Using Literal and Grammatical Statistics for Authorship Attribution

O. V. Kukushkina A. A. Polikarpov D. V. Khmelev

May 25, 2002

This paper was published in Problemy Peredachi Informatsii, vol.37, number 2, April-June, 2000, pp.96-108. Translated in "Problems of Information Transmission", pp. 172-184

Received August 8, 2000; in final form, January 11, 2001

Abstract

Markov chains are used as a model for the sequence of elements of a natural language text. This model is applied for authorship attribution of texts. An element of a text could be a letter or a grammatical class of a word. It turns out that the frequencies of usage of letter pairs and pairs of grammatical classes are stable characteristics of the author, and they could be used in disputed authorship attribution. A comparison of results with letters and grammatical classes is given. The research is carried out over 385 texts of 82 writers.

In Appendix the research of D.V. Khmelev is described, where data compression algorithms are applied to authorship attribution.

Contents

1 Introduction

In this paper the problem of identification of a text's author is presented as follows. Some large fragments of prose works by a number of authors are given. These texts are in Russian or in a different, written, phonological (non-hieroglyph, non-character) language[1]. An anonymous text known to belong to one of these authors is disputed between all of them. One has to predict the correct author. D.V. Khmelev in [1] has offered and tested statistically a technique giving correct predictions with quite a high probability. The technique chosen is based on the data about usage of subsequent elements in the text (letters, morphemes etc).

To the best of our knowledge, first attempts in the search of a technique for authorship attribution were given in [2]. Markov [3] had an almost immediately replied to [2]. This shows that the founder of the theory of Markov chains was quite interested in this field. Notice also that the first application of "events which are linked to a chain" Markov described in [4], where he studied the distribution of vowels and consonants among initial 20000 letters of "Evgenij Onegin".

Modern methods of authorship attribution are reviewed for our country in [5, Chapter. 1] and a good review of foreign papers is given in [6]. Despite the huge variety of methods described, none of them have been applied to a large number of texts. Often, these methods are not automatic and require some human intervention which makes the computational analysis of a large number of large texts almost impossible. Hence a question of generalization raises: can they be used outside the situation they were developed for?

Until recent times the only exception was the paper [7], where the chosen technique has been tested on a number of texts. They examine the proportion

[1] Written non-character (non-hieroglyph) texts are mentioned, since character written language reduces opportunities in analysis of subsequent elements because of very implicit phonological data (telling apart morphemes)

of function words used by an author and find that this proportion is stable for each author among a large number of Russian writers of eighteenth–twentieth centuries. This technique has been applied in [7] to the problem of determining plagiarism.

A new method of authorship attribution for natural language texts (actually, independent of a language considered), has been offered for the first time by Khmelev in [1].

The new technique is based on the Markov model for the sequence of letters (and any other elements) of a text, i.e., a sequence of letters is considered as a Markov chain, where each letter (element) depends on the preceding letter (element) only.

The matrices of transition frequencies of element (letter, grammatical code, etc) pairs are calculated over all texts by each of the authors. Therefore we know (approximately) the probability of transition from one letter to another for each author. The true author of an anonymous text is calculated using the principle of maximal likelihood, i.e., for each matrix we calculate the probability of the anonymous text and we choose the author with the maximal corresponding probability and the chosen author is assumed to be the true author.

This method is amazingly precise as was shown in [1], where this method was applied to a large number of various texts. The result becomes even more amusing if we recall that only frequencies of letter pairs are taken into account.

Markov models of first and higher orders have been used in a large number of works in 50-60 years of XX century in order to estimate an entropy of various kinds of texts. A lot of those works are described in [8, Ch. 4.3].

But none of works mentioned have raised the question of application of Markov chains to the problem of authorship attribution. Moreover, a generally accepted viewpoint was that characteristics of *any fiction text* measured via letters and letter pairs frequencies are close to average characteristics of the language. Hence these characteristics are indistinguishable from practical viewpoint (see [8, footnote on the p.181] and [9, 10, 3]). A principle of maximal likelihood have never been applied in authorship attribution either because of the following reasons. Firstly, computations on a large amount of data was awkward until appearance of computers and a large number of electronic texts. Secondly, a kind of a psychological barrier existed in an absence of underlying theory, since Markov chain of first order is just a first and a very bad approximation for a natural language text. This fact is pointed out in many papers concerning estimating the texts entropy via Markov chains of high order [8, p.187] Finally, in authorship attribution it was believed that there exists a set of stable quantitative characteristics for grammatical information, and these characteristics are useful in distinguishing the writers. To the best of our knowledge, until the recent times the only significant result for Russian language in this direction was obtained in [7].

In this work we develop further a validation procedure of method [1] and apply the method of [1] for different units of the analysis, i.e., we study:

(a) letter pairs in their natural sequences in a text, i.e., in words (as they appeared in the text) and spaces between them;

(b) letter pairs in sequences of letters in the vocabulary form of words. For example, the previous sentence is reduced as follows "letter pair in sequence of letter in vocabulary form of word". For Russian texts this reduction is much more significant, since Russian words have a number of various forms.

(c) pairs of most generalized ("incomplete") grammatical classes of words and parts of speech in their sequences in sentences of the text. In Russian, 14 parts of speech are traditionally assigned: nouns, verbs, adjectives etc. We introduce the other 4 categories: "end of sentence", "shortening", "unclear class", "dash sign". A category "unclear class" was introduced, since grammatical classes were assigned automatically to more than 99% of words, but some words (for example, misprinted) was not assigned automatically, and hence, their grammatical classes have not been clear.

(d) pairs of less generalized ("complete") grammatical classes of words, for example, animated noun, unanimated noun, qualitative adjective, relative adjective, possessive adjective etc.

