Journal of Siberian Federal University. Humanities & Social Sciences 8 (2018 11) 1211-1228
~ ~ ~

УДК 801.8:004

Statistical Analysis of the Slavonic *Paraenesis* by Ephrem the Syrian (on Three Electronic Copies of the 13–14th Centuries from the *Manuscript* Corpus)

Victor A. Baranov*
Kalashnikov Izhevsk State Technical University
7 Studencheskaia Str., Izhevsk, 426069, Russia

Received 18.04.2018, received in revised form 09.07.2018, accepted 23.07.2018

The work presents an experience of applying statistical methods to discovering thematically valuable words in three Old Russian (Old East Slavic) copies of the Ephrem the Syrian's Paraenesis.

The quantitative data were obtained with the help of the search forms in the historical corpus Manuscript (manuscripts.ru), namely the multitext and N-Gram modules. The basic corpus for analysis of the 28 most frequent lemmas of content words from the Paraenesis (the collection volume exceeds 100 thousand word forms) comprised five corpus collections of different genres: copies of the Menaion for May, Service Menaions for other months of the year, Sticherarion (Book of stichera), Acts and Epistles of the Apostles, and Gospels (the total amount of word forms is more than 1 million).

To evaluate the lemmas obtained with the help of the system automatic morphological analyser the statistic TF-ICTF' (version of the weighting scheme TF-IDF) and Log-Likelihood were used. The increase of the number of analysed lemmas from 10 to 28 allowed demonstration of the great statistical weight of lemmas that are used less often than the most frequent lemmas.

To eliminate the discrepancies in the statistical evaluation of lemmas there were made a comparison of the lemmas' ranks and corresponding diagrams. The analysis of the diagrams made it possible to find the core and periphery of the lists; identify the lemmas with the greatest averaged statistical weight – ПОМЫСЛЪ and СТРАХЪ and also the lemmas in the nearest periphery – ВЕКЪ, ЖИТИ, ХОТЕТИ, БРАТИЯ that represent the orientation of the Paraenesis texts to the spiritual search and motifs of punishment and apocalypse.

The conclusion concerns efficiency and effectiveness of the statistical methods as regards the evaluation of linguistic data in the historical corpora that due to the objective causes considerably cede to the modern corpora by their volume. Moreover, the materials for analysis shall involve the data from the entire corpus of the Slavonic texts (Manuscript) and the entire list of word forms (lemmas) from the analysed manuscript (subcorpus).

* Corresponding author E-mail address: victor.a.baranov@gmail.com

Keywords: corpus linguistics, linguistic statistics, medieval Slavonic manuscripts, Ephrem the Syrian's Paraenesis, key lemmas.

The article was written with the support of the Russian Foundation for Basic Research (RFBR) within the framework of the project "Linguistic statistical analysis of one-component and multi-component lexical units in the historical Manuscript corpus" (grant No. 18-012-00463).

Research area: computational linguistics.

Citation: Baranov, V.A. (2018). Statistical analysis of the slavonic Paraenesis by Ephrem the Syrian (on three electronic copies of the 13–14th centuries from the Manuscript corpus). J. Sib. Fed. Univ. Humanit. soc. sci., 11(8), 1211-1228. DOI: 10.17516/1997-1370-0302.

Introduction and statement of the problem

The use of modern text corpora for quantitative and statistical analysis of linguistic data began at the moment of their appearance, so as for today the relevant methods are thoroughly methodologically, methodically and instrumentally developed and allow solving a large number of applied and research problems.

The creation of Slavonic historical collections and corpora containing marked machine-readable copies of medieval written monuments (see the list of the main ones at the end of the article) now makes it possible to set unconventional tasks for historical linguistics, including the tasks related to the search for combinatorial and statistical features of certain linguistic units functioning in texts. The first experiments produced encouraging results.

Thus, in my work (Baranov, 2017a) I described the experience of quantitative, statistical and distributive analysis of the forms with suffixes *-ьск-* and *-ьн-* in Christianopolis (12th century, Lviv Historical Museum, OP, No. 37, 291 p.) and Tolstoy's (14th century, National Library of Russia, Q.p.I.5, 93 p.) Lists of the Slavonic Apostle (Acts and Epistles of the Apostles)[1]; the potential of several *Manuscript* modules to be used in search for linguistic regularities in the use of long (pro-) and short forms.

