

DOI: 10.17516/1997-1370-0704  
УДК 519.2:801.82(045)

## Anonymous Vs. Attributed: Cluster Analysis of Tolstovskii Sbornik Texts and Its Interpretation in Terms of Cultural Heritage

Oleg F. Zholobov<sup>a</sup>, Victor A. Baranov<sup>b</sup>  
and Maria O. Novak<sup>c,d\*</sup>

<sup>a</sup>*Kazan Federal University  
Kazan, Russian Federation*

<sup>b</sup>*Kalashnikov Izhevsk State Technical University  
Izhevsk, Russian Federation*

<sup>c</sup>*Vinogradov Russian Language Institute of RAS  
Moscow, Russian Federation*

<sup>d</sup>*FRC “Kazan Scientific Center of RAS”  
Kazan, Russian Federation*

Received 01.08.2020, received in revised form 05.11.2020, accepted 09.12.2020

**Abstract.** In the article, the quantitative analysis revealed lexical and semantic dominants and markers that distinguish the medieval anthology texts from each other. To verify whether three anonymous homilies in the thirteenth-century Tolstovskii Sbornik might be attributed to Cyril of Turov, the authors examined the statistical distance between anonymous and already attributed texts. Using the clustering method based on the ranks of the most frequent tokens and the corresponding ranks of other texts, they constructed dendrograms that showed the text grouping. This technique allowed demonstrating the statistical proximity of six Cyril of Turov’s texts, their contrast to seven Cyril of Jerusalem’s texts, and the formation of the third cluster from texts of other authors. Cluster analysis made it possible to identify in Cyril of Turov’s homilies several crucial thematic keys, as well as to establish such a feature of his preaching discourse as the widespread use of role deixis. The analysis confirmed the sharp difference between the anonymous Parable of Wisdom and Cyril of Turov’s homilies. Separate convergences of two anonymous sermons with Cyril of Turov’s homilies were discovered. However, the level of convergence in this case, as analysis has shown, contrasts sharply with the level of convergence among Cyril of Turov’s homilies. It suggests that the causes of individual convergences are not associated with one person’s authorship.

**Keywords:** 13<sup>th</sup>-century Tolstovskii Sbornik, Cyril of Turov, anonymous texts, attribution, cluster analysis, tokens’ frequency ranks, lexical and grammatical convergence.

This study is financially supported by the Russian Science Foundation, the project “Distributional-quantitative analysis of semantic changes based on large diachronic text corpora” (project no. 20-18-00206).

Research area: linguistics.

---

Citation: Zholobov, O.F., Baranov, V.A., Novak, M.O. (2020). Anonymous vs. attributed: cluster analysis of Tolstovskii sbornik texts and its interpretation in terms of cultural heritage. *J. Sib. Fed. Univ. Humanit. Soc. Sci.*, 13(12), 2056–2081. DOI: 10.17516/1997-1370-0704.

---

### The source and tasks

Kazan digital collection on the “Manuscript” website now contains a new online edition of an Old Russian manuscript, Tolstovskii Sbornik from the second half of the 13<sup>th</sup> century (National Russian Library, F.p.I.39, *SbTol* hereafter) (Kazan Collection). Owing to the fragmentation module, all the structural parts of the collection received, in addition to the general, a separate publication; thus for the first time, it became possible to compare them using computer methods and give a linguistic interpretation of this comparison.

*SbTol* has a unique complex composition. The contents of this written source determine its particular value: it includes the earliest copies of Cyril of Turov’s homilies (ff. 1r–23r, 25r–46r), Parable of Wisdom (ff. 48r–49v), a compilation attributed to John Chrysostom (ff. 49v–56v)<sup>1</sup>, two apocryphal texts, the Legend of Aphroditian (ff. 56v–62r), the Abgar Legend (ff. 62v–68v), Life of Basil the Great (ff. 68v–88v)<sup>2</sup>, and a peculiar version of Cyril of Jerusalem’s catechetical lectures (ff. 89v–184r)<sup>3</sup>. There are two anonymous sermons in between Cyril of Turov’s homilies, and also, the Parable of Wisdom adjoins them. These texts could belong to Cyril of Turov, although they do not refer to any authorship, in contrast to the other six homilies. In (*Svodnyĭ katalog* 1984: 324), the anonymous sermons on the 5<sup>th</sup> Sunday after Easter (ff. 23r–25r) and on

Pentecost (ff. 46r–48r) are attributed to Cyril of Turov, indeed. The Parable of Wisdom is associated with Cyril of Turov’s authorship in (*Svodnyĭ katalog*, 2002: 494; *Zalizniak*, 2004: 464; *Slovar’ drevnerusskogo ĭazyka*, VI: 34). In (Zholobov, 2018a, Zholobov, 2018b, Zholobov, Novak, 2018), we found cases of spelling and lexical and grammatical contrast, proving that these three texts did not belong to Cyril of Turov, contrary to the existing assumptions. The “intrigue” of the research below is the use of new IT-technologies in the search for accurate quantitative indicators in establishing authorship and their linguistic interpretation. Of great interest are the statistical parameters that determine the differences in the creative methods of two authors of homiletic texts – Cyril of Turov and Cyril of Jerusalem. The first experience of combining cluster and linguistic analysis gave the results unexpected in many ways.

At the statistical experiment stage, it is supposed to solve the following tasks:

- to find statistical differences in Cyril of Turov’s and Cyril of Jerusalem’s texts,
- to identify the degree of contrast of John Chrysostom’s text and two subcorpora with the established authorship,
- to determine the degree of proximity or remoteness of three anonymous texts to Cyril of Turov’s texts,
- to demonstrate the general grouping of Cyril of Turov’s and Cyril of Jerusalem’s works and other texts of *SbTol*.

### Quantitative and statistical methods in linguistics

Quantitative and statistical methods of data analysis have long been actively and

<sup>1</sup> The compilation nature of the homily was recently discovered in Maria Novak’s research (Novak, 2019).

<sup>2</sup> This version remained unknown to Bulgarian researchers of hagiography. Its incipit is different from those presented in (Ivanova, 2008: 410–418) that indicates a particular translation.

<sup>3</sup> The special version of catechetical lectures’ translation was preliminarily confirmed in (Novak, Penkova 2020).

productively used in various theoretical and applied linguistic studies: text systematization, text attribution, identification of topics and keynotes, and other areas. These methods are based on a statistical analysis of text units (characters, words, syntactic constructions) and/or their distribution. Started in Russia by N.A. Morozov (Morozov, 1915), the analysis based on quantitative and statistical characteristics was used toward the end of the 20<sup>th</sup> century as one of the leading research methods. As more machine-readable texts appeared, it became an integral part of linguistic research (cf., for example: (Marusenko, 1990; From Nestor to Fonvizin, 1994; Mukhin, 2011; Shaikovich, Andriushchenko, Rebetskaia, 2013; Zakharov, Khokhlova, 2014; Mitrofanova, 2015; Borunov, Malygin, 2016; Litvinova, T., Litvinova, O., 2016; Litvinova, T., Zagorovskaia, Seredin, 2016; Mimno, Blei, 2011; Bing, 2012; Blei, 2012; Daud, 2012; Guest, MacQueen, Namey, 2012; Baranov, 2018; Jurafsky, Martin, 2019: 325–415), and many other works devoted to general and peripheral issues of linguistic statistics).

### Establishing authorship

Determining the author of a piece of text is one of the traditional and long-developing areas using quantitative and statistical information. Over more than 100 years of searching for effective methods for solving this problem, a large number of methods and techniques have been proposed and tested using both various formal indicators of texts and various analysis procedures and algorithms (cf., for example, (Morozov, 1915; Marusenko, 1990; Martynenko, 2014; Martynenko, 2015; Gurova, 2016), and many others).

The work (Gurova, 2016) offers a review of various approaches and systematizes various methods used for anonymous text attribution: those based on the analysis of vocabulary, syntactic constructions, sequences of linguistic units, and complex ones (Gurova, 2016: 30). The author concludes that neither the use of lexical or syntactical quantitative characteristics of the texts (for instance, comparing the frequency of service or modal words, syntactic constructions, or sentence length) nor calculat-

ing letter combinations and applying to them, for example, Markov-based methods processes are not universal (Gurova, 2016: 31–34). The author recognizes complex methods as the most effective but notes that even using them, mathematical or linguistic methods “определения авторства могут быть лишь подспорьем для филолога: полностью доверять им нельзя” (“determining authorship cannot be entirely helpful to a philologist: one cannot completely trust them”) (Gurova, 2016: 35).

### Methods

To achieve the goal set in this paper, the cluster analysis method is used. It allows organizing the studied objects according to their characteristics into groups (clusters), where certain objects are more similar than objects from other groups, and visualize it as a dendrogram (Mandel', 1988: 10, 11, 22, 40; Manning, Raghavan, Schütze, 2011: 353–354, 379–381; Vorontsov; Klasternyi Analiz).

The taxonomy method proposed in (Tryon, 1939) is a series of algorithms (procedures) that use various methods for measuring multidimensional distances (numerical expressions of similarities and differences of objects) between objects and various methods of forming clusters (cf., for example, (Cattell, 1944; Sokal, Sneath, 1963; Mandel', 1988; Prikladnaia statistika, 1989; Zagoruiko, 1999; Paul, Gore, 2000; Manning, Raghavan, Schütze, 2011: 383–396).

The Euclidean distance, the squared Euclidean distance, Manhattan distance, the power iteration, the Chebyshev distance, and others are used as modes of measuring distances (metrics). The nearest neighbour rule (single-linkage and complete-linkage clustering), Ward's method, and k-means clustering methods are used for cluster-forming (Mandel', 1988: 30–35, 41–52 onwards; Jain, Murty, Flynn, 1999; Prikladnaia statistika, 1989: 147–181, 249–260; Avtomaticheskaiia obrabotka, 2011: 193–194; Klasternyi Analiz).

In addition to hierarchical methods, there is a group of non-hierarchical ones, which include, in particular, the k-means, where the algorithm constructs a given number of clusters using the smallest distances between objects within groups and the largest distances be-

tween clusters (Mandel', 1988: 40; Prikladnaïa statistika, 1989: 221–222, 291–293; Manning, Raghavan, Schütze, 2011: 363–370; Klasternyi Analiz).

The cluster method has its peculiarities. The researchers indicate that the results of clustering depend, in particular, on the selection and proportions of the source data, on the choice of method for metrics' evaluation, on grouping rules, on the degree of compactness, on the proportionality of the selected measure for each feature of grouping objects (Mandel', 1988: 30, 111–112; Prikladnaïa statistika, 1989: 148, 180–181, 300–301; Bureeva, 2007: 8–11; Vorontsov).