A full cross-validation was carried out over 385 texts of 82 writers. Results are presented in Tables 1 and 2. The details of authors and texts are given in Table 3; the size of texts and the number of texts per author are included.

To gauge the precision of the method one can calculate a percent of correctly classified texts. The best results were obtained in cases (a) and (b) (73% and 62% of correct classification, resp.). 61% of texts were correctly classified in case (c). Results for case (d) are much worse (4%).

In Section 2 the principles and results of preprocessing are described. In Section 3 the procedure of cross-validation is described. A detailed description of results is given in Section 4 with conclusions presented in Section 5. In Appendix written by D.V. Khmelev, another approach to authorship attribution is presented. This approach uses data compression algorithms.

2 Pre-processing

After pre-processing the source corpora of texts was presented in all forms (a)–(d).

In case (a) all words of unclear grammatical class were omitted, i.e., words, not recognized automatically, were ignored (and this is the main difference of present research from [1]). These words are ignored, since we are now able to compare results in case (a) with results in case (b). Also, text is converted into a sequence of words and delimiting spaces without any formatting. All punctuation is removed. Finally, words beginning with a capital letter are also omitted (in particular, all the first words of sentences are removed). It was shown in [1] that the last trick significantly increases the precision of authorship attribution. There is a conjecture that this improvement is implied by ignoring of names of characters, that are usually not related to the style of the author of the fiction text. A letter "ë" in Russian alphabet was glued to the letter "e", and hence we had 33 characters including the space. Each letter was encoded with its number: letter "a" corresponding to 1, letter "ya" corresponding to 32. The space is corresponding to 0. The total number of letters in all texts is 96209964. The total number of different pairs of letters found in the corpora of text is 1011 (out of possible $33 \times 33 = 1089$). Clearly, 1011 is larger than an actual number of different letter pairs to be found in Russian texts. This is a result of a number of misprints in electronic versions of books and this fact can lead to errors in computations. In order to obtain an estimate of this noise, 121 letter pairs were selected, those whose appearance in a Russian text was unlikely. In total, they were used 38495 times or 0,04% of total size of the corpora, and hence they could be disregarded.

The pre-processing of source corpora in cases (b), (c) and (d) is carried out with help of automatic classifier developed by O.V. Kukushkina at the Laboratory for General and Computational Lexicology and Lexicography of the Philological faculty at Lomonosov Moscow State University. The classifier is based on the grammar vocabulary of Zaliznyak [11] and Academic grammar [12, 13].

Case (c) is based on the information obtained from the most general grammatical class of word form, i.e., we use only the information concerning the part of the speech (particles, prepositions, interjections, copulatives and adversative conjunctions, other conjunctions, verbs, pronouns, adverbs, adjectives, nouns, numerals, predicate nouns, comparatives, modal adverbs).

Case (d) is based on the information obtained from lexical and grammatical category of the given part of speech (animated noun, unanimated noun etc)

Let us present some data from cases (b)–(d).

In case (b) the total number of letter is 110704464. There are 1029 different pairs of letters out of $33 \times 33 = 1089$. As one can see, figures are increased w.r.t. case (a). This effect is related to conversion of oblique forms to vocabulary ones.

In cases (c) and (d) the number of elements is 20262449. In case (c) the total number of different pairs of elements is 302 out of possible $18 \times 18 = 324$. In case (d) the number of different pairs of elements is 8124 out of possible $112 \times 112 = 12544$.

3 Technique and its cross-validation

An analysis of texts in each case was carried out based on software developed by Khmelev. For each case we present results of cross-validation of the method of [1]. Cross-validation is carried out as follows.

Let us recall that elements of texts are encoded with numbers from 0 to 32 (in cases (a) or (b)), 17 or 111 (in cases (c) and (d), respectively). Code 0 always corresponds to a delimiter between large blocks, i.e., in cases (a) and (b) code 0 corresponds to the space between words and in cases (c) and (d) code 0 corresponds to delimiter between sentences ("end of sentence").

Given W writers each of which has N_w texts, where $w = 0, \ldots, W-1$, we count Q_{ij}^{wn} which is the number of transitions from letter i to letter j for text n ($n = 0, \ldots, N_w - 1$) from writer w ($w = 0, \ldots, W-1$). In order to find the predicted author for text $\hat{n}$ (of known author $\hat{w}$) using information about authorship of the other texts for all authors including $\hat{w}$, we have

$$Q_{ij}^{k} = \sum_{n=0}^{N_w-1} Q_{ij}^{kn}, \quad Q_i^k = \sum_j Q_{ij}^k$$

for authors $k \neq \hat{w}$, and for author $\hat{w}$ we exclude text $\hat{n}$ from a training set

$$Q_{ij}^{\hat{w}} = \sum_{n \neq \hat{n}} Q_{ij}^{\hat{w}n}, \quad Q_i^{\hat{w}} = \sum_j Q_{ij}^{\hat{w}}.$$

We then have

$$\Lambda_k(\hat{w},\hat{n}) = -\sum_{i:Q_i^k>0}\sum_{j:Q_{ij}^k>0} Q_{ij}^{\hat{w}\hat{n}} \ln\frac{Q_{ij}^k}{Q_i^k}$$

and

$$\Lambda_{\hat{w}}(\hat{w},\hat{n}) = -\sum_{i:Q_i^{\hat{w}}>0}\sum_{j:Q_{ij}^{\hat{w}}>0} Q_{ij}^{\hat{w}\hat{n}} \ln\frac{Q_{ij}^{\hat{w}}}{Q_i^{\hat{w}}}.$$

Ignoring degenerate cases $Q_{ij}^k = 0$ and $Q_i^k = 0$, one can see that each $\Lambda_k(\hat{w},\hat{n})$ is minus the logarithm of the probability of the text $\hat{n}$ being from writer $\hat{w}$, when the text $\hat{n}$ from writer $\hat{w}$ is generated by the Markov chain with transition probabilities $P_{ij}^k = Q_{ij}^k/Q_i^k$. The hint to ignore the degenerate summands is given by results about optimal maximal likelihood estimate, presented in [14, p.224].

