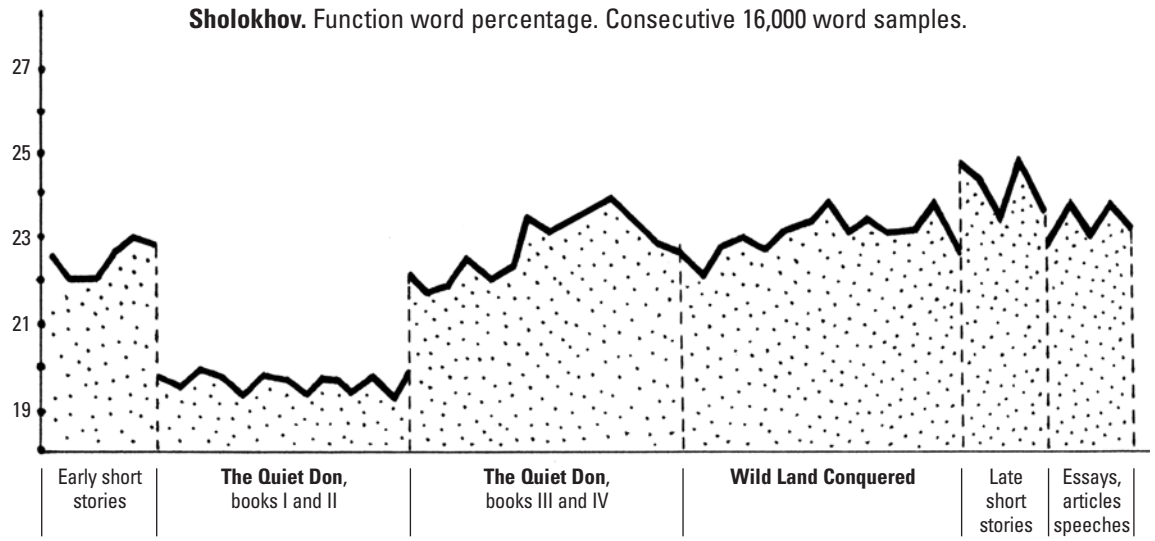

Fig. d3.9. The behaviour of the function word percentage parameter demonstrates rather obviously the high likelihood of M. A. Sholokhov not being the author of ***The Quiet Don***.

parts of ***The Quiet Don.*** Then he allegedly returned to his old narrative manner.

This is possible.

However, in this case one would have to recognize Sholokhov as a unique occurrence in the entire Russian literature, amazing enough for a special study of this phenomenon - after all, he would then become the only Russian writer out of the ones we studied who managed to change the value of his authorial invariant drastically.

Indeed, the randomly chosen 27 other authors who had written voluminous works (hailing from various centuries and literary schools) demonstrate a lifelong adherence to their literary style – inasmuch as parameter 3 is concerned, at least, which is confirmed by our calculation experiment.

As for Sholokhov – he managed to suddenly change his style for a year or two; furthermore, he managed to keep this radically new style for the whole time that the first five gigantic parts of ***The Quiet Don*** were created. And we did already mention that the percentage of function words used in the narrative is an integral factor, and it is also of an omnipresent nature – most probably beyond conscious control of the author (which proved true for the 27 other writers).

The example with the change of Chekhovian style given above doesn't count, since we were comparing his short stories to his large works, whereas in the case of Sholokhov we are concerned with his large works exclusively.

If we are to divide the general volume of function words in Sholokhov's works into prepositions, conjunctives and particles, Sholokhov I demonstrates about the same amount of prepositions as the alleged Sholokhov II; however, there are a lot more conjunctives and particles in the works of the former as compared to the latter. See for yourselves.

	Sholokhov I	The alleged Sholokhov II
Prepositions	10.62	11.61
Conjunctives	7.36	4.80
Particles	4.59	3.07

Once again, this testifies that the texts of Sholokhov I and the alleged Sholokhov II differ drastically.

One cannot fail to mention a good concurrence between our result and the independent conclusion of the critic "D" based upon completely different considerations, namely, that the books I and II, as well as the beginning of book III, weren't written by Shol-

okhov. However, "D." had also been of the opinion that about 70% of books III and IV weren't written by Sholokhov, either; our results demonstrate that a large part of book III is characterized by Sholokhov's authorial invariant value.

12. OBSERVATIONS OF A SECONDARY NATURE. Chronology and volume of Sholokhov's publications

The quantitative difference between various parts of ***The Quiet Don*** implies the need to divert our attention towards the chronology and the volume function of Sholokhov's writing. Study the table offered below carefully, as well as fig. d3.10 which serve to illustrate the annual volume distribution of Sholokhov's publications (according to the 8-volume collection of 1962).