At the same time, they emphasize that the method allows presenting data following the tasks, for which one can select appropriate algorithms for measuring distance and moving from one level of grouping to another one (Mandel', 1988: 73, 146; Zagoruïko, 1999: 60–62; Rabinovich, 2007: 74; Bureeva, 2007: 12–14; Avtomaticheskaiâ obrabotka, 2011: 208–209). For example, N.G. Zagoruïko writes: “Одной, «самой естественной», «абсолютно объективной», таксономии не существует. Все реальные объекты имеют бесконечное число свойств, и выделение некоторого конечного подмножества этих свойств – акт субъективный. Меры близости, критерии качества также выбираются субъективно. Если известна цель, для достижения которой делается таксономия (т. е. при наличии «суперцели»), то качество таксономии проверяется тем, хорошо ли она способствует достижению этой цели, удобна ли, экономична и т.д. Эта проверка носит объективный характер, но выбор суперцели опять-таки субъективен и для одной суперцели данная таксономия будет хорошей, для другой – нет” (“The one and only, ‘the most natural,’ ‘objective,’ taxonomy does not exist. All real objects have an infinite number of properties, and the selection of a finite subset of these properties is a subjective act. Proximity measures, quality criteria are also selected subjectively. If we know for what goal a taxonomy is made (that is, if there is a ‘super goal’), then the quality of the taxonomy is verified whether it contrib-

utes well to this goal, whether it is convenient, economical, and so on. This verifying is objective. However, the choice of a super goal is again subjective, and for one super goal, this taxonomy will be good; for another, it will not”) (Zagoruïko, 1999: 59). Also: “Надо помнить, что выбранная метрика, как и выбранное пространство, является *единственной*, и никакая другая такого же результата не гарантирует. Поэтому очень полезно сделать расчеты несколько раз с разными метриками и найти устойчивые общие черты в разбиениях. Окончательный критерий кластер-анализа – критерий *практической полезности результата*; в случае успеха одновременно считаются удачными и расстояние, и алгоритм” (“One should remember that the selected metric, like the selected space, is the only one, and no other guarantees the same result. Therefore, it is advantageous to make calculations several times with different metrics to find stable common features in the partitions. The final criterion for cluster analysis is the criterion of the practical utility of the result; if successful, both distance and the algorithm are considered successful at the same time”) (Mandel', 1988: 146)<sup>4</sup>.

For methodology, “познание сущности объекта сводится к выявлению тех его качественных свойств, которые и определяют данный объект, отличают его от других. По этой причине задача построения естественных классификаций в известной мере смыкается с традиционной для статистически задачей построения типологических группировок...” (“knowledge of the essence of an object leads to identifying those of its properties that determine this object, distinguishing it from others. For this reason, the task of constructing natural classifications is, to some extent, interfaced with the statistically traditional task of constructing typological groupings...”) (Mandel', 1988: 138). Besides: «Однако объекты

<sup>4</sup> There is also a piece of evidence that «одна и та же пара алгоритма кластеризации и метрики дает различные результаты в зависимости от программы кластеризации» (“the same pair of clustering algorithm and metric gives different results depending on the clustering program”) (Rabinovič, 2007: 74).

могут быть однокачественными в одном отношении и разнокачественными в другом, причем выбор этих отношений (целей, точек зрения) *полностью* находится в руках исследователя» (“However, objects can be one-quality in one respect and different-quality in another, and the choice of these relations (goals, points of view) is entirely in the hands of a researcher”) (Mandel', 1988: 138).

The quoted statements and the logic of the study require choosing features that correspond to our goal. To verify whether anonymous pieces from *SbTol* might be attributed to Cyril of Turov, we should exclude from the analysis the features that unite anonymous and attributed texts, namely graphical and spelling peculiarities, and common text topics. Also, we should rely on the primary expert grouping on Cyril of Turov's and Cyril of Jerusalem's works based on their differentiating features.

### Argumentation

To initialize cluster analysis, we can use the information either about the properties (characteristics) of objects or their pairwise mutual distances, in both cases presented in the form of matrices (Prikladnaia Statistika, 1989: 143; Vorontsov).

To compare texts and register the degree of their closeness using cluster analysis, one can use various quantitative characteristics, such as the number of tokens, the order of their sequence in the lists ranked by quantitative or statistical value, the average length of linguistic units, lists of the most important words or combinations, and discrepancies knots.

There are cases of using cluster methods for the analysis of medieval Slavonic manuscripts. The works of D.M. Mironova (Mironova, 2015; Mironova, 2017) demonstrate the efficiency of automatic clustering of medieval manuscripts of one work (based on several dozen verses of the Gospel of Matthew (Matthew, 14: 14–34) from 525 Slavonic copies of the Gospels from the 11<sup>th</sup> – 16<sup>th</sup> centuries). The author also proposes optimal procedures for highlighting textually significant differences and selecting parameters for comparing manuscripts.

### Source data selection and algorithm search

The properties of cluster methods make it possible to select experimentally such characteristics of the analyzed objects, which provide a result that most closely corresponds to the intuitively (expertly) established grouping of objects.

The most popular are hierarchical clustering algorithms, in particular, due to the visibility of its results on dendrograms and the possibility (moreover, desirability) of their comparison, even if they are obtained by various methods (Mandel', 1988: 73).

There are several approaches to finding an objective grouping of objects: a) data arrays with the unknown structure are analyzed by various methods, with results comparing, b) data are analyzed using algorithms verified on similar data arrays, c) algorithms are verified on artificial arrays, and other ways (Mandel', 1988: 108).

In this paper, we use the first approach, which also involves comparing the analysis results to the existing (expert) grouping into two groups – texts of Cyril of Turov and texts of Cyril of Jerusalem. Searching for the actual results presumes, at the same time, the search for the most appropriate method of analysis and the need for experiments with various data sets.

Algorithms for grouping objects into clusters have various properties. One can assume that texts written by the same author have some similar features represented by numerical values, and can choose a grouping method that puts these texts into a separate group. After finding such a method, it is possible to determine the location of anonymous texts in relation to established clusters.

### Texts to analyze

Our material contains six texts of Cyril of Turov and seven texts of Cyril of Jerusalem.

The sermons of Cyril of Turov are:

On Sunday of St. Thomas the Apostle (without beginning) (hereafter CT\_Thom, ff. 1r–5v), On Descent from the Cross (hereafter CT\_Desc, ff. 5v–16r), On Sunday of the Paralytic (hereafter CT\_Paral, ff. 16r–23r), On Sunday of the Blind Man (hereafter CT\_Blind, ff. 25r–32r), On Ascension of the Lord (hereafter CT\_Asc, ff. 32r–37v), On Nicaea Council Fa-

thers' commemoration (hereafter CT\_Fath, ff. 37v–46r).

The catechetical lectures of Cyril of Jerusalem are: 1st – 3rd (hereafter CJ1, ff. 89v–92v; CJ2, ff. 92v–97v; CJ3, ff. 97v–103r) and 13th – 16th (hereafter CJ13, ff. 158v–170v; CJ14, ff. 170v–176v; CJ15, ff. 176v–181v; CJ16, ff. 181v–184v).

The volumes of the texts of the two authors are approximately equal: 39.899 and 44.524 tokens, respectively.

The volume of three anonymous texts is 5.215 tokens (2.173, 1.572, and 1.470).

The analysis also involves one text partially attributed to John Chrysostom, the Nativity sermon (hereafter Chrys\_Nat, ff. 49v–56v, 5.391 tokens), as well as The Legend of Aphroditian (hereafter Aphr, ff. 56v–62r, 4.649 tokens), The Abgar Legend (hereafter Abg, ff. 62v–68v, 5.016 tokens), and Life of St. Basil the Great (hereafter Life\_Bas, ff. 68v–88 v, 19.421 tokens).

The pieces of Cyril of Turov, Cyril of Jerusalem, and the Nativity sermon, the authors of which are known, as well as four texts with non-established authorship, act as expert texts.

### Machine-readable text transcription

A peculiarity of the text transcription on the “Manuscript” website is its maximum, as far as possible, correspondence to the original: transcriptions transfer the manuscripts letter to letter, line to line, and page to page. Since several scribes might have written a manuscript, the use of accurate transcription, conveying all the features of the original, would lead to analysis based on the graphic characteristics of writing, not the linguistic features of the texts. Therefore, when preparing the data, we level the graphic and orthographic variability as much as possible, using the modern Cyrillic alphabet in tokens' lists and eliminating all the diacritics.

### Tools

The “Manuscript” corpus's statistics module (<http://manuscripts.ru/mns/!cred2.stat>) presents a query form and a sample visualization form. It allows creating comparable subcorpora; entering a wildcard of linguistic units;

entering a quantitative (information about the absolute or relative use of units) or a statistical (statistical measures Log-Likelihood, TF \* ICTF, Weirdness) ranking measure; selecting a contrasting subcorpus; sorting and displaying the lists of results in a table. We can use each subcorpus's numerical data (unit number, its rank, absolute or relative quantity, and the value of the statistical measure) to view the contents of tables and study them using the traditional comparative method, as well as export them to other programs for processing and evaluating numerical data. In this work, we carry out correlation and cluster analysis using the statistical software package Statistica (StatSoft-Dell / TIBCO Software Inc.), which is one of the professional programs.

### Data extraction

For clustering, texts should be represented by some featured spaces whose vectors are sets of numerical values. For such a vector, we selected a set of quantities that described the relation of a text to other texts, namely, the values of pairwise correlations, which could be represented as an  $n*m$  matrix.

The correlations between the texts were calculated using the Spearman's rank method based on the information about the ranks of tokens in a regularized list.

We used various methods to identify how much thematic and semantic features of texts influence the analysis and minimize the peculiarities of writing: the generalization of tokens, their sorting based on quantitative parameters, and taking into account their part-of-speech characteristics.

### Experimental technique

Using the query form of the “Manuscript” corpus, based on the *SbTol* transcription, we prepared 20 samples, including six sermons of Cyril of Turov, seven lectures of Cyril of Jerusalem, three controversial anonymous works, and four other texts.

The prepared samples were loaded into the statistics module, the necessary query parameters established<sup>5</sup>. See Fig. 1:

<sup>5</sup> Unit type: token; step type: sampling; measure: relative quantity; accuracy: 1 (diacritics removed, ligatures are dis-