We also define a rank $R_k(\hat{w},\hat{n})$ to be the rank of $\Lambda_k(\hat{w},\hat{n})$ in $\{\Lambda_k(\hat{w},\hat{n}), k = 0,\ldots,W-1\}$, where the smallest rank is 0, i.e., $R_k(\hat{w},\hat{n}) \in \{0,\ldots,W-1\}$, and the smallest number has the smallest rank. If the text is assigned to the correct author, then $R_{\hat{w}}(\hat{w},\hat{n}) = 0$.

If the text is assigned to an other author, and the correct author is second among other pretenders, then $R_{\hat{w}}(\hat{w},\hat{n}) = 1$ etc.

A result of cross-validation is a set of ranks

$$\{R_{\hat{w}}(\hat{w},\hat{n})\}_{\hat{w}\in\{0,\ldots,W-1\},\hat{n}\in\{0,\ldots,N_{\hat{w}}-1\}}.$$

The precision of the technique for authorship attribution is measured by this set of numbers. The proportion of correct predictions is the proportion of zero ranks. When the true author was close to being correctly predicted, we have small ranks. Another measure of the precision of the technique for authorship attribution is given by the average rank

$$M = \frac{1}{\sum_{\hat{w}=0}^{W-1} N_{\hat{w}}} \sum_{\hat{w}=0}^{W-1}\sum_{\hat{n}=0}^{N_{\hat{w}}-1} R_{\hat{w}}(\hat{w},\hat{n}). \tag{1}$$

We shall also present results of cross-validation in case of analysis of individual letters (in cases (a) and (b)) and individual grammatical classes (in cases (c) and (d)). All the calculations are the same, but we additionally calculate

$$Q^k = \sum_i Q_i^k \quad \text{and} \quad Q^{\hat{w}} = \sum_i Q_i^{\hat{w}}$$

Table 1: Full cross-validation of the authorship attribution based on the sequence of letters

Case (a)				Case (b)		
Words as appeared in the text				Words in vocabulary form		
R	Letters pairs	Individual		R	Letter pairs	Individual
0	282/385	27/385		0	240/385	12/385
1	21/385	55/385		1	29/385	40/385
2	9/385	25/385		2	17/385	19/385
3	5/385	24/385		3	9/385	16/385
4	5/385	17/385		4	6/385	16/385
≥ 5	63/385	237/385		≥ 5	84/385	282/385
M	3,38	12,69		M	4,77	17,88

and the following quantaties $\Gamma_k(\hat{w}, \hat{n})$ are used instead of $\Lambda_k(\hat{w}, \hat{n})$:

$$\Gamma_k(\hat{w}, \hat{n}) = - \sum_{i:Q_i^k>0} \left(\sum_j Q_{ij}^{\hat{w}\hat{n}} \right) \ln \frac{Q_i^k}{Q^k}$$

and

$$\Gamma_{\hat{w}}(\hat{w}, \hat{n}) = - \sum_{i:Q_i^{\hat{w}}>0} \left(\sum_j Q_{ij}^{\hat{w}\hat{n}} \right) \ln \frac{Q_i^{\hat{w}}}{Q^{\hat{w}}}.$$

4 Description of the results

The results of the research are presented in Tables 1, 2. The column R corresponds to ranks $0, \ldots, 4$ and ranks exceeding 4. Row of rank 0 contains the proportion of correctly assigned texts. Row of rank 1 contains the proportion of texts such that the correct author was the second among the other pretenders etc. Finally, row of rank ≥ 5 contains the proportion of texts such that the correct author was on the place not better than 6.

Line M contains the average rank, determined by (1).

Straight away we can notice that the frequencies of individual letters and individual grammatical classes give a low but significantly non-random level

Table 2: Full cross-validation of the authorship attribution based on the sequence of grammatical classes

Case (c)				Case (d)		
Generalized grammatical classes				"Complete" grammatical classes		
R	Pairs	Individual		R	Pairs	Individual
0	235/385	128/385		0	15/385	6/385
1	31/385	43/385		1	21/385	12/385
2	16/385	29/385		2	9/385	6/385
3	8/385	15/385		3	12/385	6/385
4	11/385	17/385		4	19/385	8/385
≥ 5	84/385	153/385		≥ 5,	309/385	347/385
M	5,43	10,13		M	17,76	31,93

of correct authorship attribution (except the case of individual "complete" grammatical classes, the reasons of bad results are to be studied).

All calculation with pairs of elements (letters or grammatical classes) give better results w.r.t. calculations with individual letters or grammatical classes.

Let us now study the difference in results of analysis using information about usage of letter pairs in words as they appeared in the text and letter pairs in vocabulary forms of words. Comparison shows that the success rate of the authorship attribution is better in the case of natural sequences of letters in the text (there are 73% against 62% of correct authorship predictions and average rank is 3,38 against 4,77). Perhaps, an "equalization" of different forms of a word reduced to the vocabulary form eliminates some information useful for authorship attribution and leads to the fall in success rate in case (b) w.r.t. case (a).

Comparison of results of authorship attribution using the information about the pairs of letters and pairs of generalized ("incomplete") grammatical classes (cases (a) and (c)) shows that the success rate in both cases is quite high, i.e., we have 73% and 61% correctly classified texts, respectively. The average rank of correct authorship attribution falls in 2 ranks from letter pairs to generalized grammatical classes. Perhaps, a relatively lower efficiency of pairs of grammatical classes (in comparison with letter pairs for words as they appeared in the text) on the same data set is concerned with smaller magnitude of sample size on the grammatical classes (let us recall that on the same corpora of texts there are 96 million of letter usages in case (a) and there are 20 million of grammatical classes usages in case (c)).