Years	Publication volume (printed pages per year)	Sholokhov's age
1924-1927	4.6	19-22
1928	47.6 (!)	23
1929-1931	5.6	24-26
1932	24.3	27
1933-1936	No publications	28-31
1937-1938	8.1	32-33
1939	No publications	34
1940	8.1	35
1941	No publications	36
1942	1.3	37
1943-1944	2.7	38-39
1945-1948	No publications	40-43
1949	2.7	44
1950-1953	No publications	45-48
1954	5.6	49
1955	2.9	50
1956	3.9	51
1957	3.9	52
1958-1960	2.9	53-55

Sholokhov is supposed to have been born in 1905. However, in 1994 there was a series of programmes on St. Petersburg television where this date was declared dubious, with the theory proposed that Sholokhov was really born later than it is presumed officially. Since we did not study this issue, we shall adhere to the official point of view.

It is also presumed (see Annexes to Volume VIII of Sholokhov's works, Moscow, 1962) that Sholokhov began the creation of ***The Quiet Don*** in the end of 1925, being a mere 20 years of age. In 1928, when Sholokhov was only 23, parts 1-5 of ***The Quiet Don*** had already been published; their volume is gigantic – 47.6 printed sheets. This text was printed in record terms: the first part was printed in the first 1928 issue of the ***Oktyabr*** magazine, and the last – in the tenth issue the same year.

Therefore, the manuscript could only have been received by the editing board in 1927, or possibly even earlier. Should this prove true, and we hardly have a reason to doubt it, the completion of such a voluminous (47.6 printed sheets) and mature work as the first two books of ***The Quiet Don*** required a single year at best – 1926. Sholokhov himself wrote that he "started to sketch out ***The Quiet Don*** in autumn 1925, but stopped after having written about 3-4 printed sheets" (M. Sholokhov, ***Autobiography***, quoted according to ***The Creation of the "Quiet Don"*** by V. V. Goura, Moscow, 1980, pages 95-96. See also [f18], page 134.

Therefore, according to the critics of Sholokhov's writing, when he was only 20 or 21 years of age, with neither general (4 years at the gymnasium) nor special education, nor experience, nor fame, nor access to the archives of war (and the novel contains a great amount of factual information pertinent to the time of the war), he managed to create a fundamental and highly literary work in record terms.

It is hard to consider such argumentation demonstrative; still, one does get a feeling that something is out of place here.

L. Kolodny, Sholokhov's apologist, wrote that "Mikhail Sholokhov began independent life in 1918, at the age of 13. He took part in the civil war as member of a 216-bayonet party. Sholokhov had been tried for "excess of jurisdiction", but saved from the sentence by the fact of his being underage… as for the four years of his gymnasium studies, we should also recollect Ivan Bunin, whose period of education had been even shorter – a mere three years; nevertheless, he became a Nobel laureate just like Sholokhov" ([f17]).