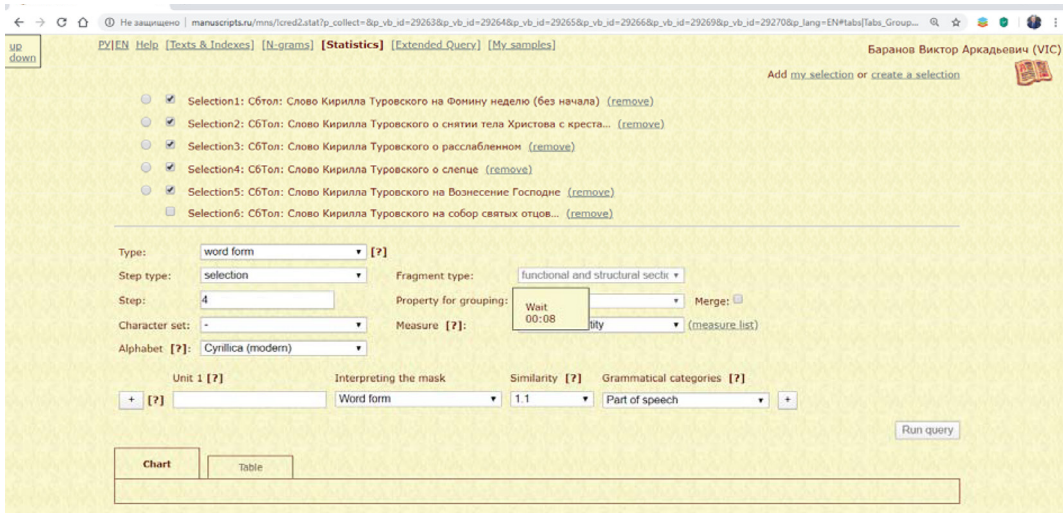


Fig. 1. The query form of the statistics module

Units	№	Rank	Total	СбТол: Слово Кирилла Туровского на Фомину неделю (без начала)				СбТол: Слово Кирилла Туровского о снятии тела Христова с креста...				СбТол: Слово Кирилла Туровского о расслабленном				СбТол: Слово Кирилла Туровского о слепце				СбТол: Слово Кирилла Туровского о Вознесении Господне	
				№	R	F	Freq	№	R	F	Freq	№	R	F	Freq	№	R	F	Freq		
и	1	1	753	0.06718	1	1	95	0.07600	1	1	184	0.06703	1	1	125	0.06692	1	1	107	0.06203	1
ль	2	2	222	0.01981	2	2	20	0.01600	2	2	42	0.01530	3	3	39	0.02088	3	3	32	0.01855	3
от	5	5	148	0.01320	3	2	20	0.01600	4	3	41	0.01494	9	9	16	0.00857	6	5	26	0.01507	7
не	4	4	162	0.01445	4	3	15	0.01200	9	8	28	0.01020	2	2	50	0.02677	2	2	35	0.02029	10
сь	8	8	123	0.01097	5	3	15	0.01200	3	3	41	0.01494	20	14	10	0.00535	7	6	21	0.01217	4
бо	7	7	131	0.01169	6	4	13	0.01040	6	5	38	0.01384	5	5	25	0.01338	10	8	16	0.00928	6
на	3	3	193	0.01722	7	4	13	0.01040	5	4	40	0.01457	8	8	17	0.00910	5	4	30	0.01739	2
наше	23	22	37	0.00330	9	5	12	0.00960	39	20	8	0.00291	31	17	7	0.00375	32	15	6	0.00348	53
кю	12	12	67	0.00598	8	5	12	0.00960	36	20	8	0.00291	14	12	12	0.00642	15	12	10	0.00580	28
да	13	13	64	0.00571	10	6	9	0.00720	17	13	16	0.00583	10	9	16	0.00857	24	15	6	0.00348	13
огоже	29	27	32	0.00285	14	6	9	0.00720	24	17	11	0.00401	57	20	4	0.00214	73	18	3	0.00174	378
же	6	6	139	0.01240	11	6	9	0.00720	7	6	32	0.01166	4	4	27	0.01445	11	9	15	0.00870	5
77	76	33	0.00204						50	23	5	0.00182	15	12	12	0.00642	44	16	5	0.00260	476

Fig. 2. The result web form: a table of tokens regularized by frequency criterion

The result of the query is a table that includes all tokens of samples, and the information about each of them: their absolute and relative number, their index number, and their rank in every text (see Fig. 2):

In the web form, there is a possibility to resort the table columns. It allows choosing each of the texts as the main one and sorting its tokens in descending order of their quantity,

and then establishing the correspondence of the sequence of forms for each pair of texts (see Tables 1 and 2).

The length of the lists is 100 forms for tokens, the results saved in the files of the Statistica program.

In the Statistica program, using Spearman's rank correlation method, based on text

closed); the alphabet: modern Cyrillic.

Table 1. Correlation of quantitative characteristics of the most frequent tokens in Cyril of Turov's sermon "On Sunday of St. Thomas the Apostle", with the corresponding values of other texts (a fragment)

Tokens	Cyril of Turov's sermon "On Sunday of St. Thomas the Apostle" (without a beginning)					Cyril of Turov's sermon "On Sunday of the Paralytic"					Cyril of Turov's sermon "On Sunday of the Blind Man"									
	№ <sup>2</sup>	R <sup>3</sup>	F <sup>4</sup>	Freq <sup>5</sup>		№	R	F	Freq		№	R	F	Freq						
Sample volume	4293 <sup>1</sup>					9925					6605					6009				
и	1	1	95	0,076	1	1	186	0,068	1	1	125	0,067	1	1	107	0,062				
въ	2	2	20	0,016	2	2	42	0,015	3	3	39	0,021	3	3	32	0,019				
о(т)	3	2	20	0,016	4	3	41	0,015	9	9	16	0,009	6	5	26	0,015				
не	4	3	15	0,012	9	8	28	0,010	2	2	49	0,026	2	2	35	0,020				
съ	5	3	15	0,012	3	3	41	0,015	20	14	10	0,005	7	6	21	0,012				
бо	6	4	13	0,010	6	5	37	0,013	5	5	25	0,013	10	8	16	0,009				
на	7	4	13	0,010	5	4	40	0,015	8	8	17	0,009	4	4	30	0,017				
нѣня	9	5	12	0,010	39	20	8	0,003	30	17	7	0,004	31	15	6	0,003				
яко	8	5	12	0,010	36	20	8	0,003	14	12	12	0,006	15	12	10	0,006				
да	10	6	9	0,007	17	14	16	0,006	10	9	16	0,009	25	15	6	0,003				
егоже	14	6	9	0,007	23	17	11	0,004	64	20	4	0,002	72	18	3	0,002				
же	11	6	9	0,007	7	6	32	0,012	4	4	27	0,014	11	9	15	0,009				
мя	13	6	9	0,007	58	23	5	0,002	15	12	12	0,006	37	16	5	0,003				
о	12	6	9	0,007	8	7	29	0,011	17	13	11	0,006	8	7	17	0,010				
азъ	15	7	7	0,006	265	26	2	0,001	74	20	4	0,002	11386	21	0	0,000				
вся	16	8	6	0,005	18	15	14	0,005	82	21	3	0,002	280	20	1	0,001				
есмь	17	8	6	0,005	298	26	2	0,001	25	17	7	0,004	8834	21	0	0,000				
къ	19	8	6	0,005	14	12	19	0,007	12	11	13	0,007	21	14	7	0,004				
ми	20	8	6	0,005	24	18	10	0,004	16	12	12	0,006	562	20	1	0,001				
ребра	18	8	6	0,005	85	24	4	0,001	3761	24	0	0,000	3729	21	0	0,000				

<sup>1</sup> The number of tokens used in the text.

<sup>2</sup> A token's ordinal number.

<sup>3</sup> A token's rank.

<sup>4</sup> A token's absolute quantity.

<sup>5</sup> A token's relative quantity.



Table 2. Correlation of quantitative characteristics of the most frequent words of Cyril of Turov's sermon "On Descent from the Cross" with the corresponding values of other texts (a fragment)

Tokens	Cyril of Turov's sermon "On Descent from the Cross"				Cyril of Turov's sermon "On Sunday of St. Thomas the Apostle" (without a beginning)				Cyril of Turov's sermon "On Sunday of the Paralytic"				Cyril of Turov's sermon "On Sunday of the Blind Man"			
	№	R	F	Freq	№	R	F	Freq	№	R	F	Freq	№	R	F	Freq
Sample volume	4293				9925				6605				6009			
и	1	1	186	0,068	1	1	95	0,076	1	1	12,5	0,067	1	1	107	0,062
въ	2	2	42	0,015	2	2	20	0,016	3	3	39	0,021	3	3	32	0,019
о(т)	4	3	41	0,015	3	2	20	0,016	9	9	16	0,009	6	5	26	0,015
съ	3	3	41	0,015	5	3	15	0,012	20	14	10	0,005	7	6	21	0,012
на	5	4	40	0,015	7	4	13	0,010	8	8	17	0,009	4	4	30	0,017
бо	6	5	37	0,013	6	4	13	0,010	5	5	25	0,013	10	8	16	0,009
же	7	6	32	0,012	11	6	9	0,007	4	4	27	0,014	11	9	15	0,009
о	8	7	29	0,011	12	6	9	0,007	17	13	11	0,006	8	7	17	0,010
не	9	8	28	0,010	4	3	15	0,012	2	2	49	0,026	2	2	35	0,020
его	10	9	22	0,008	26	9	5	0,004	18	13	11	0,006	9	8	16	0,009
еси	11	10	21	0,008	118	12	2	0,002	22	15	9	0,005	19	14	7	0,004
нъ	13	11	20	0,007	24	9	5	0,004	6	6	22	0,012	5	4	30	0,017
тя	12	11	20	0,007	1870	14	0	0,000	50	19	5	0,003	27	15	6	0,003
б(с)ы	15	12	19	0,007	11383	14	0	0,000	35	18	6	0,003	125	19	2	0,001
къ	14	12	19	0,007	19	8	6	0,005	12	11	13	0,007	21	14	7	0,004
тело	16	13	17	0,006	2008	14	0	0,000	2235	24	0	0,000	2213	21	0	0,000
да	17	14	16	0,006	10	6	9	0,007	10	9	16	0,009	25	15	6	0,003
вся	18	15	14	0,005	16	8	6	0,005	82	21	3	0,002	280	20	1	0,001
ли	19	16	12	0,004	360	13	1	0,001	28	17	7	0,004	22	14	7	0,004
ти	20	16	12	0,004	2040	14	0	0,000	29	17	7	0,004	90	18	3	0,002

forms<sup>6</sup>, we established the correlation distances for each text concerning other texts (see Table 3), saving the results in temporary files.

Then, we collected the data in an  $n \times m$  matrix (see Table 4), where each text gained a description with a set of correlation values relative to all other texts.

In the Statistica program, we obtained a series of dendrograms (see Fig. 3).

texts: the distribution of Cyril of Turov's and Cyril of Jerusalem's texts into two clusters gave an integration of the Ward combination method and the 1-r Pearson proximity measure (see Fig. 4).

1.2. The addition of three anonymous texts to 13 texts demonstrates the inclusion of the former into the subcluster of Cyril of Jerusalem's works, but not Cyril of Turov's (see

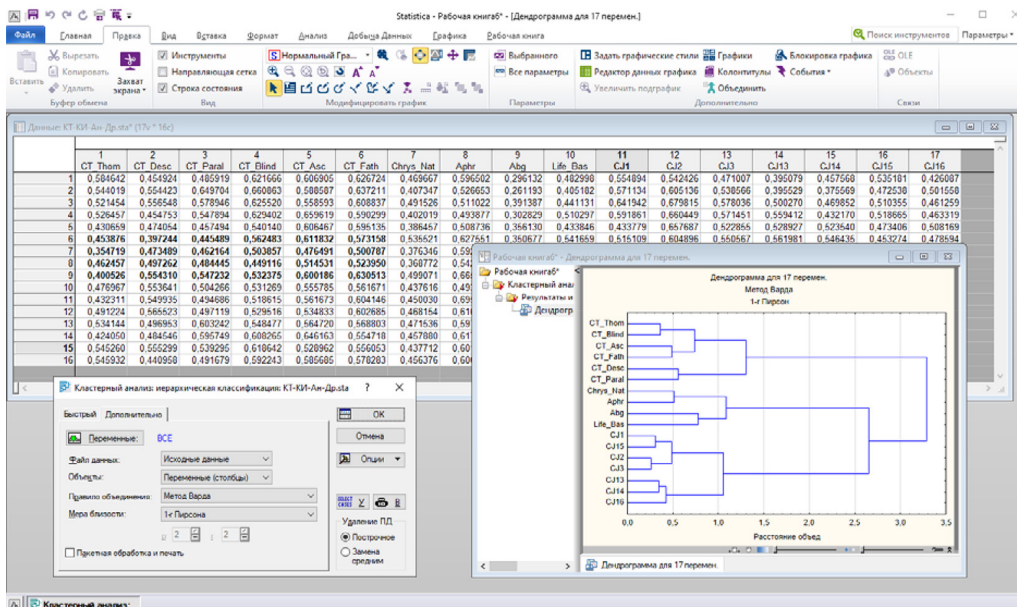


Fig. 3. Building dendrograms in the Statistica program

Experimentally, for the texts of Cyril of Turov and Cyril of Jerusalem, we selected such combinations of proximity measures and association rules that gave two distinct clusters.