Other works (Baranov, 2017b; Baranov, 2018) show the effectiveness and prospects of using statistical methods to identify significant function and content words in small collections of the historical corpus (in several collections of different genres, namely collections of Apostle's lists, Service Menaions for May, Menaions for other months of the year, sticherarions and Gospels). In particular, it was proved

[1] The transcripts were prepared with the help of M.O. Novak (Kazan (Volga region) Federal University), then they were marked out and put into the database of the *Manuscript* corpus, now being available at: http://manuscripts.ru/mns/portal.main?p1=68.

that the set and range of the most frequent function and content words (word forms and lemmas) are different and specific for each collection and can be considered an essential characteristic as regards texts of a certain genre. One of the main research results was the application of statistical methods to identify topically and semantically meaningful words.

It is well known that statistical methods of evaluating the significance of linguistic units in corpora may generate varying results. To eliminate discrepancies in the research I used the methodology of comparing the values of statistical measures for analysed words, so I was able to identify the core and periphery of the most frequently used words and show their individuality for each of the collections. For instance, in the Apostle, the core words were as such – *законъ*, *божии*, in the Menaion for May *песнь*, *славити*, *радовати* were found to be the core words, in the Menaions for other months – *Хрьстосъ*, *дьньсь*, *съпасти*, *милость*, in the sticherarion and in the Gospels – *ити*, *рещи*, *Исусъ*.

Such representative results of the statistical study of the medieval manuscripts collections from the *Manuscript* corpus are inasmuch productive as the analysis shall include the texts from other collections with the subsequent singling out their lexical and thematic core.

Material and method

Amount of linguistic material

The work is based on the subcorpus material comprising three lists of the Ephrem the Syrian[1] *Paraenesis* (the collection of teachings) – Pogodin's (National Library of Russia, Pog., 71a, circa 1269–1289, 328 p.), Typographic (Russian State Archive of Ancient Acts, Typ., 38, 1270s–1280s, 143 p.), Troitsky (Russian State Library, Tr., 7, the middle of the 14th century, 245 p.).[2]

There were analysed most frequently used lemmas of content words, the list of which was obtained using the N-Gram module. A list of the ten most frequent lemmas of the *Paraenesis* subcorpus is given in Appendix 1.

[1] The start page of the subcorpus on the *Manuscript* site, available at: http://manuscripts.ru/mns/portal.main?p1=62; multitext query form, available at: http://manuscripts.ru/mns/srch.simple?p_ed_id=94394362; query form for n-grams, available at: http://manuscripts.ru/mns/cred_ngr.stat?p_vb_id=23280&p_collect=94394362&p_lang=RU

[2] The transcriptions were prepared with the help of O.F. Zholobov (Kazan (Volga region) Federal University), then they were marked out and put into the database of the *Manuscript* corpus, now being available at: http://manuscripts.ru/mns/portal.main?p1=62.

The quantitative index of lemmas in the multitext query form gives information on the total number of word forms in the three lists, the number of lemmatized word forms and the number of lemmas:

– number of word forms in the subcorpus – 249 124;
– number of lemmatized word forms – 191 762[1];
– number of lemmas at present – 4 822.

The use of statistical measures also requires information on the total number of word forms in the collections with which comparisons are made and on the number of corresponding lemmas in them. Information on the five already studied collections is given in my previous works (Baranov, 2017b; Baranov, 2018) (the data being rectified as regards those collections whose texts are currently updated):

– number of word forms in the subcorpus of the Menaion for May – 99 613;
– number of word forms in the subcorpus of the Service Menaions for other months of the year – 187 748;
– number of word forms in the subcorpus of the sticherarion – 104 905;
– number of word forms in the subcorpus of the Gospels – 522 793;
– number of word forms of the subcorpus of the Apostles – 90 442;
– total number of word forms in five collections – 1 005 501.

Statistical evaluation of the most common content words

To assess the significance of lemmas most often used in the *Paraenesis* collection, two statistical measures were used: *TF-ICTF* and *Log-Likelihood*.