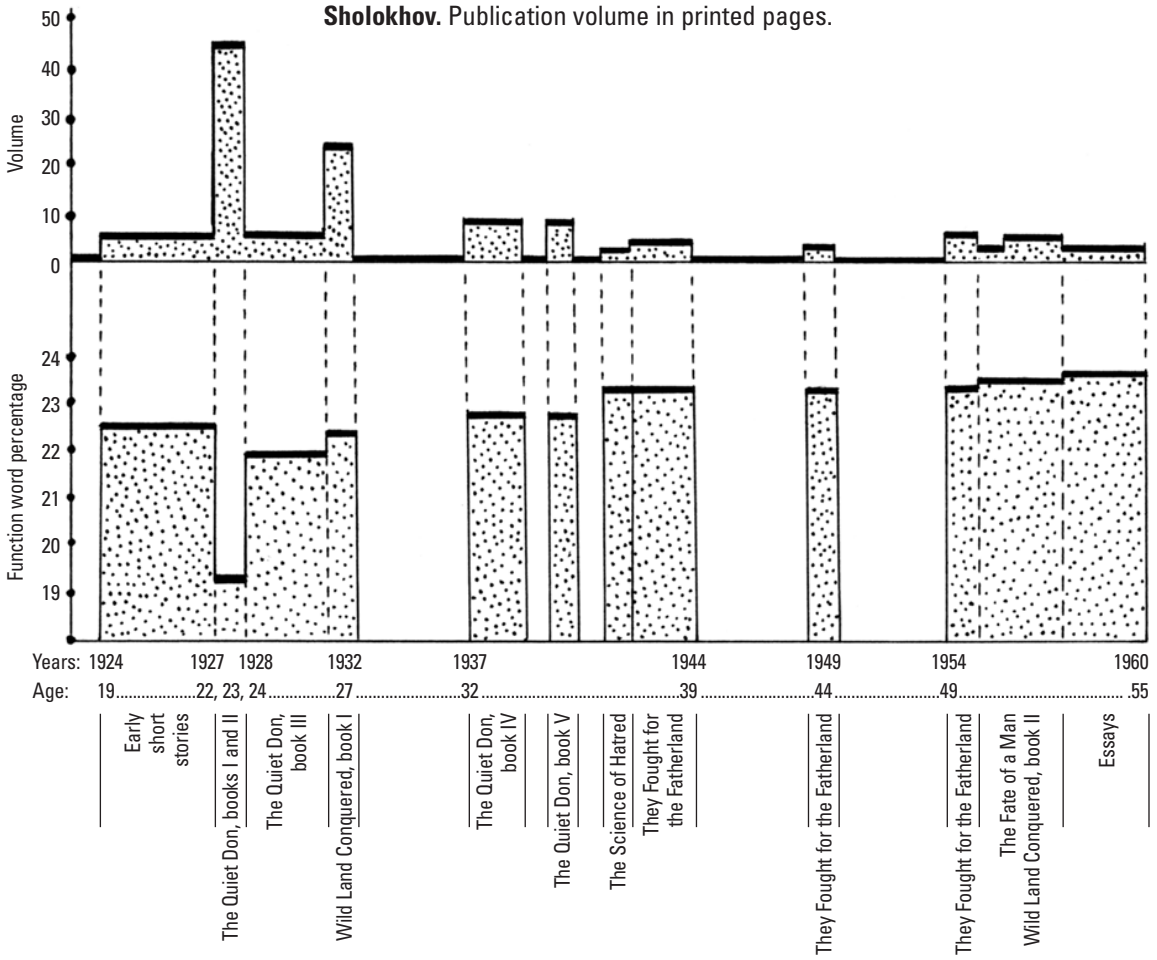

Fig. d3.10. A comparison of function word evolution and yearly publication volumes of M. A. Sholokhov. It is amazing that the greatest yearly volume (***The Quiet Don***, books 1 and 2) is characterized by the smallest percentage of function words used.

As one sees from the volume table and the graph in fig. d3.10, the average annual production rate of Sholokhov over the 40 years of his literary career fluctuated around 3.5 printed sheets; if we are to exclude the text under suspicion, it shall go even further to 2 printed sheets per year.

Such annual volume is exceptionally small in comparison to other professional writers. Chekhov produced around 14 printed sheets a year, Leo Tolstoy – around 13, and Emile Zola would manage around 21. All of this makes the sporadic one-year activity burst that allowed Sholokhov produce a mind-bogglingly great amount of high-quality prose (47.6 printed sheets) over the course of a single year (1926), at the very young age of 20 or 21. His subsequent productivity was a lot lower, and the same is true for the period that preceded 1926.

However, all of these considerations are of a secondary nature and are by no means presented as veracious independent argumentation. The fact that the original manuscripts of ***The Quiet Don***'s first two books still aren't located anywhere (to the best of our knowledge) also cannot serve as independent argumentation. The manuscripts of books III and IV, which can be safely declared beyond suspicion, are kept in the archives of "The Pushkin House" in

St. Petersburg, whereas the manuscripts of the first two books that interest us are presumed missing – allegedly perished in a fire. On the other hand, in May 1995 a news programme on the "Ostankino" channel reported the original manuscripts of ***The Quiet Don***'s first two books found at last. It would be interesting to further elucidate this issue; still, it bears no relation to the results of our statistical research.

13. THE ANALYSIS OF SEVERAL TEXTS BY F. D. KRYUKOV

Since some of the researchers are convinced that the Cossack writer Fyodor Kryukov was the co-author of ***The Quiet Don***, it would be apropos to study this issue as well. Unfortunately, we did not have the later, more fundamental works of Kryukov written during World War I and the Civil War, at our disposal. In general, as it is pointed out in [f18], the biography of F. D. Kryukov had remained all but unknown to the Soviet reader until 1990. A. G. Makarov and S. E. Makarova are of the opinion that "Soviet literary critics played an important role in keeping Kryukov obscure – specialists in the field of Sholokhov's writing in particular" ([f18], page 14).