**The experiment**

1.1. The experiment used lists of the 100 most frequent tokens. We found a correlation between the texts, as described above, and summarized the data in an  $n \times m$  matrix. The matrix data underwent the cluster analysis.

Experimentally, we selected a combination of metric and method of unification, which most closely matched the expert grouping of

Fig. 5). At the same time, two texts, the anonymous sermon on the 5<sup>th</sup> Sunday after Easter and the anonymous sermon on Pentecost, form a separate subcluster, and anonymous Parable of Wisdom form a subcluster with the 3<sup>rd</sup> lecture of Cyril of Jerusalem.

1.3. For comparison, the texts of Cyril of Turov and Cyril of Jerusalem were analyzed with the addition of four texts of unknown authors (Fig. 6).

All four texts formed a special subcluster close to the subcluster of Cyril of Jerusalem's texts.

1.4. The construction of the dendrogram using all texts gave the result shown in Fig. 7.

The controversial texts: the anonymous sermon on the 5<sup>th</sup> Sunday after Easter and the

<sup>6</sup> When analyzing lemmas from a list of 200 forms, only function words and adverbs were selected.

Table 3. The values of Spearman's rank correlation of Cyril of Turov's sermons "On Sunday of St. Thomas the Apostle" and "On Descent from the Cross" (the first 100 most frequent text forms)

On Sunday of St. Thomas the Apostle	Tokens number	Spearman's R	t(N-2)	p-level	On Descent from the Cross	Tokens number	Spearman's R	t(N-2)	p-level
CT_Desc	100	0,584642	7,133888	0,000000	CT_Thom	100	0,454924	5,057124	0,000002
CT_Paral	100	0,544019	6,418412	0,000000	CT_Paral	100	0,554423	6,594905	0,000000
CT_Blind	100	0,521454	6,049761	0,000000	CT_Blind	100	0,556548	6,631483	0,000000
CT_Asc	100	0,526457	6,129900	0,000000	CT_Asc	100	0,454753	5,054725	0,000002
CT_Fath	100	0,430659	4,723807	0,000008	CT_Fath	100	0,474054	5,329840	0,000001
An_East	100	0,429214	4,704371	0,000008	An_East	100	0,426611	4,669470	0,000010
An_Pent	100	0,405248	4,388232	0,000029	An_Pent	100	0,414578	4,509945	0,000018
An_Wisd	100	0,394699	4,252584	0,000048	An_Wisd	100	0,433699	4,764839	0,000007
Chrys_Nat	100	0,453876	5,042444	0,000002	Chrys_Nat	100	0,397244	4,285122	0,000043
Aphr	100	0,354719	3,755762	0,000293	Aphr	100	0,473489	5,321640	0,000001
Abg	100	0,462457	5,163404	0,000001	Abg	100	0,497262	5,673857	0,000000
Life_Bas	100	0,400526	4,327258	0,000036	Life_Bas	100	0,554310	6,592956	0,000000
CJ1	100	0,476967	5,372196	0,000001	CJ1	100	0,553641	6,581489	0,000000
CJ2	100	0,432311	4,746080	0,000007	CJ2	100	0,549935	6,518229	0,000000
CJ3	100	0,491224	5,582869	0,000000	CJ3	100	0,565523	6,788134	0,000000
CJ13	100	0,534144	6,254790	0,000000	CJ13	100	0,496953	5,669171	0,000000
CJ14	100	0,424050	4,635272	0,000011	CJ14	100	0,484546	5,483481	0,000000
CJ15	100	0,545260	6,439228	0,000000	CJ15	100	0,555299	6,609963	0,000000
CJ16	100	0,545932	6,450544	0,000000	CJ16	100	0,440958	4,863652	0,000004

Table 4. The matrix of correlation values (vectors) of Cyril of Turov's and Cyril of Jerusalem's texts

no.	CT_Thom	CT_Desc	CT_Paral	CT_Blind	CT_Asc	CT_Fath	CJI	CJ2	CJ3	CJ13	CJ14	CJ15	CJ16
1	0,584642	0,454924	0,485919	0,621666	0,606905	0,626724	0,554894	0,542426	0,471007	0,395079	0,457568	0,535181	0,426087
2	0,544019	0,554423	0,649704	0,660863	0,588587	0,637211	0,571134	0,605136	0,538566	0,395529	0,375569	0,472538	0,501558
3	0,521454	0,556548	0,578946	0,625520	0,558593	0,608837	0,641942	0,679815	0,578036	0,500270	0,469852	0,510355	0,461259
4	0,526457	0,454753	0,547894	0,629402	0,659619	0,590299	0,591861	0,660449	0,571451	0,559412	0,432170	0,518665	0,463319
5	0,430659	0,474054	0,457494	0,540140	0,606467	0,595135	0,433779	0,657687	0,522855	0,528927	0,523540	0,473406	0,508169
6	0,476967	0,553641	0,504266	0,531269	0,555785	0,561671	0,515109	0,604896	0,550567	0,561981	0,546435	0,453274	0,478594
7	0,432311	0,549935	0,494686	0,518615	0,561673	0,604146	0,605303	0,645678	0,641308	0,408796	0,516876	0,577996	0,485932
8	0,491224	0,565523	0,497119	0,529516	0,534833	0,602685	0,676646	0,691917	0,594379	0,569276	0,532308	0,585933	0,478010
9	0,534144	0,496953	0,603242	0,548477	0,564720	0,568803	0,697112	0,696091	0,684136	0,589209	0,559987	0,576867	0,523542
10	0,424050	0,484546	0,595749	0,608265	0,646163	0,554718	0,655101	0,657863	0,625271	0,686842	0,555666	0,664347	0,526790
11	0,545260	0,555299	0,539295	0,618642	0,528962	0,556053	0,640105	0,623886	0,561580	0,620035	0,681245	0,575210	0,530149
12	0,545932	0,440958	0,491679	0,592243	0,585685	0,578283	0,623243	0,682011	0,627238	0,602598	0,484626	0,581881	0,494365

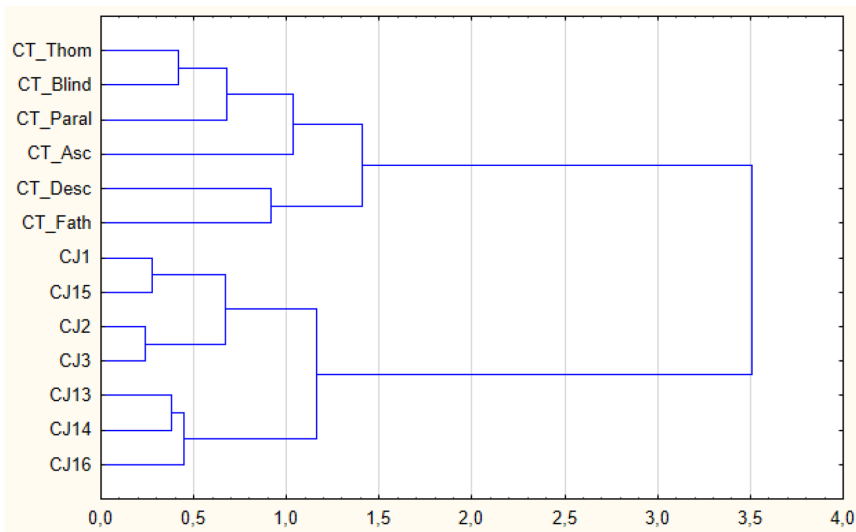


Fig. 4. The dendrogram of Cyril of Turov's and Cyril of Jerusalem's texts (Ward's method, Pearson's 1-r correlation)

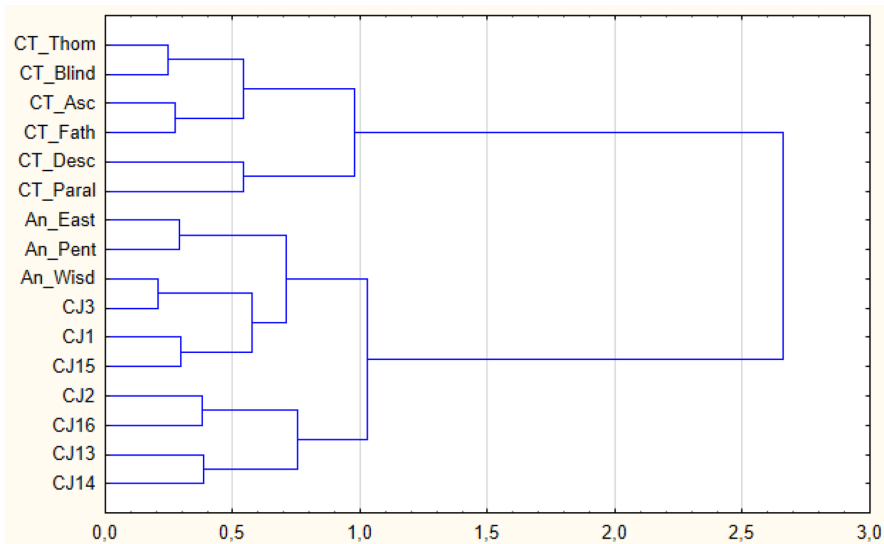


Fig. 5. The dendrogram of the texts of Cyril of Turov, Cyril of Jerusalem, and three controversial texts (Ward's method, 1-r Pearson proximity measure)

anonymous sermon on Pentecost became a subcluster in the Cyril of Turov's works cluster, close to the subcluster of Cyril of Turov's sermon on Descent from the Cross and Cyril of Turov's sermon on Sunday of the Paralytic. The anonymous Parable of Wisdom was included in the subcluster of works of various authors, close to the subcluster of Cyril of Jerusalem's works.