Measure TF-ICTF' as a variant of the statistical measure TF-IDF

One of the most commonly used statistical measures for finding meaningful (key) words in a document is TF-IDF (term frequency – inverse document frequency)[2]. In terms of its application the greatest weight (the value of a measure) is attributed to such linguistic units that are most often used in the analysed document but are not found in other corpus documents:

$$TF\text{-}IDF = \frac{f}{F} * \log\frac{D}{d},$$

[1] The lemmatization of the *Paraenesis* lists was carried out automatically with the help of a morphological analyzer and the *Manuscript* database of the grammatical dictionary of the Old East Slavic language.

[2] The statistical measure TF-IDF for evaluation of the significance of the words in the document (cf., one of the earlier works (Salton and Yang, 1973)) is based on the IDF method proposed by Sparck 1972; see also (Roelleke, 2013; Robertson, 2004) about the development of the statistical measure TF-IDF, its potential, variants and usage.

where f – number of word forms / lemmas / terms to be analysed in the document; F – number of all word forms / lemmas in the document; D – number of documents in the corpus; d – number of corpus documents in which the word form / lemma being analysed occurs.

It is clear that the evaluation with the help of this measure is ineffective when a corpus has a small number of texts. There are many modifications of this measure, which involve taking the logarithm of TF, introducing additional coefficients into calculations, taking into account not only documents, but also linguistic units in them, etc. (see, for example, Roelleke, 2013).

In this study, as well as in [Baranov, 2018], I used the TF-ICTF′ measure to analyse the material of the *Paraenesis* lists:

$$TF\text{-}ICTF' = (0{,}5 + 0{,}5\frac{f_d}{F_d}) * \log\frac{F_D - F_d}{f_D - f_d},$$

where f_d – number of analysed word forms / lemmas (terms) in the document (in this case – in the collection of documents); F_d – number of all the word forms / lemmas in the document (collection of documents); F_D – total number of word forms / lemmas in all corpus documents (in all collections); f_D – number of words / lemmas analysed in all corpus documents (in all collections).

Measure Log-Likelihood

Another frequently used measure is the Log-Likelihood function (similarity index), which allows one to evaluate the significance of a word by comparing its frequency in the analysed document with the statistically average (expected) frequency in the corpus (cf., for example, (Rayson, Garside, 2000: 3; Liashevskaia, Sharov, 2009: 8–9)):

$$LL = 2(a \ln\left(\frac{a}{c\frac{a+b}{c+d}}\right) + b \ln\left(\frac{b}{d\frac{a+b}{c+d}}\right)),$$

where a – absolute number of the analysed word form / lemma in the document (collection); b – absolute number of the analysed linguistic unit in other documents (collections) from the corpus; c – length of the document (collection) in which the unit is analysed; d – length of other documents (collection) of the corpus.

As can be seen from both formulas, the same parameters are used to evaluate the significance of linguistic units: a) the number of the analysed unit in the document (subcorpus), b) the number of all units in the subcorpus, c) the number of

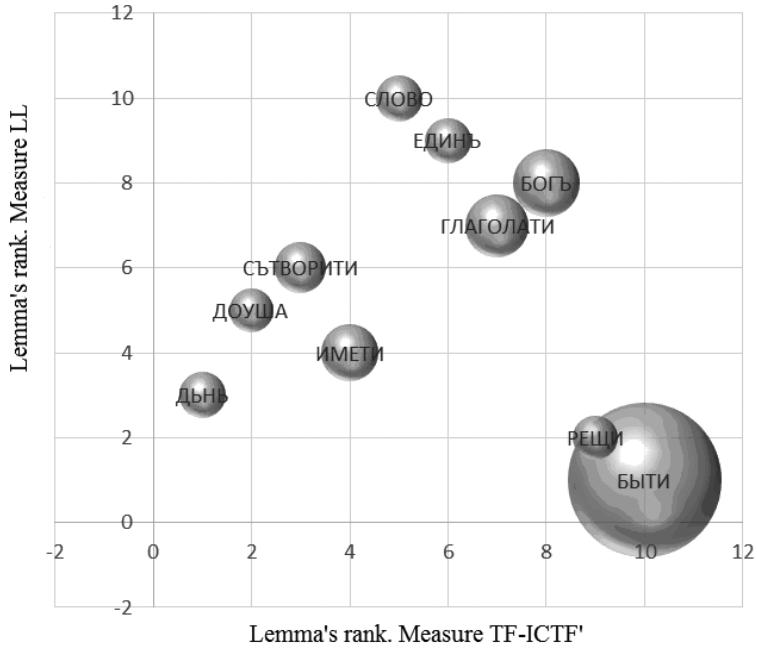

Fig. 1. The distribution of the 10 most frequent lemmas in the three *Paraenesis* lists according to their ranks on the basis of the statistical measures TF-ICTF′ and LL

the analysed unit in all other subcorpora, and d) the number of all units in all other subcorpora.