Also, it requires attention that the usage of individual grammatical classes in case (c) (see Table 2) is significantly more effective than usage of individual letters: there are 33% of correct authorship attributions in case (c) against 7% in case (a). The average rank in case (c) is two and a half times less than the average rank in case (a). Perhaps, this result is concerned with more specific information (more precisely characterising stable structure characteristics of texts for each author) given by individual grammatical codes in contrast to individual letters.

"Complete" grammatical classes were the most ineffective in authorship attribution both in the case of pairs frequencies and in the case of individual frequencies. The reasons for these results are to be studied.

The details about authors of the corpora of texts, the number of texts per author, the dispersion of text sizes with minimal, average (in brackets) and maximal sizes, results of authorship attribution in all four cases ((a)–(d)) with minimal, average (in brackets) and maximal rank of the text examined on its author are given in Table 3.

The corpora also containes some translated texts (S. Lem, I. Khmelevskaja, B. Rajnov). It is interesting that the quality of attribution of these texts is not worse than texts written by authors whose mothers language is Russian (although corpora contains Lem's translation by different interpreters)

Table 3. The number of and size of texts in corpora by author

w	Writer	N_w	The size of texts (thousands letters)	(a)	(b)	(c)	(d)
0	O. Avramenko	7	223,7(279,5)395,1	0(0,0)0	0(0,0)0	0(0,0)0	10(13,9)22
1	A. Bol'nykh	7	0,8(185,0)298,8	0(4,1)24	0(5,9)37	1(12,1)79	4(15,3)75
2	K. Bulychev	16	3,3(129,5)458,9	0(6,8)59	0(6,8)53	0(9,0)65	3(16,3)69
3	A. Volkov	9	5,2(186,8)610,5	0(20,4)50	0(26,3)51	0(12,3)57	5(23,1)65
4	G. Glazov	6	184,5(263,7)326,1	0(0,0)0	0(0,0)0	0(0,0)0	9(13,2)19
5	O. Grinevskij	2	96,1(127,4)158,6	0(0,0)0	0(0,0)0	0(0,0)0	0(0,5)1
6	N.V. Gogol'	4	97,7(213,3)334,0	0(1,0)4	0(1,5)6	0(8,0)32	14(22,8)42
7	N. Gumilev	2	70,1(70,6)71,0	0(0,0)0	0(0,0)0	0(0,0)0	0(0,0)0
8	F.M. Dostoevskij	4	88,6(175,5)268,9	0(0,0)0	1(2,0)3	0(0,3)1	10(13,3)18

Continued on the next page

w	Writer	N_w	The size of texts (thousands letters)	(a)	(b)	(c)	(d)
9	M. and S. Djachenko	6	23,3(325,1)553,2	0(0,0)0	0(0,2)1	0(0,0)0	7(11,5)17
10	S. Esenin	2	44,6(131,5)218,4	0(0,0)0	0(0,0)0	0(0,0)0	0(0,5)1
11	A. Etoev	6	2,7(57,9)114,8	0(1,0)4	0(4,3)19	0(2,2)13	2(3,0)5
12	I. Efremov	2	256,5(396,5)536,5	0(0,0)0	0(0,0)0	0(0,0)0	1(2,0)3
13	A. Zhitinskij	3	253,6(793,2)1207,6	0(0,3)1	0(0,0)0	0(9,0)26	18(26,3)42
14	A. Kabakov	5	69,0(225,5)418,4	0(0,0)0	0(0,2)1	0(2,6)11	15(19,2)26
15	S. Kazmenko	5	132,8(400,5)1148,3	0(1,2)5	0(0,8)4	0(0,2)1	16(21,8)28
16	V. Kaplan	7	19,3(91,9)305,2	0(4,1)25	0(5,6)24	0(5,0)23	9(23,0)40
17	A. Kac	2	81,7(461,4)841,0	0(0,0)0	0(0,0)0	0(0,0)0	1(2,5)4
18	V. Klimov	4	58,5(107,5)179,9	0(7,0)15	0(7,0)20	0(1,5)6	3(6,8)9
19	E. Kozlovskij	2	848,6(868,4)888,2	0(0,0)0	0(2,0)4	15(42,0)69	40(56,5)73
20	I. Krashevskij	3	380,6(555,2)803,1	0(0,0)0	0(0,0)0	0(0,0)0	6(8,3)10
21	I. Kublickaja	2	170,2(226,2)282,3	0(0,0)0	0(0,0)0	0(0,0)0	1(2,0)3
22	L. Kudrjavcev	4	108,3(190,5)348,2	0(0,3)1	0(1,5)5	0(0,0)0	8(13,5)24
23	V. Kunin	4	296,3(407,9)610,3	0(0,0)0	0(2,5)7	0(3,5)5	10(15,5)23
24	A. Kurkov	7	17,5(121,0)276,9	0(3,1)10	0(11,6)28	0(1,9)3	4(11,9)19
25	A. Lazarevich	4	11,3(101,3)274,7	0(14,3)47	5(20,3)54	2(11,0)18	4(9,3)15
26	A. Lazarchuk	6	141,4(434,2)786,9	0(0,0)0	0(0,8)2	0(0,0)0	19(25,5)34
27	Ju. Latynina	3	116,8(970,7)2511,8	0(3,3)10	0(13,0)36	0(0,0)0	4(15,3)23
28	S. Lem	8	11,6(238,6)535,2	0(0,9)5	0(1,1)8	0(5,1)27	11(26,5)44
29	N. Leonov	3	273,1(282,7)295,7	0(0,0)0	0(0,0)0	0(0,0)0	3(4,0)5
30	S. Loginov	14	1,3(153,4)916,2	0(15,9)36	4(18,1)37	0(18,9)49	14(35,6)59
31	E. Lukin	5	26,9(144,6)367,9	0(3,2)15	0(0,0)0	0(4,4)19	8(16,0)39
32	L. and E. Lukiny	4	105,2(239,9)564,7	0(0,3)1	2(3,8)6	0(0,5)2	4(12,8)20
33	S. Luk'janenko	15	6,0(277,6)542,9	0(3,0)22	0(6,9)76	0(9,1)58	9(25,4)73
34	N. Markina	2	93,6(179,8)266,0	0(0,0)0	0(0,0)0	1(2,5)4	0(1,5)3
35	A. Melikhov	2	457,6(536,4)615,2	0(0,0)0	0(2,5)5	0(0,0)0	17(17,5)18
36	V. Mikhajlov	2	84,2(169,3)254,5	0(0,0)0	0(0,0)0	0(0,0)0	1(2,5)4
37	A. Molchanov	2	206,5(302,4)398,3	0(0,0)0	0(0,0)0	0(0,5)1	4(5,5)7
38	V. Nabokov	6	102,0(310,6)599,8	0(2,0)11	0(0,8)3	0(3,0)15	5(12,7)18
39	M. Naumova	4	5,2(161,1)337,8	0(7,8)31	0(11,8)47	0(17,8)69	4(10,0)17
40	Ju. Nesterenko	2	71,1(212,0)352,8	0(1,0)2	1(3,5)6	0(0,0)0	1(3,5)6
41	Ju. Nikitin	3	656,9(681,4)702,2	0(11,3)34	0(17,0)51	0(0,7)1	5(6,7)8
42	S. Pavlov	2	375,6(414,5)453,4	0(0,0)0	0(0,5)1	0(0,0)0	6(6,5)7
43	A.S. Pushkin	2	57,1(113,7)170,3	0(0,0)0	0(0,0)0	0(0,0)0	0(0,0)0