We could only analyse several of Kryukov's early short stories – ***The Thirst***, ***The Mother***, ***Half an Hour***, and ***A Step and No Movement***. All of them were written by Kryukov before World War I, in 1905-1907, and pertain to the dawn of his literary career. Let us therefore state in advance that one shouldn't have any aspirations concerning this meagre material.

The results obtained were arranged into a table.

Kryukov's works	General amount of words	Amount of function words	Function word percentage
The Thurst	5,528	1,161	21.00
Half an Hour	4,391	924	21.04
The Mother	14,965	3,159	21.17
A Step and No Movement	18,699	3,954	21.14
Total:	43,583	9,198	21.11

One sees that the sample volumes available to us are minute; therefore, the result might prove unstable. Nevertheless, the percentage of function words in Kryukov's writing is rather stable and fluctuates minimally.

The small volume of text under study, as well as the rather poor vocabulary of Kryukov's early works, and also the fact that some of these short stories have got nothing to do with the Cossacks, do not permit to make a conclusion about Kryukov's either being a co-author of ***The Quiet Don*** or having no relation to the book whatsoever.

However, the cited results permit the assumption that Kryukov's co-authorship is more than unconfirmed rumour. As one sees from the function word percentage rates, the difference between Kryukov's works and the first two books of ***The Quiet Don*** equals a mere 1.56% = 21.11 – 19.55. The difference between Sholokhov I and the same books of ***The Quiet Don*** (or the alleged Sholokhov II) is a lot greater and equals 3.48% = 23.03 - 19.55. This implies that the style of Kryukov isn't all that different from that of ***The Quiet Don*** quantitatively.

M. A. Sholokhov's index is a lot further from the first two books of the novel than that of F. D. Kryukov.

However, until later texts written by Kryukov about the history of the Don Cossacks are studied, one can make no definite conclusions about Kryukov being in any relation to the creation of the first two books of ***The Quiet Don***. Nevertheless, we have no reasons to refute his participation, either.

Let us conclude with providing the portraits of the two authors – F. D. Kryukov's is in fig. d3.11, and Sholokhov's – in fig. d3.12.

Fig. d3.11. A portrait of F. D. Kryukov. Taken from [501].

Fig. d3.12. A portrait of M. A. Sholokhov. Taken from [501].

14. A DETAILED TABLE OF FUNCTION WORD DISTRIBUTION IN M. A. SHOLOKHOV'S TEXTS

The first column contains the sample number; the second refers to the sample volume in words, the third – to the amount of function words in the sample, and the fourth – to the percentage of function words in the sample.

1	2	3	4
EARLY SHORT STORIES			
1	16,000	3,614	22.59
2	16,000	3,520	22.00
3	16,000	3,522	22.01
4	16,000	3,617	22.61
5	16,000	3,680	23.00
6	2,142	495	22.64
Total:	**82,142**	**18,448**	**22.46**
THE QUIET DON Books I and II, parts 1-5; also the beginning of part 6			
1	16,000	3,154	19.71
2	16,000	3,122	19.51
3	16,000	3,178	19.88
4	16,000	3,137	19.61
5	16,000	3,078	19.24
6	16,000	3,152	19.70
7	16,000	3,135	19.59
8	16,000	3,080	19.25
9	16,000	3,152	19.70
10	16,000	3,097	19.36
11	16,000	3,158	19.74
12	16,000	3,068	19.18
13	16,000	3,168	19.91
Total:	**208,000**	**40,697**	**19.55**
THE QUIET DON Books III and IV, part 6 continued and parts 7-8			
14	16,000	3,534	22.09
15	16,000	3,485	21.78
16	16,000	3,515	21.97
17	16,000	3,609	22.52
18	16,000	3,520	22.01
19	16,000	3,559	22.23
20	16,000	3,752	23.45
21	16,000	3,715	23.22
22	16,000	3,747	23.42
23	16,000	3,758	23.49
24	16,000	3,815	23.84
25	16,000	3,719	23.24
26	16,000	3,626	22.70
27	5,563	1,115	20.43
Total:	**213,563**	**48,466**	**22.69**
WILD LAND CONQUERED Books I and II			
1	16,000	3,645	22.78
2	16,000	3,549	22.18
3	16,000	3,657	22.82
4	16,000	3,679	22.99
5	16,000	3,647	22.75
6	16,000	3,689	23.05
7	16,000	3,730	23.30
8	16,000	3,800	23.78
9	16,000	3,707	23.14
10	16,000	3,735	23.34
11	16,000	3,693	23.08
12	16,000	3,686	23.03
13	16,000	3,786	23.66
14	290	65	22.42
Total:	**208,290**	**48,058**	**23.07**
LATE WORKS (*The Science of Hatred, The Fate of a Man* and *They Fought for the Fatherland*)			
1	16,000	3,980	24.87
2	16,000	3,920	24.50
3	16,000	3,752	23.45
4	16,000	3,959	24.70
5	1,424	338	23.73
Total:	**65,242**	**15,949**	**24.37**
ESSAYS, ARTICLES, CAUSERIES AND SPEECHES			
1	16,000	3,682	23.01
2	16,000	3,797	23.73
3	16,000	3,685	23.03
4	16,000	3,797	23.73
5	3,493	805	23.03
Total:	**67,495**	**15,766**	**23.35**