Analysis and comparison of measurement results

Statistical weights of the most frequent lemmas in the Paraenesis lists

The values (weights) of the 10 most frequent lemmas in accordance with TF-ICTF′ and LL measures are given in Appendix 1[1].

The table shows that the measures assess the significance of lemmas in different ways (sometimes in quite the opposite ways). For example, the most frequent lemma *БЫТИ*[2] (the absolute number is 7016, the relative one is 0.02816) has the lowest weight (rank 10) when the measure TF-ICTF′ is applied (0.86179) and the largest (rank 1) when the measure LL is applied (422.180).

Comparison of statistical weights of lemmas

Both measures are used to search in the corpus for such words that are assumed to be the key ones. Different estimates by measuring the significance of

[1] The values of each lemma are computed using the number of the corresponding lemma in each of the additional subcorpora analysed.

[2] Church Slavonic verb *to be*.

a particular lemma makes it necessary to seek ways of simultaneously averaging values and taking them into account. One possible way is to construct a diagram in which the values are 1) two lemma's ranks obtained by reorganization of the list in accordance with the value of the measures (see the *Rank* values in Appendix 1) and 2) the relative number of lemmas in the *Paraenesis* lists corpus. These parameters can be used to construct a three-dimensional diagram, where *x* axis is given the rank of the lemma in accordance with its weight TF-ICTF′, *y* axis corresponds to LL value, and the ball size corresponds to the relative number of the lemma in the subcorpus (see Fig. 1).

The diagram allows one to assess a) the compliance / incompliance of estimates of statistical measures with each other: in case of compliance the lemma is on the line with the coordinates 0-0 – 12-12 or close to it, if there is a discrepancy the lemma is not on the line, and the greater incompliance, the farther from the diagonal is the lemma; b) the greater / lesser significance of the lemma for the *Paraenesis* text against other corpora: if the lemma has greater significance, it is close to the origin, whereas the lemma with the smaller significance is far from the origin.

Among the first ten lemmas, both measures identically evaluate the words of *ГЛАГОЛАТИ* (meaning 'to speak, to say, to utter') and *БОГЪ* (meaning 'God'); the most significant (within the first 5 ranks) are *ДЬНЬ* (meaning 'day'), *ИМЕТИ* (meaning 'to have'), *ДОУША* (meaning 'soul').

Comparison of different in number lists of the most frequent lemmas

The maximum objective analysis shall adhere to two conditions at least: a) the search for meaningful words is carried out not only among the most frequent ones, but among all lemmas of the subcorpus; b) the reference corpus shall be the most representative (large and balanced) one. Failure to comply with these conditions in this work leads to the need to find ways to refine the obtained results. Non-compliance happens first due to the reduction of the lemmas' number to the most frequent ones, second, due to inability to comply with the latter requirement at present for objective reasons. One of these methods to improve the situation is to increase the number of lemmas analysed.

Let us consider how the composition of significant lemmas changes when their number increases.

The list of the 28 lemmas most frequently used in the *Paraenesis* lists, their relative frequency and statistical weight are given in Appendix 2 (positions from 11 to 28). In Appendix 3, the ranks of the first 10, 12, 16, 20, 24, 28 lemmas are given in accordance with each of the statistical measures.