Continued on the next page

w	Writer	N_w	The size of texts (thousands letters)	(a)	(b)	(c)	(d)
44	B. Rajnov	5	267,7(363,6)420,3	0(0,0)0	0(0,0)0	0(0,6)3	12(13,8)15
45	L. Reznik	2	79,6(97,8)115,9	0(0,0)0	0(0,0)0	0(0,0)0	1(1,0)1
46	N. Rerikh	4	84,5(305,6)608,7	0(5,5)22	0(3,5)14	0(2,3)9	0(2,5)9
47	N. Romaneckij	7	5,5(203,2)530,6	0(5,4)21	0(7,7)20	0(1,0)4	15(27,7)45
48	A. Romashov	2	87,7(88,1)88,4	0(0,0)0	0(0,0)0	2(3,0)4	0(0,0)0
49	V. Rybakov	7	9,7(119,5)366,1	0(9,0)24	0(9,0)21	0(16,4)36	15(25,6)41
50	M.E. Saltykov-Schedrin	2	101,6(170,4)239,1	0(0,0)0	0(0,0)0	0(0,0)0	1(2,5)4
51	R. Svetlov	3	29,2(241,0)425,4	0(0,0)0	3(13,3)20	0(0,7)2	3(8,7)17
52	A. Sviridov	4	13,4(224,0)601,5	10(27,5)65	0(11,8)41	0(11,0)44	6(26,5)56
53	V. Segal'	3	60,5(132,0)259,7	0(0,0)0	0(0,0)0	0(0,3)1	1(4,7)8
54	K. Serafimov	2	75,3(130,8)186,4	0(0,0)0	0(0,0)0	0(0,0)0	1(1,0)1
55	I. Sergievskaja	2	50,7(79,8)108,9	0(0,0)0	0(1,0)2	0(0,0)0	0(0,5)1
56	K. Sitnikov	8	13,0(66,1)274,3	0(0,0)0	0(2,3)18	0(0,5)4	3(8,3)22
57	S. Snegov	3	385,8(411,1)438,4	0(0,0)0	0(0,0)0	0(0,0)0	5(5,0)5
58	S.M. Solov'ev	2	159,9(1251,6)2343,3	0(0,0)0	0(0,0)0	0(0,0)0	0(2,5)5
59	A. Stepanov	6	83,7(219,6)390,3	0(0,0)0	0(0,0)0	0(0,0)0	4(5,0)8
60	A. Stoljarov	2	137,2(241,9)346,7	0(5,0)10	9(11,0)13	0(7,5)15	1(2,5)4
61	A. and B. Strugackie	30	37,1(230,4)579,5	0(1,9)24	0(2,5)23	0(5,9)54	15(36,4)79
62	A. Strugackij	2	51,9(101,6)151,3	0(0,0)0	0(0,0)0	0(0,5)1	1(1,0)1
63	B. Strugackij	2	260,7(279,6)298,4	0(0,0)0	0(0,0)0	0(0,0)0	7(8,0)9
64	E. Til'man	3	307,8(390,0)464,7	0(0,0)0	0(0,0)0	0(0,0)0	11(11,3)12
65	A. Tolstoj	2	97,9(113,8)129,7	0(0,0)0	0(0,0)0	0(0,0)0	0(0,0)0
66	L.N. Tolstoj	2	199,9(712,5)1225,1	0(0,0)0	0(0,0)0	0(0,0)0	1(1,5)2
67	D. Truskinovskaja	9	82,6(235,9)478,6	0(0,8)3	0(2,7)12	0(3,8)28	13(23,8)53
68	A. Tjurin	19	1,3(222,0)832,7	0(2,3)20	0(1,2)13	0(2,6)25	24(34,0)55
69	E. Fedorov	2	221,3(667,2)1113,1	0(0,0)0	0(0,0)0	0(0,0)0	1(1,5)2
70	E. Khaeckaja	3	204,1(309,0)414,3	1(10,3)22	12(31,3)42	54(57,0)62	28(39,3)54
71	D. Kharms	3	13,9(104,1)185,5	0(0,0)0	0(0,0)0	0(12,0)29	5(16,7)30
72	V. Khlumov	4	183,3(242,9)395,5	0(3,8)15	0(11,3)38	6(15,3)38	26(34,0)46
73	I. Khmelevskaja	5	203,7(345,7)459,1	0(0,0)0	0(0,0)0	0(0,4)2	9(12,8)18
74	V. Chernjak	3	201,6(373,7)501,0	0(0,0)0	0(2,7)8	0(11,7)35	2(11,0)25
75	A.P. Chekhov	3	247,9(335,3)414,5	0(0,0)0	0(0,0)0	4(13,3)20	18(18,7)20
76	V. Shinkarev	3	56,2(78,9)100,1	0(11,7)29	6(13,3)22	4(23,7)61	5(5,7)6
77	V. Shukshin	2	66,7(187,7)308,8	0(0,0)0	0(0,0)0	0(0,0)0	1(2,5)4
78	S. Scheglov	3	55,2(103,0)146,1	0(4,0)12	0(13,7)41	0(3,0)9	2(2,3)3