### The linguistic interpretation of the experiment

The experiment results given above may, at first glance, seem paradoxical. Using statistical tools (a combination of association rules and proximity measures) we managed to construct a quantitative picture of the distribution of homiletic texts that are homogeneous in the

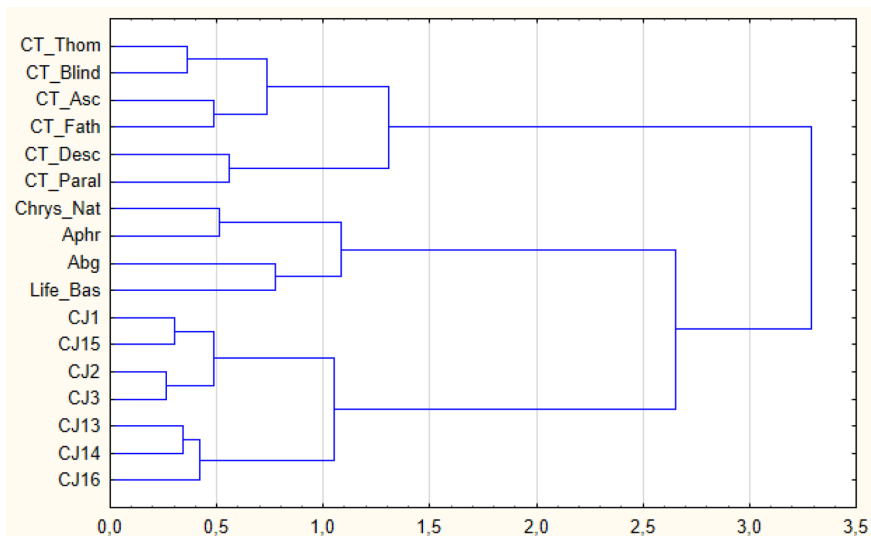


Fig. 6. The dendrogram of the texts of Cyril of Turov, Cyril of Jerusalem and four texts of other authors (Ward's method, 1-r Pearson proximity measure)

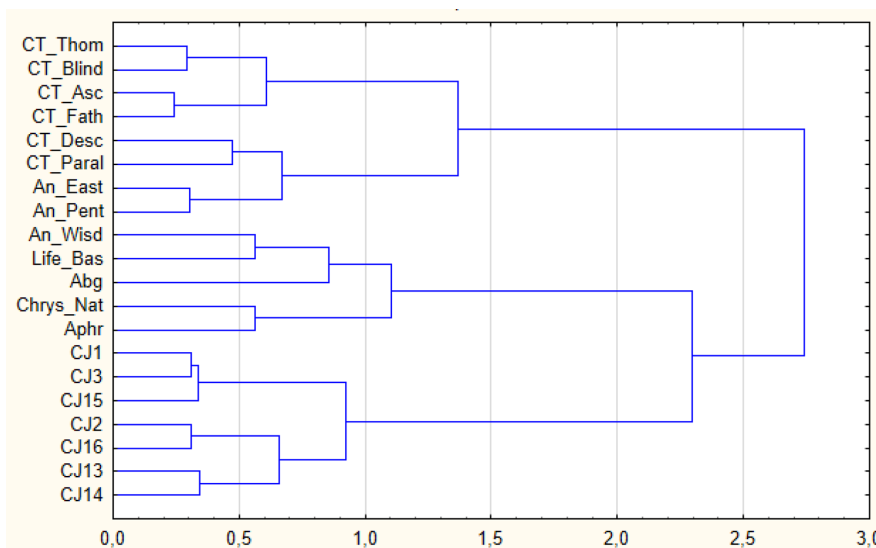


Fig. 7. The dendrogram of the texts of Cyril of Turov, Cyril of Jerusalem, three controversial texts, and four texts of unknown authors (Ward's method, 1-r Pearson proximity measure)

genre-and-stylistic sense (see Fig. 4, 5), and the texts turned out to be distinctly separated there, i.e., grouped in different clusters and subclusters.

A distinct statistical contrast between Cyril of Turov's original homilies and Cyril of Jerusalem's translated lectures made it possible to reevaluate both the degree of originality of

Cyril of Turov's sermons and the V.V. Kolesov's idea (which seemed to be an exaggeration) about russification of Cyril of Turov's language<sup>7</sup>.

<sup>7</sup> “Художественное открытие Кирилла и заключается в самом раннем в истории русского литературного языка и весьма последовательном сближении двух языковых стихий – церковнославянской и русской, в

Although different clustering conditions give similar results, there is still some dissimilarity. Consideration of specific lexical and grammatical forms, whose quantitative characteristics were analyzed, allows us to understand both lexical and grammatical reasons for the inclusion of controversial texts in certain subclusters and differences in the classification results. Comparison of the linguistic parameters, which base the classification in Fig. 7, allows us to interpret in what respects and how accurately the statistical data correspond to the linguistic picture of the convergence and divergence in the texts. In the comparison below, we consider the full-meaning units: nouns, adjectives, verbs, and pronouns. This comparison indicates not just lexical convergence or divergence, but a specific morphological and syntactic realization, i.e., a special kind of thematic proximity or remoteness of texts since we carry out statistical calculations based not on lemmas but word forms (tokens). The coincidence of tokens, and not lexemes, of course, should emphasize the particular proximity between the texts.

“On Descent from the Cross” and “On Sunday of the Paralytic” homilies form a subcluster; the frequent nouns accurately reflect the key themes of both sermons. At the same time, there are not many exact matches of tokens in them, and each list, being unique, characterizes Cyril of Turov’s preaching style. Cf., with ranks indicated (Table 5).

In these lists of frequent tokens, along with expected coincidences of the *Богъ* word forms (see below), we also register non-trivial ones. For instance, the form *словомъ* (Instr. Sg.) turned out to be quite frequent, as well as various declension forms of the word *земля*. *Словомъ* is a new form instead of the original \*s-stem *словесьмь*. In Cyril of Turov’s preaching strategy, the form *словомъ* is necessary when confessing in both homilies the life-giving, miraculous power of the divine Logos’s ut-

чрезвычайно тонком понимании их специфики и пределов использования в художественной речи” (“Cyril’s artistic discovery is the persistent convergence of two language elements, Church Slavonic and Russian, with an extremely subtle understanding of their specificity and limits of use in the artistic speech, which phenomenon was the earliest in the history of the Russian literary language”) (Kolesov, 1981: 38).

terances, when a word becomes a deed (слово юго дѣломъ бѣы СТ\_Paral, 16.2). Cf.:

- 8 и мѣртвы-  
 9 я словомъ въскрѣсивъша твоюго  
 10 бжѣтва мановениюмъ СТ\_Desc, 8.1;
- 14 Како ли  
 15 въ моемъ хоудѣмъ положю та гро-  
 16 бѣ · нбѣнии кроугъ оутвердивъша-  
 17 го словомъ · ѡ на хѣровимѣхъ съ w̄-  
 18 цмъ ѡ съ стымъ почивающаго дхмъ СТ\_  
 Desc, 11.1;
- 15 Блжнѣ юси и ѡсифе · иже вса ѡживи-  
 16 въшаго словомъ · ѡ водами покры-  
 17 въшаго твердѣ нбѣною · сего іако мь-  
 18 ртвѣца каменемъ покрылѣ юси  
 19 въ гробѣ СТ\_Desc, 14.2;
- 11 Егоже нѣнѣ хѣ бл-  
 12 гый члвколюбецъ словомъ ищѣли ·  
 13 врачъ бо юсть дшамъ нашимъ и тѣ-  
 14 ломъ · и слово юго дѣломъ бѣы СТ\_Paral,  
 16.2;
- 3 Лазора  
 4 оуже раскысѣвъша въ гробѣ · и че-  
 5 тыри дни ймоуща въ мѣртвыхъ ·  
 6 словомъ жива створихъ · и тобѣ  
 7 нынѣ глѡ въстани ѡ възми ѡдрѣ  
 8 свой · ѡ йди въ домъ свой СТ\_Paral, 20.2;
- 8 не насытите ли са въ · л̄ · ѡ ·  
 9 ѡ · л̄ · зраще мене на ѡдрѣ йсполу-  
 10 мѣртва лежаща · нынѣ же въставѣ-  
 11 шю ми бжнѣмъ словомъ · ѡслѣпосте  
 12 оумомъ · ѡ ѡ своюй храмлюще прѣ-  
 13 тыкаетеца неправдѣ СТ\_Paral, 21.1.

It is noticeable that syntagmas with the form *словомъ* are not repetitive, and in the last context, this form occasionally expands a Dativus absolutus. Only once an instrumental form is used as a comitative one since it expands the lexeme’s meaning, referring to the Tablets of the Law:

- 18 двѣ бо ѡ силома киво-  
 19 тъ съ бнѣмъ словомъ принесе нѣ  
 20 въ своюмъ оубоіаца поставити ю-

Table 5

CT_Desc	CT_Paral
тело 13	члвка 14
иосифе 20	одрь 17
ба (= бога) 21	г(с)ь 18
гробе 21	купель 18
бъ (= богъ) 22	одра 18
кр(с)те 22	бе (= боже) 18
руце 23	ба (= бога) 20
страха 23	болезни 20
х(с)а (= христа) 23	бу (= богу) 20
х(с)е (= христе) 23	бъ (= богъ) 20
адъ 24	вода 20
гробъ 24	купели 20
животъ 24	недуга 20
земля 24	одре 20
и(с)съ (= иисус) 24	члвкъ 20
кр(с)та 24	англь 21
миръ 24	блг(д)ти 21
мъртвьца 24	бмь (= богомъ) 21
ребра 24	воду 21
словомъ 24	г(с)ди (= господи) 21
смрти 24	горе 21
сна (= сына) 24	земля 21
телесе 24	крщения 21
адама 25	народа 21
англа 25	недугъ 21
	недугы 21
	слово 21
	словомъ 21

21 го домоу · ты же не скинию̄ съ зако-  
22 номъ · нъ самого ба̄ приємъ ѿ крь-  
23 ста CT\_Desc, 14.1.

The form *земля* (Nom.Sg. and the homonymous bookish form Gen. Pl., different from its East-Slavic correlate *земль*) is predictably frequent in both homilies. If lemmatized, this frequency would grow even higher, since in “On Descent from the Cross” sermon, there are also three Acc. Sg. forms *землю*, two Instr. Sg. forms *землею*, and one Loc. Sg. form *земли*; in “On Sunday of the Paralytic” sermon there are two Acc. Sg. forms and one Loc. Sg. form. The different meanings of the word *земля* in both Cyril of Turov’s homilies convey a universal nature of the events of sacred history. The paired formula *небо и земля* ‘heaven and earth’ used in homilies is inseparable from the mythological and folklore poetic tradition. Cf.:

18 како та̄ тьрнїемъ вѣнчаша  
19 и зълчи съ ѿцтомъ напоиша · ѿ ѿ-  
20 ще и прч̄таи ти ребра копиємъ про-  
21 бодоша · оужасноуа нбо̄ и земля  
22 трепещеть · ию̄дѣйска не тьрпаше  
23 дерзновениа̄ · слнце помьрче · ѿ ка-  
24 мение̄ распадеса · жидовскою̄ ѿка-  
25 менїение̄ іавляюще CT\_Desc, 6.2;

26 спостражи бл̄говѣрно̄ · соугу-  
1 баго ти ради вѣнца · ѿгоже по въскрь-  
2 сении х̄вѣ въсприимеши · ѿ всѣхъ  
3 конецъ земля чьстьноюю̄ славою и  
4 поклонение̄ · и на нбси бесконьчъ-  
5 ною̄ жизнь CT\_Desc, 8.1–2;

24 тобѣ всю тварь на рабо-  
25 тоу створиохъ · нбо̄ ѿ земля тобѣ слу-  
26 жита · ѿно влагою̄ · а си плодомъ :  
1 Тебе ради слнце свѣтомъ ѿ теплото-



2 ю слоужить · и лоуна съ звѣздами  
 3 ношь вѣблаѣть · тебе дѣла вѣблаци  
 4 дѣждьмь землю напаіають · и зе-  
 5 мла всакоу траву сѣменитоу · и  
 6 дрѣва плодовитаіа · на твою слоужь-  
 7 боу възращають CT\_Paral, 19.2–20.1.

The word *земля* in the sense ‘terra firma as a part of the universe’ appears in the following context:

3 оу-  
 4 вы мнѣ і<sup>с</sup>се мнѣ драгоу іма · како  
 5 стоитъ земля чююши тл на себе на  
 6 крѣтѣ висаща · иже на водахъ тоу въ  
 7 начатѣцѣ вѣсновавъ іеси CT\_Desc, 8.1.