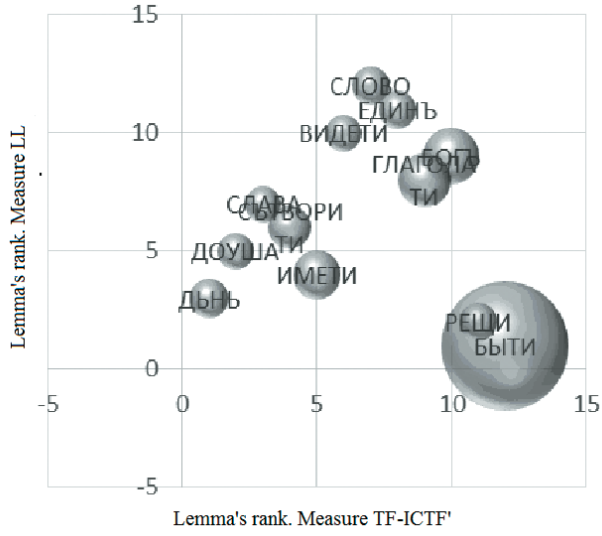

Fig. 2. Distribution of the 12 most frequent lemmas in the three *Paraenesis* lists according to their ranks on the basis of the statistical measures TF-ICTF′ and LL

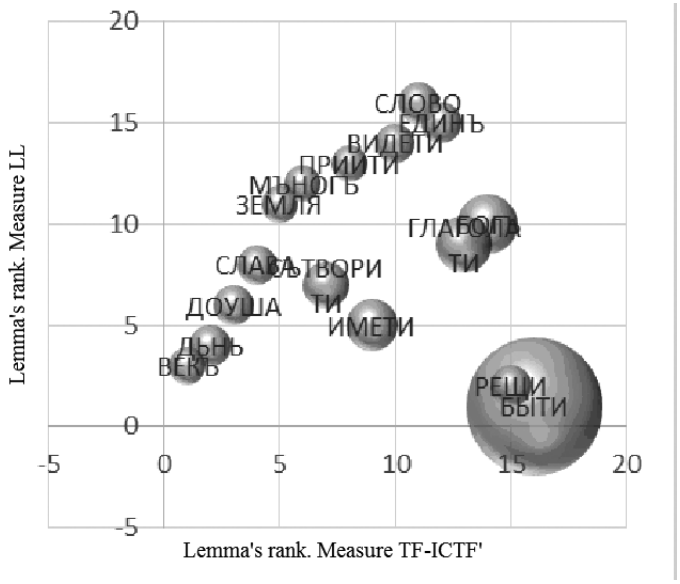

Fig. 3. Distribution of the 16 most frequent lemmas in the three *Paraenesis* lists according to their ranks on the basis of the statistical measures TF-ICTF′ and LL

Let us consider the change in the rank of certain lemmas with the widened list of analysed words.

The lemma of *БЫТИ*. In accordance with the LL measure, this lemma has the highest rank in all lists, except for the list of 24 and 28 lemmas, in which it occupies the second and third places, respectively. The measure TF-ICTF′ in all cases assigns the lemma the lowest rank.

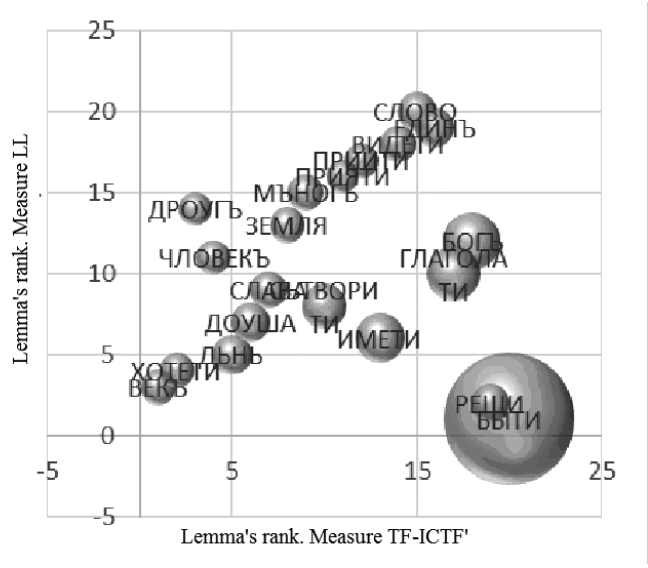

Fig. 4. Distribution of the 20 most frequent lemmas in the three *Paraenesis* lists according to their ranks on the basis of the statistical measures TF-ICTF′ and LL