Continued on the next page

w	Writer	N_w	The size of texts (thousands letters)	(a)	(b)	(c)	(d)
79	A. Schegolev	3	105,6(318,0)561,7	0(0,0)0	0(0,0)0	0(0,0)0	12(18,7)32
80	V. Jugov	6	66,7(149,2)304,3	0(0,5)2	0(0,7)2	0(1,8)10	8(10,7)18
81	V. Jan	2	507,3(553,9)600,4	0(0,0)0	0(0,0)0	0(0,0)0	2(2,0)2

5 Conclusion

The main result of the carried-out research is that the usage of grammatical information in authorship attribution is not only useful, but efficient and is even comparable with the usage of information about letter pairs frequencies (the effectiveness of the last method was shown before in [1]).

At the same time it is still amazing that the usage of such a seemingly simple unit as a pair of subsequent letters in the text gives more precise results than the usage of such a language categories as individual grammatical codes and their pairs. Perhaps, letter pairs contain a kind of converted and incomplete information about structure of morphemes of words as they appear in a text (prefixes, roots, suffixes and inflexions). Therefore, a lot of information about words changing and formation is contained in statistics of usage of letter pairs and this leads to quite a high efficiency of letter pairs statistics for authorship attribution.

In other words, the letter pairs frequencies take into account the vocabulary used by the author and by implication it takes into account information about preferred grammatical structures. Although the differences in usage of particular pairs of letters are likely to be non-significant (since they are converging to average frequencies for the language, as it was noticed by Markov [3] long ago), the maximal "likelihood" takes into account the "total" effect in changing of the usage of a pair of letters and nevertheless it provides a high precision in assignment of a text to the correct author, as it was shown before in [1] and as it was approved in this research by full cross-validation.

However, the further experiments with grammatical classes of words with more precise grammatical analysis would probably lead to a higher success rate in assigning the correct author than it was achieved in this research. Perhaps, the usefulness of usage of grammatical information is shown in the observation that the usage of information on individual generalized grammatical classes is significantly more effective than the usage of information on just the individual grammatical classes.

Since results provided using of different units (letters and generalized grammatical classes) are compatible with each other, one can assume that future well-developed methods for authorship attribution will use different representation of the text, obtained with these units, for mutual cross-validation of results.

A Appendix. Application of Data Compression Algorithms in Authorship Attribution

In this Appendix it is shown how one could use data compression algorithms for authorship attribution. The results of cross-validation for this new method are also given. Here we use the corpora of texts used in [1] and in the main body of the article.

The corpora of texts in [1] is obtained from 82 authors. One randomly-chosen text from each author is held out to make up a test set. The other texts are used as the learning sample. Afterwards all the control texts were classified and the correct author was assigned in 69 cases. To estimate how good this result is in terms of probability of correct authorship attribution let us consider the following hypothetical situation. Suppose that we have a black box such that given two texts it produces 1 if these texts are of the same author and 0 if these texts are definitely of different authors. Suppose that the level of alpha and beta errors for each trial is $0 < p < 1$. One can apply the black box assigning an anonymous text to the correct author as follows. Given 82 alternatives and one anonymous text, this black box should produce 81 zeros corresponding to comparison of the control text to learning samples of wrong authors and only one 1 corresponding to comparison of the control text to the correct author. Of course, all other outcomes are considered as misclassification. Assume that results for each of these pairwise comparisons are independent of each other, then, the probability of correct classification is $(1-p)^{82}$. For alpha-beta error $p = 0{,}05$ we have $(1 - 0.05)^{82} \approx 0.015$, $p = 0{,}01$ corresponds to $(1 - 0{,}01)^{82} \approx 0{,}439$ and for $p = 0{,}005$ we obtain $(1-0{,}005)^{82} \approx 0{,}663$. Notice that $69/82 \approx 0{,}84$ and if we want to surpass the method of Khmelev (2000) in quality of recognition, we should require that, for example, $p = 0{,}001$ and it turns out that $(1-0{,}001)^{82} \approx 0{,}921$. Therefore 69 correct classifications with choice among 82 writers should be considered as an extremely good result. With certain reservations this argumentation is applicable to the results of the main text of the article as well.