The word *земля* in the general sense of human habitat is a natural continuation of the previous usage<sup>8</sup>:

21 съ бо наша  
 22 болѣзни понесе · и за ны пострада ·  
 23 раноу іго вси йцѣлѣхомь · зане  
 24 прѣдана б<sup>с</sup>ы на смрѣ дша іго · и  
 25 съ незаконники вѣмѣненъ б<sup>с</sup>ы ·  
 26 истрѣбимъ бо рѣша память іго  
 1 вѣ земля живуущихъ · и іма іго  
 2 не поманетьса к тому CT\_Desc, 10.1–2;

5 къ крщ<sup>с</sup>нию бо аще и всеіа земля при-  
 6 доуть члвци · не оумалитьса б<sup>и</sup>а  
 7 блг<sup>т</sup>ь CT\_Paral, 17.2.

Figure 7 presents a classification where the anonymous sermons on the 5<sup>th</sup> Sunday after Easter and on Pentecost enter the subcluster of Cyril of Turov’s homilies and join the subcluster, which includes “On Descent from the Cross” and “On Sunday of the Paralytic” sermons. See the contents and ranks of the matching noun forms in four texts: “On Descent from the Cross,” “On Sunday of the Paralytic,” the anonymous “On the 5<sup>th</sup> Sunday after Easter,” and the anonymous “On Pentecost” (Table 6).

The results of the comparison seem unexpected, demonstrating an extremely narrow lexical and syntactic base of convergence. Even

lesser convergence we observe among adjective forms. Cf. (Table 7).

Verb forms (including participles) demonstrate a similar situation. Rare convergence here is almost entirely limited to individual forms of the existential verb *быти*, which acts as a copula and is associated with temporal and role deixis. Cf. (Table 8).

In this short list, the exception is the aorist form *пече* as an essential element in the dramatic discourse of homilies since it presents someone’s speech.

The distribution of pronoun forms presents a completely different picture. Cf. (Table 9).

The frequent convergence of the tokens here is significant: in Cyril of Turov’s homilies, there are 17 convergences (!), between CT\_Desc and An\_Pent – 6, between CT\_Desc and An\_Pent – 4, between CT\_Paral and An\_East – 7, between CT\_Paral and An\_Pent – 8, between two anonymous sermons – 5.

Thus, consideration of convergence among pronoun forms gives unexpected results: the main basis for the closeness between the texts is formed precisely by pronoun forms, not thematic but mainly deictic ones.

K. Bühler, one of the founders of semiotics, attributed personal pronouns *me* and *you* to the “deictic field of language”, to the role deixis, linking these lexemes with the indication a sender and an addressee of the “signal exchange” [Bühler 2000: 74 fol.]. He developed the deictic coordinate system of a subject with the well-known “here-now-me” scheme, and also made a distinction between the dramatic and epic types of deixis [Ibid: 94 fol., 125–127].

In the encyclopedic description of deixis in Russian, the role-playing deixis, in addition to the personal pronouns of the 1<sup>st</sup> and 2<sup>nd</sup> persons, designating the addresser and the addressee, also includes possessive pronouns and an indication the speech object (third person of a pronoun) is assigned to a separate variety of deixis [Russkii iazyk: entsiklopediia 2020: 139–140]. Indicative pronouns and prepositions are characterized as deictic per se since their lexical meaning is identical to deictic. Anaphora is associated with the syntagmatic type of deixis.

<sup>8</sup> Cf. the entry *земля* in (Slovar’ drevnerusskogo jazyka III: 371–376).

Table 6

CT_Desc	CT_Paral	An_East	An_Pent
-	бе (= боже) 19	-	бе 9
ба (= бога) 21	ба 20	-	-
-	бу (= богу) 20	бу 11	бу 10
бъ (= богъ) 22	бъ 20	-	бъ 9
земля 24	земля 21	-	-
словомъ 24	словомъ 21	-	-
-	-	братъе 13	братъе 10
of 25 tokens	of 28 tokens	of 19 tokens	of 25 tokens

Table 7

CT_Desc	CT_Paral	An_East	An_Pent
бию (= божию) 23	-	-	бию 8
бия (= божия) 24	бия 20	бия 13	бия 9
of 4 tokens	only one rank form!	of 12 tokens	of 11 tokens

Table 8

CT_Desc	CT_Paral	An_East	An_Pent
еси 10	еси 15	-	-
б(с)ы (= бысть) 12	б(с)ы 18	-	-
рече 18	рече 21	-	р(ч)е (= рече) 8
есть 19	есть 12	есть 11	есть 8
-	есмь 17	есмь 12	-
-	быхъ 19	быхъ 12	-
-	-	-	-
-	-	будеть 11	будеть 9
-	-	блг(с)ви 13	блг(с)ви 9
-	-	будемъ 13	будемъ 10
of 9 tokens	of 16 tokens	of 23 tokens	of 21 tokens

In the above-mentioned statistic data for four texts, the first place belongs to the role deixis, namely to personal pronouns of the 1<sup>st</sup> and 2<sup>nd</sup> persons (both singular and plural), as well as related possessive pronouns. The convergence between CT\_Desc and CT\_Paral in this part contains 10 units (!); between CT\_Desc and An\_East – only one unit; between CT\_Desc and An\_Pent – 2; between CT\_Paral and An\_East – 3; between CT\_Paral and An\_Pent – 3; An\_East and An\_Pent – 3 units.

To a far lesser degree, the convergence in the homilies of Cyril of Turov is due to the deictic, i.e., demonstrative pronouns.

We observe approximately the same situation concerning other texts. The convergence

between CT\_Desc and CT\_Paral is 3 units; between CT\_Desc and An\_East – 3; between CT\_Desc and An\_Pent – 0; between CT\_Paral and An\_East – 1; between CT\_Paral and An\_Pent – 0; between An\_East and An\_Pent – 1. The anaphoric deixis expressed in rank tokens *его, егоже, иже* is at the same level: between CT\_Desc and CT\_Paral are 3 units; between CT\_Desc and An\_East – 1; between CT\_Desc and An\_Pent – 1; between CT\_Paral and An\_East – 1; between CT\_Paral and An\_Pent – 1 unit.

The forms of the generalizing quantifier *вьсь* are also statistically significant, not so much for Cyril of Turov's homilies, but relations with anonymous texts. Cf.: between

Table 9

CT_Desc	CT_Paral	An_East	An_Pent
его 9	его 13	его 11	-
-	кто 18	кто 10	-
-	всемъ 19	-	всемъ 10
тя 11	тя 19	-	-
вся 15	вся 21	вся 12	вся 8
ти 16	ти 17	-	ти 9
-	вси 18	-	вси 10
егоже 17	егоже 20	-	егоже 9
ми 18	ми 12	ми 12	-
сего 18	-	сего 11	-
всехъ 19	всехъ 19	-	-
-	азъ 20	азъ 11	азъ 8
-	васъ 21	васъ 11	васъ 10
-	самъ 21	-	самъ 9
мне 20	мне 17	-	-
мои 20	мои 18	-	-
насъ 22	-	-	насъ 9
се 22	се 17	се 12	-
иже 23	иже 19	-	-
мя 23	мя 12	-	-
си 23	си 18	-	-
тебе 23	тебе 16	-	-
тому 23	тому 20	-	-
моего 24	моего 17	-	-
намъ 24	намъ 21	-	-
того 24	-	того 12	-
-	-	то 8	то 7
-	-	вамъ 11	вамъ 6
of 30 tokens	of 31 tokens	of 17 tokens	of 21 tokens

CT\_Desc and CT\_Paral – 2 units; between CT\_Desc and An\_East – 1; between CT\_Desc and An\_Pent – 1; between CT\_Paral and An\_East – 1; between CT\_Paral and An\_Pent – 3 units. In this regard, the sermon “On Sunday of the Paralytic” turned out to be closest to the anonymous homily on Pentecost.

The large statistical weight of pronouns in texts is universal. In modern dictionaries, pronouns occupy only 0.1% of lexis, and in texts, do much more, about 11%. In this regard, they are close to function words, and together, they occupy the “statistical” top “of the frequency dictionary” (Russkii iazyk: entsiklopediia, 2020: 342).

At the same time, the noted significant differences in the frequency of some pronouns and the level of convergence reveal the essential features of the discursive and pragmatic

nature of texts. An extremely high level of role deixis characterizes Cyril of Turov’s homilies, which indicates the great weight of the dramatic elements in his preaching strategy, or the peculiar place of the “dramatic method” in his texts (using K. Bühler’s terms).

The remaining texts of the collection, according to the dendrogram in Figure 7, form a separate cluster, which grows closer to the cluster of Cyril of Jerusalem homilies and consists of the following three subclusters: An\_Wisd, Life\_Bas, and Abg; Chrys\_Nat and Aphr. Convergences among pronouns also take a leading place in these cases; however, the importance of role deixis in these convergences presents a different picture. Its quantitative status is sharply reduced, though in such texts as the Life of Basil and the Tale of Aphroditian, it is represented by a consider-

able amount, 7 and 5 units, respectively. Cf. (Table 10).

The presented pronouns convergence supposes to contradict the union of subclusters shown on the dendrogram in Figure 7. The distribution of pronouns in “The Life of Basil the Great” is closer to that in the apocryphal “The Abgar Legend” than in the anonymous “Parable of Wisdom”: 9 and 7 units, respectively. This contradiction is even more tangible when juxtaposing Life\_Bas with the anonymous “Legend of Aphroditian”: there are 12 matches, although the texts belong to different subclusters.

We should also note that between Life\_Bas and Aphr, there are more similarities concerning nouns and verbs, while there are no similarities among adjectives. Cf. (Table 11).

This contradiction certainly can be smoothed out by the greater proximity of the rank values of the matching tokens. Thus, it reveals the specifics of the quantitative measurement of the texts’ convergence and divergence based on the rank status of corresponding to-

kens. Another explanation for this contradiction could be the greater proximity of Life\_Bas and An\_Wisd, considering function words: prepositions, conjunctions, and particles. Their comparison, however, gives the opposite result again. Cf. (Table 12).

Here again, the similarities between Life\_Bas and Abg prevail: the convergence between An\_Wisd and Life\_Bas is 11 units, and between Life\_Bas and Abg – 15 units.