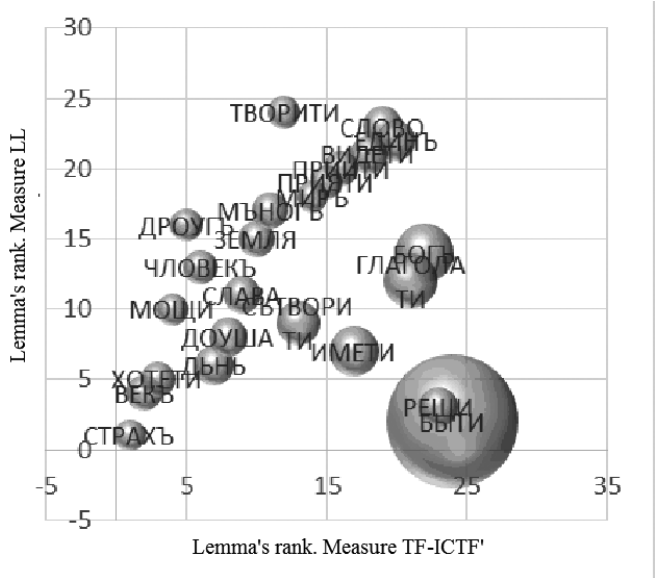

Fig. 5. Distribution of the 24 most frequent lemmas in the three *Paraenesis* lists according to their ranks on the basis of the statistical measures TF-ICTF′ and LL

The lemma of *СЛОВО* (meaning ‘word’). The weight of this lemma with respect to the LL measure is the lowest one or one of the lowest in all lists. The TF-ICTF measure has it in the middle of the list of 10 words, in other lists the lemma falls even below the middle of the lists. But lemmas, occurring in the body less often, gradually occupy higher ranks.

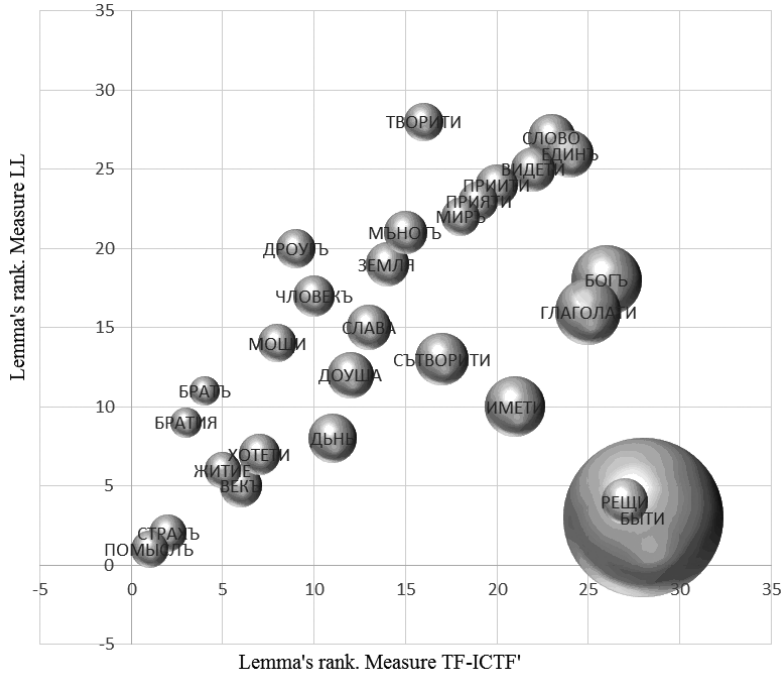

Fig. 6. The distribution of the 28 most frequent lemmas in the three *Paraenesis* lists according to their ranks on the basis of statistical measures TF-ICTF′ and LL

A clear change in the ranks of lemmas and their correlations in accordance with the measures is shown in Fig. 2–5 and Fig. 6.

The diagrams demonstrate the following trends that occur when the number of the analyzed units increases:

– lemmas, located in the core of small lists, move to the periphery with an increase in the number of words analysed (see the place of the lemmas *ДЬНЬ, ДОУША, ИМЕТИ*);

– their place is occupied by the lemmas of *ПОМЫСЛЪ* (meaning ‘thought’) and *СТРАХЪ* (meaning ‘fear’), which occupy the 26th and 24th positions in the list of 28 words, respectively, as well as the lemmas of the nearest periphery – *ВЕКЪ* (meaning ‘century, age’, 15th position), *ЖИТИЕ* (meaning ‘existence’, 25th position), *ХОТЕТИ* (meaning ‘to wish, to want’, 17th position), *БРАТИЯ* (meaning ‘brethren, fraternity’, 27th position);