Here by a text we understand a sequence of letters from some alphabet $\mathcal{A}$. Denote by $|B|$ the length of text B. Let us call a *concatenation* of the texts B and A the sequence S of length $|B| + |A|$ such that first $|B|$ letters of S coincide with B and the last $|A|$ letters of S coincide with A. We shall write $S = B.A$.

Now let us give an "ideal" definition of relative complexity following definition of Kolmogorov complexity (see [15, 16]): the relative complexity $K(A \mid B)$ of text A w.r.t. text B is the length of the shortest program in binary alphabet which translates text B to the text A. Unfortunately, $K(A \mid B)$ is not computable and it is not clear how one can use it.

A first approximation to $K(A \mid B)$ (however, it is enough for authorship attribution as we shall see later) could be obtained from data compression algorithms. Let us define the *relative complexity* $C(A \mid B)$ of text A w.r.t. text B as follows. Compress text B to a text B' and text $S = B.A$ to a text S'. Now put $C(A \mid B) = |S'| - |B'|$. This definition is ambiguous since we have not fixed the data compression algorithm. All algorithms used in this research are described later.

We shall apply $C(A \mid B)$ for authorship attribution. Given texts from n authors, form a test set by holding out one control text from each author U_1, $\ldots$, U_n. The other texts of each author are concatenated in texts T_1, $\ldots$, T_n.

The predicted author for text U_i is determined as follows. Firstly, find the rank R_i of the number $C(U_i \mid T_i)$ in the set $\{C(U_i \mid T_1), \ldots, C(U_i \mid T_n)\}$, where ranks take values from 0 to $n-1$. If rank R_i equals 0, then the control text U_i is correctly assigned.

Like done in [1] one can introduce different measures of precision for the method of disputed authorship resolution. For example,

1. The simplest measure — the number of zero ranks R_i;
2. More generalized measure is given by an *average rank*

$$M = \frac{1}{n} \sum_{i=1}^{n} R_i.$$

Cross-validation of different data compression algorithms is carried out on the corpora of texts used in [1] and in the main body of this paper. Let us recall, that the corpora consists of 385 texts from 82 writers. The total size of texts is about 128 million letters. Some pre-processing of texts was carried out. Firstly, all words carried over to the next line were restored. Further, words beginning with capital letter were omitted. The other words were kept in the order as they appeared in the text. They were delimited by new-line character. A control text U_i was selected from the texts of each of $n = 82$ authors. The other texts of each author were concatenated into texts T_i, $i = 1, \ldots, 82$. The size of each U_i was not less that 50–100 thousand letters.

Let us consider lossless data compression algorithms. The following algorithms have been most popular recently: Huffman coding, arithmetic coding, Burrows-Wheeler technique [17] and many variations of Lempel-Zip coding [18]. Some algorithms are specially aimed on text coding: these are PPM [19] (this algorithm uses Markov model of small order) and DMC [20] (this algorithm uses dynamical Markov Coding). Each algorithm has a huge number of variations and parameters (for example, there exist so-called dynamic Huffman coding, also the size of vocabulary varies etc). Moreover, there exists a lot of "mixed" algorithms, when a text obtained by PPM compression is additionally compressed by Huffman coding.

All these algorithm are found in a large variety of compression programs, whose number exceeds 150. Each of those programs implements different data compression algorithm. More variety appears because of multiple versions of data compression programs with different data compression algorithms. All programs selected for this research are given in Table 4.

Most of these programs with their descriptions can be obtained from Archiver index[2] supported by Jeff Gilchrist[3]. Program `compress` was taken from SunOS 5.6 operating system. Program `dmc` is available by ftp[4]. Notice that the program `dmc` has an option of maximal memory to use. In this research this option was set to 100000000 bytes. LDS 1.1 is a lossless data compression sources kit and it is also available by ftp[5]. Program `ppm` is available from the personal page of its author[6].

The results of application of these programs are given in Table 5. The last line of Table 5 contains the results of application of Markov chains [1] to the same corpora of texts. Computations for data shown in Table 5 were

[2] http://web.act.by.net/~act/act-index.html

[3] Jeff Gilchrist, jeffg@cips.ca

[4] http://plg.uwaterloo.ca/~ftp/dmc/

[5] ftp://garbo.uwasa.fi/pc/programming/lds_11.zip

[6] http://www.cs.waikato.ac.nz/~wjt/

Table 4: Data compression programs

Program	Author	Algorithm used
1. `7zip` version 2.11	Igor Pavlov	Arithm. coding, LZ + arithm. coding, PPM
2. `arj` version 2.60	RAO Inc.	LZSS + Huffman
3. `bsa` version 1.9.3	Sergey Babichev	LZ
4. `bzip2`	Julian Seward	Burrows-Wheeler + Huffman
5. `compress`	Sun Inc.	LZW
6. `dmc`	Gordon V. Cormack	DMC
7. `gzip`	Jean-loup Gailly	Shannon-Fano, Huffman
8. `ha` version 0.999c	Harri Hirvola	Sliding window dictionary + Arithm. coding, Finite contex model + Arithm. coding
9. `huff1`	William Demas (LDS 1.1)	static Huffman
10. `lzari`	Haruhiko Okumura (LDS 1.1)	LZSS+Arithm. coding
11. `lzss`	Haruhiko Okumura (LDS 1.1)	LZSS
12. `ppm`	William Teahan	PPM
13. `ppmd5` version F	Dmitry Shkarin	PPM
14. `rarw` version 2.00	Eugene Roshal	LZ77 variant + Huffman
15. `rar` version 2.70	Eugene Roshal	LZ77 variant + Huffman
16. `rk` version 1.03α	Malcolm Taylor	Reduced-offset LZ, PPMZ

performed under different operation systems on different types of computers and took about three weeks of non-stop computing.