**Conclusion**

The linguistic analysis of the convergence and divergence in Tolstovskii Sbornik texts as a whole confirmed the effectiveness of the statistical methods applied. The statistical analysis made it possible to identify several thematic keys crucial for Cyril of Turov’s homilies, as well as to establish the significance of the role deixis and the diverse use of other types of pronouns for his preaching discourse. The great importance of role deixis in Cyril of Turov’s homilies creates the basis for evaluation of the

Table 10

An_Wisd	Life_Bas	Abg	Chrys_Nat	Aphr
то 4	то 23	то 16	-	-
того 5	-	того 16	-	-
тя 5	-	тя 17	-	-
еже 6	еже 10	-	еже 19	еже 14
всемъ 7	всемъ 18	всемъ 29	-	-
ми 7	ми 21	-	-	-
ся 7	ся 18	-	ся 21	ся 11
ти 7	ти 20	-	-	-
всехъ 8	всехъ 22	-	-	всехъ 17
-	се 20	се 11	се 21	се 14
-	ему 8	ему 14	ему 22	ему 12
-	ту 23	ту 16	-	-
-	его 19	его 18	его 19	его 15
-	иже 9	иже 18	иже 17	иже 16
-	мя 28	мя 18	-	-
-	ны 29	ны 18	-	ны 16
-	насъ 29	-	насъ 21	насъ 17
-	-	вси 16	вси 22	вси 16
-	намъ 29	-	намъ 22	намъ 11
-	-	весь 18	-	весь 18
-	-	нашего 18	нашего 20	-
-	что 24	-	-	что 16
ты 6	-	-	-	ты 17
-	тебе 27	-	-	тебе 17
of 21 tokens	of 28 tokens	of 20 tokens	of 12 tokens	of 24 tokens

Table 11

An_Wisd	Life_Bas	Abg
бъ (= богъ) 8	бъ 25	-
г(с)ь (= господь) 8	г(с)ь 29	-
-	ба (= бога) 12	ба 17
-	леть 16	леть 18
-	бе (= боже) 22	бе 13
-	градъ 25	градъ 9
-	града 26	града 11
-	еп(с)ль 26	еп(с)ль 17
-	множество 29	множество 18
of 36 tokens	of 20 tokens	of 34 tokens
есть 3	есть 17	есть 14
рече 7	рече 17	рече 18
-	б(с)ы (= бысть) 22	б(с)ы (= бысть) 22
-	быти 28	быти 18
of 19 tokens	of 16 tokens	of 14 tokens

Table 12

An_Wisd	Life_Bas	Abg
же 2	же 2	же 3
аще 3	аще 24	-
а 4	а 11	а 12
не 4	не 5	не 13
въ 5	въ 3	въ 2
да 5	да 10	да 18
в 6	в 6	в 10
о(т) 6	о(т) 6	о(т) 8
с 6	с 18	-
о 7	о 7	о 12
бо 8	бо 10	бо 17
-	на 4	на 4
-	яко 18	яко 6
-	къ 18	къ 13
-	к 23	к 15
-	за 26	за 17
-	по 13	по 17
of 11 tokens	of 23 tokens	of 19 tokens

dramaturgical mode involvement in other texts. Besides, it is an essential indicator of the original preaching discourse.

The analysis demonstrated that the variety and quantitative level of pronouns use could determine the proximity and distance of texts from each other to a much greater extent than the use of nouns, adjectives, and verbs. In quantitative and statistical terms, the level of convergence may not completely coincide with lexical and syntactic parallels, since these cas-

es emphasize the rank status of existing convergences. At the same time, we should take into account that the special diagnostic significance of pronouns is mostly explained by their universal high frequency in texts of various types; it primarily testifies to the nature of the discursive and pragmatic organization of texts, not to their thematic affinity. In the future, for an accurate evaluation of the thematic texts' proximity, it is necessary to use the lemmatization mechanism, which, however, should not

be used in all cases. When lemmatizing verb forms, there is a risk of losing valuable information related to deictic categories of person, time, and taxis.

Thus, the linguistic and statistical analysis confirmed the sharp difference between the anonymous Parable of Wisdom and Cyril of Turov's homilies. It is another proof that the Parable belongs to another, unknown author. At the same time, statistical data allowed us to detect peculiar similarities between anonymous sermons on the 5<sup>th</sup> Sunday after Easter and on Pentecost with Cyril of Turov's homilies. Nevertheless, the level of convergence in this case sharply contrasts with the level of convergence among Cyril of Turov's homilies themselves, and this proves that the reasons for the convergence are not connected with one person's authorship. Indeed, they are limited by some similarities only in the use of units with universally high frequency, namely, pronouns of different lexical and semantic categories. The dynamics of function words' distribution will require a separate study in the future since such words have the highest frequency in texts of any type.

#### Sources and abbreviations

Abg – The Abgar Legend (“The Holy Mandyllion Transference”)

An\_East – (an anonymous) sermon on the 5<sup>th</sup> Sunday after Easter

An\_Wisd – (an anonymous) Parable of Wisdom

An\_Pent – (an anonymous) sermon on Pentecost

Aphr – Legend of Aphroditian

Life\_Bas – Life of St. Basil the Great

Chrys\_Nat – John Chrysostom's Nativity sermon

CJ1 – the 1st catechetical lecture of Cyril of Jerusalem

CJ2 – the 2nd catechetical lecture of Cyril of Jerusalem

CJ3 – the 3rd catechetical lecture of Cyril of Jerusalem

CJ13 – the 13th catechetical lecture of Cyril of Jerusalem

CJ14 – the 14th catechetical lecture of Cyril of Jerusalem

CJ15 – the 15<sup>th</sup> catechetical lecture of Cyril of Jerusalem

CJ16 – the 16th catechetical lecture of Cyril of Jerusalem

CT\_Asc – Cyril of Turov's sermon “On Ascension of the Lord”

CT\_Paral – Cyril of Turov's sermon “On Sunday of the Paralytic”

CT\_Fath – Cyril of Turov's sermon “On Nicaea Council Fathers”

CT\_Blind – Cyril of Turov's sermon “On Sunday of the Blind Man”

CT\_Desc – Cyril of Turov's sermon “On Descent from the Cross”

CT\_Thom – Cyril of Turov's sermon “On Sunday of St. Thomas the Apostle”

Kazan Collection – Kazan collection of Slavic-Russian written sources from the 12<sup>th</sup> – 14<sup>th</sup> centuries. Kazan Federal University, the laboratory of palaeoslavistics, with the support of IAS “Manuscript”, 2007–2020, available at: <http://manuscripts.ru/mns/portal.main?p1=54> (accessed 15 May 2020).

SbTol – Sermons and teachings collection (“Tolstovskii Sbornik”), the 2<sup>nd</sup> half of the 13<sup>th</sup> century (Russian National Library, F.p.I. 39), 184 ff. [Online resource] / Oleg Zholobov et al.; “Manuscript” project, available at: [http://manuscripts.ru/mns/main?p\\_text=96362255](http://manuscripts.ru/mns/main?p_text=96362255) (accessed 15 May 2020).

#### References

Aivazian, S.A. (ed.) (1989). Prikladnaia statistika: klassifikatsiia i snizhenie razmernosti [Applied Statistics: classification and dimension reduction]. In *Finansy i statistika* [Finance and Statistics], Moscow, 608 p.

Baranov, V.A. (2018). Statistical Analysis of the Slavonic Paraenesis by Ephrem the Syrian (on Three Electronic Copies of the 13–14<sup>th</sup> Centuries from the Manuscript Corpus). In *Journal of Siberian Federal University. Humanities & Social Sciences*, 11(8), 1211-1228. DOI: 10.17516/1997-1370-0302.

Bing, L. (2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers, 167 p.

- Biuler, Karl (2000). *Teoriia iazyka: reprezentativnaia funktsiia isyuka* [Language Theory: representative function of language]. Moscow, Progress, 528 p.
- Blei, D. (2012). Probabilistic Topic Models. In *Communications of the ACM*, 55(4), 77–84.
- Bol'shakova, E.I., Klyshinskii, E.S., Landè, D.V., Noskov, A.A., Peskova, O.V., Ilagunova, E.V. (2011). *Avtomaticheskaiia obrabotka tekstov na estestvennom iazyke i komp'iuternaia lingvistika* [Automatic processing of natural language texts and computer linguistics]. Moscow, 272 p.
- Borunov, A.B., Malygin, V.T. (2016). Statisticheskie metody i priemy lingvisticheskogo izucheniia iazyka pisatel'ia (na materiale khudozhestvennoi angloiazychnoi prozy) [Statistical methods and techniques of the linguistic study of the writer's language (based on the material of English fiction prose)]. In *Mir lingvistiki i kommunikatsii* [The world of linguistics and communication], 1(46), 4, 56–65, available at: <https://www.elibrary.ru/item.asp?id=28141197> (accessed 15 May 2020).
- Bureeva, N.N. (2007). *Mnogomernyi statisticheskii analiz s ispol'zovaniem PPP «STATISTICA»* [Multidimensional statistical analysis with STATISTICA program use]. Nizhniy Novgorod, 112 p.
- Cattell, Raymond B. (1944). A Note on Correlation Clusters and Cluster Search Methods. In *Psychometrica*, 9, 169–184.
- Daud, A. (2012). Using Time Topic Modelling for Semantics-Based Dynamic Research Interest Finding. In *Knowledge-Based*, 26, 154–163.
- Guest, G., MacQueen, K., Namey, E. (2012). *Applied Thematic Analysis*. Thousand Oaks, California, Sage, 295 p.
- Gurova, E.I. (2016). Metodiki atributsii avtorstva v sovremennoi otechestvennoi filologii [Authorship attribution techniques in modern Russian philology]. In *Novyi filologicheskii vestnik* [New philological journal], 38 (2), 29–44, available at: <https://www.elibrary.ru/item.asp?id=27250173> (accessed 15 May 2020).
- Ivanova, K. (2008). *Biblioteka Hagiographica Balcano-Slavica*. Sofija, 720 p.
- Jain, A., Murty, M., Flynn, P. (1999). Data Clustering: A Review. In *ACM Computing Surveys*, 31 (3), 264–323.
- Jurafsky, D., Martin, J. (2019). *Speech and Natural Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 613 p., available at: <https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf> (accessed 15 May 2020).
- Klasternyi analiz [Cluster analysis]. In *StatSoft: Elektronnyi uchebnik po statistike* [Statsoft: Electronic textbook on statistics], available at: <http://statsoft.ru/home/textbook/modules/stcluan.html> (accessed 15 May 2020).
- Kolesov, V.V. (1981). *K kharakteristike poeticheskogo stilia Kirilla Turovskogo* [On characteristics of Kirill Turovsky's poetic style]. In *Trudy otdela drevnerusskoi literatury* [Proceedings of the Department of Old Russian Literature], 36, 37–49.
- Kniazevskaia, O.A. (ed.). (2002). *Svodnyi katalog slaviano-russkikh rukopisnykh knig, khраниashchiesia v Rossii, stranakh SNG i Baltii: XIV vek, 1 (Apokalipsis–Letopis' Lavrent'ievskaiia)* [A consolidated catalog of Slavic-Russian manuscripts stored in Russia, the CIS and Baltic countries: 14<sup>th</sup> century, 1 (Apokalipsis–Letopis' Lavrent'ievskaiia)]. Moscow, Indrik, 768 p.
- Krys'ko, V.B. (ed.). (1988–). *Slovar' drevnerusskogo iazyka (XI–XIV vv.)* [Dictionary of the Old Russian Language (11<sup>th</sup>–14<sup>th</sup> cc.)], 10–, Moscow, Azbukovnik.
- Litvinova, T.A., Litvinova, O.A. (2016). Diagnostirovanie pola avtora pis'mennogo teksta na russkom iazyke: korpusno-statisticheskii podchod [Diagnosing author's gender in a Russian written text: corpus and statistical approach]. In *Iazyk. Pravo. Obshchestvo* [Language. Law. Society], Penza, Penza University, 151–154.
- Litvinova, T.A., Zagorovskaia, O.V., Seredin, P.V. (2016). Diagnostirovanie pola avtora pis'mennogo teksta na osnove kolichestvennykh parametrov: kognitivnyi podkhod [Diagnosing author's gender in a Russian written text: cognitive approach]. In *Voprosy kognitivnoi lingvistiki* [Cognitive Linguistics Issues], 49(4), 51–59.
- Mandel', Igor' D. (1988). Klasternyi analiz [Cluster analysis]. In *Finansy i statistika* [Finance and Statistics], Moscow, 176 p.