– the lowering of the lemma’s rank in accordance with the TF-ICTF′ measure occurs more quickly as the list of lemmas increases (see, for example, the rank lowering for the lemmas of *ГЛАГОЛАТИ* and *БОГЪ* from 7th and 8th positions to 25th and 26th if TF-ICTF′ is applied, and from 7th and 8th positions to only 16th and 18th places if LL measure is applied);

– the opposite estimation by means of measuring the significance of the lemmas of *БЫТИ* and *РЕЩИ* (meaning 'to say, to enunciate') is preserved (compare their remoteness from the central line of all the diagrams).

Composition and thematic focus of the Ephrem the Syrian's Paraenesis lists and statistically significant lemmas

As we know, Ephrem the Syrian's *Paraenesis* has been one of the most popular menology and service books in the Slavic world and in Medieval Russia (Zholobov, 2007a: 7; Zholobov, 2007b: 31-35; Zholobov, 2008: 52). The earliest ancient Old East Slavic lists (Pogodin's, Typographic, Troitsky, Academic, Frolov's), being somewhat different in composition and structure, time and place of creation, contain several dozens of teachings by Ephrem the Syrian, the Life of Ephrem or excerpts from it, and also other works, for example, the story of Abraham and Joseph the all-comely[1]. The main theme is instruction on the soul salvation (Zholobov, 2007a: 10; Zholobov, 2007b: 33). Important for understanding the theme and focus of the *Paraenesis* texts is the statement by O. Zholobov: "It is very interesting that the oldest surviving Pogodin's list of Ephrem the Syrian's *Paraenesis* was made by laymen, although Ephrem the Syrian's teachings were intended for monks, so they demonstrate vivid ascetic, mystical and eschatological motifs. In the culture which gave birth to this source no distinction was made between the spiritual admonition of cloisterers and the religious upbringing of the flock who lead the ordinary life" (Zholobov, 2007a: 12).

It is not accidental that in the course of the analysis the high statistical values of the lemmas – *ПОМЫСЛЪ* and *СТРАХЪ*, as well as of the lemmas – *ВЕКЪ, ЖИТИ, ХОТЕТИ, БРАТИЯ* were identified, as these words fall in the range of semes representing the themes of monastic and worldly life, the focus on spiritual search and the motifs of punishment and the end of the world.

The convergence of the general theme of the works in the *Paraenesis* subcorpus and the results of the analysis undertaken in the study, during which the thematically significant lemmas were found, proves the objectivity of the results of the statistical methods application and promotes using them to search for thematically and semantically significant linguistic units characteristic of a certain type or genre of medieval works.

[1] The author thanks Professor O.F. Zholobov who kindly provided information on the composition of the Old East Slavic lists.

Conclusion

As shown by the analysis, the statistical methods often used to search for keywords in documents that are contained in large corpora are quite effective when applied to the study of small in terms of volume historical corpora when one needs to identify the topic, the content of documents, to compare the contents of documents among themselves.

Relatively small size of the historical corpus also has its advantages: the availability of information about the type and genre of texts allows at the stage of their selection for analysis to assume that they undoubtedly have individual linguistic characteristics, so statistical analysis is necessary not for assigning the text to a particular thematic group, but for elucidating quantitative-statistical and thematic-content characteristics that oppose the text (collection / subcorpus) to others.

It was this methodology that was used in this work, the material for which were the three Slavonic lists of Ephrem the Syrian's *Paraenesis* selected for obtaining primary quantitative data on the basis of their metacharacteristics with the help of appropriate corpus modules. The volume of five different collections (more than 1 million word forms) used for comparison provided a sufficient experimental base. The application of two statistical measures made it possible to identify several maximum statistical weights of lemmas of the content words from the 28 most frequently used in the lists. Bringing in the general information on the content of the *Paraenesis* texts leads to conclusion that core *lemmas ПОМЫСЛЪ, СТРАХЪ,* as well as lemmas – close to core ones *ВЕКЪ, ЖИТИ, ХОТЕТИ, БРАТИЯ,* bear the thematic and semantic significance for the collection.