It follows from the data of Table 5 that data compression algorithms assign correct author to control text quite often. Therefore they are undoubtfully useful. Notice that application of the program `rarw` yields even better results than results obtained with help of Markov chains in [1]. Although such a superiority could be related to some statistical mistake, it is the best result achieved recently.

Perhaps an explanation of such a splendid results is as follows. Data compression algorithms actually adapt well to the control text after pre-

Table 5: The quality of authorship attribution for data compression algorithms

Program	Rank						
	0	1	2	3	4	≥ 5	M
`7zip`	39	9	3	2	3	26	6,43
`arj`	46	5	2	7	2	20	5,16
`bsa`	44	9	3	1	1	24	5,30
`bzip2`	38	5	5	1		33	13,68
`compress`	12	1	1	3	2	63	24,37
`dmc`	36	4	3	4	4	31	9,81
`gzip`	50	4	1	2	1	24	4,55
`ha`	47	8	1	3	3	20	5,60
`huff1`	10	11	4	4	2	51	15,37
`lzari`	17	5	4	2	6	48	14,99
`lzss`	14	3	1	1	3	60	20,05
`ppm`	22	14	2	1	3	40	10,39
`ppmd5`	46	6	6	2		22	5,96
`rar`	58	1	1	1		21	7,22
`rarw`	**71**	**3**		**2**	**1**	**5**	**1,44**
`rk`	52	9	3	1		17	4,20
Markov Chains (see [1])	69	3	2	1		7	2,35

processing of texts of correct author and adaptation is not so good if the texts of a different author have been pre-processed. The disadvantage of this method with respect to method of Markov chains [1] is that data compression algorithms are not so clear and that they are not even available for study in commercial software. Nevertheless, many programs among presented in Table 4 have an open source code and well-described open algorithms. A further study of those programs could make fairly clear the reasons for the efficiency of this relative complexity approach to authorship attribution.

Notice that the authorship attribution technique described here does not require any special programs when one chooses the correct author among small number of pretenders. A great advantage of the method presented is its availability on almost every computer, since most of compression programs mentioned here are widely spread and some of them (like `gzip` or `rar`) are implemented on all types of computers and on all operation systems.

REFERENCES

1. Khmelev D.V. Using Markov chains for authorship attribution, *Vestn. MGU, ser. 9, Filolog.,* 2000, no. 2, pp. 115–126.

2. Morozov N.A. Linguistic spectrums *Izv. otd. russkogo jazyka i slovesnosti Imp.akad.nauk*, 1915, vol. 20, no. 4.

3. Markov A.A. On some application of statistical method, *Izv.Imp.Akad.Nauk., Ser. 6,* 1916, no. 4, pp. 239–242.

4. Markov A.A., An example of statistical study on text of Eugeny Onegin illustrating the linking of events to a chain, *Izv.Imp.akad.nauk, Ser. 6,* 1913, no. 3, pp. 153–162.

5. *Ot Nestora do Fonvizina. Novye metody opredelenija avtorstva* (From Nestor to Fonvizin. New methods for authorship attribution), Moscow: Progress, 1994.

6. Holmes D.I. The Evolution of Stylometry in Humanities Scholarship, *Literary and Linguistic Computing*, 1997, vol. 13, no. 3, pp. 111–117.

7. Fomenko V.P., Fomenko T.G. Author's Quantative Invariant for Russian Fiction Texts, *Metody kolichestvennogo analiza tekstov narrativnykh istochnikov*, Moscow: Inst. Istorii SSSR, 1983, pp. 86–109.

8. Yaglom A.M. and Yaglom I.M. *Verojatnost' i informacija*, M.: Nauka, 1960. Translated under the title *Probability and Information*, Boston: Reidel, 1983.

9. Dobrushin R.L. Mathematical methods in linguistics, *Matematicheskoe prosveschenie*, 1959, no.6.

10. Shannon C.E. and Weaver W. *The Mathematical Theory of Communication*, Urbana: Univ. of Illinois Press, 1949.

11. Zaliznjak A.A. *Grammaticheskij slovar' russkogo jazyka* (Grammatical dictionary for Russian language), Moscow: Rus. jaz., 1977.

12. *Grammatika sovremennogo russkogo literaturnogo jazyka* (The grammar of contemporary Russian language), Moscow: Nauka, 1970.

13. *Russkaja grammatika* (Russian Grammar), 2 vols, Moscow: Nauka. 1980.

14. Ivchenko G.I., Medvedev Yu.I. *Matematicheskaja statistika* (Mathematical Statistics), Moscow: Vyssh. shk., 1992.

15. Li M., Vitányi P. *An Introduction to Kolmogorov Complexity and Its Applications*, New York: Springer, 1997.

16. Kolmogorov A.N. Three approaches to the quantitative definition of information, *Probl. peredachi inform*, 1965, vol. 1. no. 1, pp. 3–11.

17. Burrows M. Wheeler D.J. A block-sorting lossless data compression algorithm. *Digital SRC Research Report* 124. 1994. `ftp://ftp.digital.com/pub/DEC/SRC/research-reports/SRC-124.ps.gz`

18. Lempel A., Ziv J. On the Complexity of Finite Sequences, *IEEE Trans. on Inform. Theory.*, 1976, vol. 22, no. 1, pp. 75–81.

19. Cleary J.G., Witten I.H. Data Compression Using Adaptive Coding and Partial String Matching, *IEEE Trans. on Commun.*, 1984, vol. 32, no. 4, pp. 396–402.

20. Cormack G.V., Horspool R.N. Data Compression Using Dynamic Markov Modelling, *Computer J.*, 1987, vol.30, no. 6, pp. 541–550.