Manning, Ch.D., Raghavan, P., Schütze, H. (2011). *Vvedenie v informatsionnyi poisk* [Introduction to information search]. Moscow, Williams, 528 p.

Martynenko, G.Ia. (2014). Stilemetriia: vznikovenie i stanovlenie v kontekste mezhdistitsiplinarnogo vzaimodeistviia, 1, Pervye shagi: XIX vek [Style metrics: the emergence and formation in the context of interdisciplinary interaction, 1, First steps: 19<sup>th</sup> century]. In *Strukturnaia i prikladnaia lingvistika* [Structural and applied linguistics], 10, 3–23.

Martynenko, G.Ia. (2015). Stilemetriia: vznikovenie i stanovlenie v kontekste mezhdistitsiplinarnogo vzaimodeistviia, 2, Pervaia polovina XX veka: rasshirenie mezhdistitsiplinarnykh kontaktov stilemetrii [Style metrics: the emergence and formation in the context of interdisciplinary interaction. Part 2. The first half of the 20<sup>th</sup> century: the expansion of interdisciplinary contacts of style metrics]. In *Strukturnaia i prikladnaia lingvistika* [Structural and applied linguistics], 11, 9–28.

Marusenko, M.A. (1990). *Atributsiia anonimnykh i psevdonimnykh literaturnykh proizvedenii metodami raspoznavaniia obrazov* [Attribution of anonymous and pseudonymous literary works by pattern recognition methods]. Leningrad, Leningrad University, 164 p.

Milov, L.V. (ed.). (1994). *Ot Nestora do Fonvizina: novye metody opredeleniia avtorstva* [From Nestor to Fonvizin: new methods for determining authorship]. Moskva, Progress, 443 p.

Mimno, D., Blei, D. (2011). Bayesian Checking for Topic Models. In *Empirical Methods in Natural Language Processing*, 227–237.

Mironova, D.M. (2015). Primenenie klasternogo analiza v tekstologii [The use of cluster analysis in textology]. In *Strukturnaia i prikladnaia lingvistika* [Structural and applied linguistics], 11, 155–160.

Mironova, D.M. (2017). *Avtomatizirovannaia klassifikatsiia drevnykh rukopisei (na materiale 525 spisok slavianskogo Evangelii ot Matfeia XI–XVI vv.)* [Automated classification of ancient manuscripts (based on 525 Slavic copies of the Gospel of Matthew from 11<sup>th</sup> – 16<sup>th</sup> centuries)]. St. Peterburg, 315 p.

Mitrofanova, O.A. (2015). Tematicheskoe modelirovanie korpusa «Narodnykh russkikh skazok A. N. Afanas'ieva» [Thematic modelling of the “Folk Russian Tales by A. N. Afanasiev”]. In *Strukturnaia i prikladnaia lingvistika* [Structural and applied linguistics], 11, 146–154.

Moldovan, A.M. (ed.). (2020). *Russkii iazyk: entsiklopediia* [Russian language: Encyclopaedia]. Moscow, AST Press.

Morozov, N.A. (1915). Lingvisticheskie spektry: Sredstvo dlia otlichii plagiatov ot istinnykh proizvedenii togo ili drugogo izvestnogo avtora: Stilemetricheskii etiud [Linguistic spectra: A tool for distinguishing plagiarism from the true works of one or another well-known author: stylistic study]. In *Izvestiia Otdeleniia russkogo iazyka i slovesnosti Imp. Akademii nauk* [Bulletin of the department of the Russian language and literature of the Imperial Academy of Sciences], 20(4), 93–134.

Mukhin, M.J. (2011). *Leksicheskaia statistika i idioshtil' avtora: korpusnoe ideograficheskoe issledovanie (na materiale proizvedenii M. Bulgakova, V. Nabokova, A. Platonova i M. Sholokhova)* [Lexical statistics and an author's idiom: corpus ideographic research (based on the works of M. Bulgakov, V. Nabokov, A. Platonov, and M. Sholokhov)]. Yekaterinburg, 43 p.

Novak, M.O. (2019). Slovo na Rozhdestvo Khristovo v Tolstovskom Sbornike XIII v.: lingvotekstologicheskaiia kharakteristika [A Nativity Sermon in the 13<sup>th</sup>-century Tolstovskiy Sbornik: Textology and Language Features]. In *Vestnik Volgogradskogo gosudarstvennogo universiteta. Seriia 2, Iazykoznanie* [Science Journal of VolSU. Linguistics], 18 (4), 6–17. DOI: <https://doi.org/10.15688/jvolsu2.2019.4.1>

Novak, M.O., Penkova, Ia.A. (2020). Oglasitel'nye poucheniia Kirilla Ierusalimskogo v Tolstovskom Sbornike XIII v. [Cyril Of Jerusalem Catechetical Lectures in Tolstovskii Sbornik from the 13<sup>th</sup> Century]. In *Drevniia Rus'. Voprosy medievistiki* [Old Russia. The Questions of Middle Ages], 81(3), 108–118.

Paul, A., Gore, Jr. (2000). Cluster Analysis. In *Handbook of Applied Multivariate Statistics and Mathematical Modelling*. Academic Press, 2000. 297–321.

Rabinovich, B.I. (2007). Klasternyi analiz detalizatsii telefonnykh peregovorov [Cluster analysis of details of telephone conversations]. In *Sistemy i sredstva informatiki* [Systems and means of informatics], 17, 52–78, available at: <https://clck.ru/M3w75> (accessed 15 May 2020).



Sokal, R.R., Sneath, P.H.A. (1963). *Principles of Numerical Taxonomy*. San Francisco and London, W. H. Freeman and Co., XVI + 359 p.

Shaikevich, A.Ia., Andriushchenko, V.M., Rebetskaia, N.A. (2013). Distributivno-statisticheskii analiz iazyka russkoi prozy 1850–1870-x gg. [Distributive and statistical analysis of the Russian prose language from 1850-1870s]. In *Iazyki slavianskoi kul'tury* [Slavic culture languages], 1. Moscow, 504 p.

Shmidt, S.O. (ed.). (1984). *Svodnyi katalog slaviano-russkikh rukopisnykh knig, khраниashchichsia v SSSR. XI–XIII vv.* [A consolidated catalogue of Slavic-Russian manuscript books stored in the USSR. 11<sup>th</sup>–13<sup>th</sup> centuries]. Moscow, Nauka, 406 p.

Tryon, R. (1939). *Cluster Analysis: Correlation Profile and Orthometric (Factor) Analysis for the Isolation of Unities in Mind and Personality*. Ann Arbor, Mich., Edwards Brothers, inc., available at: <https://catalog.hathitrust.org/Record/001306510> (accessed 15 May 2020).

Vorontsov, K.V. Klasterizatsiia [Clustering], In: *Mashinnoe obuchenie* [Machine learning], available at: <https://clck.ru/FmWDj> (accessed 15 May 2020).

Zakharov, V.P., Khokhlova, M.V. (2014). Avtomaticheskoe vyiiavlenie terminologicheskikh slovosochetanii [Automatic identification of terminological phrases]. In *Strukturnaia i prikladnaia lingvistika* [Structural and applied linguistics], 10, 182–200.

Zagoruiko, N.G. (1999). *Prikladnye metody analiza dannykh i znanii* [Applied data and knowledge analysis methods]. Novosibirsk, available at: [https://www.rfbr.ru/rffi/ru/books/o\\_36814#1](https://www.rfbr.ru/rffi/ru/books/o_36814#1) (accessed 15 May 2020).

Zalizniak, A.A. (2004). Drevnenovgorodskii dialekt [Old Novgorod dialect]. In *Iazyki slavianskoi kul'tury* [Slavic culture languages], Moscow, 872 p.

Zholobov, O.F. (2018). Slovo-pritcha o premudrosti v spiskakh XII–XVI vv. [A parable about wisdom in the manuscripts from the 12<sup>th</sup>–16<sup>th</sup> centuries]. In *Nauchnoe nasledie V.A. Bogoroditskogo i sovremennyi vektor issledovaniia Kazanskoi lingvisticheskoi shkoly* [The scientific heritage of V.A. Bogoroditskii and the modern research vector of the Kazan Linguistic School], 1. Kazan', 85–90.

Zholobov, O.F. (2018). On Contrasting Orthographic Systems in the Manuscript of the 13<sup>th</sup> Century (to the Internet Edition of the Tolstoy Sbornik). In *Drevnīaia Rus'. Voprosy medievistiki* [Old Russia. The Questions of Middle Ages], 73(3), 77–89.

Zholobov, O.F., Novak, M.O. (2018). Verb Forms Functioning in Cyril Turovskii's Homilies (In Comparison to the Tale of Igor's Campaign). In *Zeitschrift für Slawistik*, 63(1), 74–89.

## Анонимность vs. атрибутированность: кластерный анализ текстов Толстовского сборника и его интерпретация в аспекте культурного наследия

О.Ф. Жолобов<sup>а</sup>, В.А. Баранов<sup>б</sup>,

М.О. Новак<sup>в,г</sup>

<sup>а</sup>Казанский федеральный университет

Российская Федерация, Казань

<sup>б</sup>Ижевский государственный технический университет

имени М. Т. Калашникова

Российская Федерация, Ижевск

<sup>в</sup>Институт русского языка им. В. В. Виноградова РАН

Российская Федерация, Москва

<sup>г</sup>ФИЦ “Казанский научный центр РАН”

Российская Федерация, Казань

---

**Аннотация.** В статье с помощью количественного анализа выявлены лексико-семантические доминанты и маркеры, отличающие тексты средневекового сборника друг от друга. С целью выяснения принадлежности трех анонимных произведений Толстовского сборника XIII века Кириллу Туровскому исследуется статистическое расстояние между анонимными и авторскими текстами. С помощью метода кластеризации, данными для которого являются ранги наиболее частотных текстовых форм и соответствующие им ранги других текстов, построены дендрограммы, показывающие группировку исследуемых произведений. Последовательно демонстрируется статистическая близость шести текстов подкорпуса Кирилла Туровского и их контрастность семи текстам Кирилла Иерусалимского, формирование третьего кластера из текстов других авторов. Кластерный анализ позволил выявить в гомилиях Кирилла Туровского несколько наиболее важных для автора тематических ключей, а также установить такую особенность его проповеднического дискурса, как широкое использование ролевого дейксиса. Анализ подтвердил резкое отличие анонимного Слова о премудрости от гомилий Кирилла Туровского. Были обнаружены отдельные схождения двух анонимных поучений с гомилиями Кирилла Туровского. Однако уровень схождений в этом случае, как показал анализ, резко контрастирует с уровнем схождений в гомилиях самого Кирилла Туровского. Это свидетельствует о том, что причины отдельных схождений не связаны с единым авторством.

**Ключевые слова:** Толстовский сборник XIII века, Кирилл Туровский, анонимные тексты, атрибуция, кластерный анализ, ранговая корреляция, лексическая и грамматическая конвергенция.

Работа выполнена при финансовой поддержке Российского научного фонда (РНФ) в рамках проекта «Дистрибутивно-количественный анализ семантических изменений на основе больших диахронических корпусов» (проект № 20–18–00206).

Научная специальность: 10.02.00 – языкознание.