Bibliography to Annex 3

[f1] Fuchs, W. *Nach allen Regeln der Kunst.* Diagnosen über Literatur, Musik, bildende Kunst. Die Werke, ihre Authoren und Schopfer. Deutsche Verlags-Anstalt., Stuttgart, 1968.

[f2] Fuchs, W. *Mathematical Theory of Word Formation*. – London, 1955.
[f3] Morozov, N. A. *Linguistic Spectra.* Izvestiya Akademii Nauk, Department of Russian and Philology. Books 1-4, Volume XX, 1915.
[f4] Meier, H. *Deutsche Sprachstatistik.* Hildesheim, 1964.
[f5] Markov, A. A. *In re a Possible Use of the Statistical Method.* Izvestiya Akademii Nauk, Series 6, Volume X, Issue 4, 1916.
[f6] Akhmanova, O. S. etc. *On the Exact Methods of Linguistic Studies.* Moscow, 1961.
[f7] Froumkina, R. M. *Statistical Methods of Lexical Studies.* Moscow, 1964.
[f8] Golovin, B. N. *Language and Statistics.* Moscow, 1971.
[f9] Meshcherskiy, N. A. *History of Literary Russian Language.* Leningrad, 1981.
[f10] G. Kjetsaa, S. Gustavsson, B. Beckman, S. Gil. *The Authorship of the Quiet Don.* Russian translation. Moscow, Kniga, 1989. Original: G. Kjetsaa, S. Gustavsson, B. Beckman, S. Gil. *The Authorship of the Quiet Don.* Solum Forlag A. S., Oslo, Humanities Press, New Jersey.
[f11] "D." *The Stirrup of the "Quiet Don". Mysteries of the Novel.* Paris, YMCA Press, 1974.
[f12] R. Medvedev. *Who Wrote the "Quiet Don"?* Paris, Christian Bourg. Edit., 1975.
[f13] Kietsaa, G. *The Battle for the "Quiet Don".* Seanado-Statica, 22, 1976.
[f14] Kietsaa, G. *The Battle for the "Quiet Don".* – Pergamon Press, USA. 1977.
[f15] Yermolayev, G. *Mysteries of the "Quiet Don".* Slavic and European Journal. 18.3.1974.
[f16] Yermolayev, G. *The Authorship of the "Quiet Don".* Slavic and European Journal. 20.3.1976.
[f17] Kolodny, L. *Maelstrom on the "Quiet Don". Fragments of the Past: Sources of a XX Century Slander.* The Moskovskaya Pravda Newspaper, 5 and 7 March 1989.
[f18] A. G. Makarov and S. E. Makarova. *The Melancholy Thistle. Towards the Sources of the "Quiet Don".* Moscow, 1991. Copying facility of the National Egazprom Research Institute.
[f19] Fomenko, A. T. *A Number of Determinate Statistical Traits of Information Density Distribution in Texts, Including Scale.* Semiotika i informatika, Moscow, National Institute of Scientific and Technological Information, 1980, Issue 15, pp. 99-124.
[f20] Fomenko, A. T. *Duplicates in Mixed Sequences and a Frequency Duplication Principle. Methods and Applications. Probability Theory and Mathematical Statistics.* Proceedings of the Fourth Vilnius Conference (24-29 June 1985). VNU Science Press, Utrecht, The Netherlands. 1987, Volume 1, pages 439-465 (in English).
[f21] *Mysteries and Riddles of the "Quiet Don".* Almanac. G. Porfiryev ed., Samara, P. S. Press, 1996.
[f22] M. T. Mezentsev. *The Fate of Novels. (In Re the Discussion of the "Quiet Don's" Authorship Issue).* Samara, P. S. Press, 1994.
[f23] Makarov, A. G. and Makarova, S. E. *Around the "Quiet Don". From Mythology to the Search of Truth.* Moscow, Probel Publications, 2000.