Removing the two limitations specified at the beginning (extending the database from 5 analysed collections to all the corpus documents with the subsequent analysis of all the units, not only the most frequent ones, in the relevant manuscript or group of lists) will certainly increase the accuracy and improve informative results. But even the limited amount of material, that is possible to use currently, gives, indisputably, indicative results – a list of words thematically and semantically meaningful for the analysed collection.

Electronic collections and corpora of ancient Russian written artefacts

Church-Slavonic corpus, In *The National Corpus of the Russian language,* available at: http://www.ruscorpora.ru/search-orthlib.html

Corpus of birch bark manuscripts, In *The National Corpus of the Russian language,* available at: http://www.ruscorpora.ru/search-birchbark.html

Corpus of birch bark manuscripts in Old East Slavic, available at: http://gramoty.ru/

Old East Slavic corpus, In *The National Corpus of the Russian language,* available at: http://www.ruscorpora.ru/search-old_rus.html

Old Russian corpus, In *The National Corpus of the Russian Language,* available at: http://www.ruscorpora.ru/search-mid_rus.html

Old Russian Texts, In *Pragmatic Resources in Old Indo-European Languages*, available at: http://foni.uio.no:3000

Old Russian Texts, In *TITUS*, available at http://titus.uni-frankfurt.de/indexe.htm

Povest' vremennykh let [The Tale of Past Years], D. Birnbaum (ed.), D. Ostrowski et al. (eds.), available at: http://pvl.obdurodon.org/

Regensburg Russian Diachronic Corpus, available at: http://rhssl1.uni-regensburg.de/SlavKo/korpus/rrudi-new

St. Petersburg corpus of hagiographic texts, available at: http://project.phil.spbu.ru/scat/page.php?page=project

The Manuscript corpus, available at: manuscripts.ru

Sources

Collection of teachings of Ephrem the Syrian, 1270s–1280s. In *Russian State Archive of Ancient Acts*, Typ., 38, 143 p., available at http://manuscripts.ru/mns/main?p_text=86892772 (accessed 31.12.2017).

Collection of teachings of Ephrem the Syrian, circa 1269–1289. In *National Library of Russia*, Pog., 71a, 328 p., available at http://manuscripts.ru/mns/main?p_text=88512667 (accessed 31.12.2017).

Collection of teachings of Ephrem the Syrian, middle of the 14th century. In *Russian State Library*, Tr., 7, 245 p., available at http://manuscripts.ru/mns/main?p_text=87272654 (accessed 31.12.2017).

References

Baranov, V.A. (2017a). Kolichestvennyi i statisticheskii analiz srednevekovykh slavianskikh tekstov: instrumentarii korpusa "Manuskript" i metodika ego ispol'zovaniia [Quantitative and statistical analysis of medieval Slavonic texts: tools of the *Manuscript* corpus and the methodology of its use], In *Tsifrovaia gumanitaristika: resursy, metody, issledovania: materialy mezhdunarodnoi nauchnoi konferentsii v dvukh chastiakh*

[Digital Humanitaristics: resources, methods, research: Materials of the international scientific conference in 2 vols.] (Perm, May 16–18, 2017), 40–49. Perm State National Research University, available at: https://goo.gl/zyb75Q (accessed 31.12.2017).

Baranov, V.A. (2017b). Statisticheski znachimye slova kak kharakteristika srednevekovogo slavianskogo teksta (na material kollektsii Apostolov istoricheskogo korpusa "Manuskript" Statistically significant words as a characteristic of the medieval Slavonic text (on the basis of the Apostles' collection from the historical *Manuscript* corpus), In *Gumanitarnoe obrazovanie I nauka v tekhnicheskom vuze: Sbornik dokladov Vserossiiskoi nauchno-prakticheskoi konferentsii s mezhdunarodnym uchastiem [Education and research in the humanities in a technical university: Proceedings of the All-Russian scientific and practical conference with the international participants]* (Izhevsk, October 24-27, 2017), 359–369. Izhevsk, Publishing House of Izhevsk State Technical University named after M.T. Kalashnikov, available at: https://goo.gl/9QegRb (accessed 31.12.2017).

Baranov, V.A. (2018). *Opyt primeneniia kolichestvennykh i statisticheskikh metodov dlia poiska znachimykh slov v istoricheskom korpuse (na materiale srednevekovykh slavianskikh gimnograficheskikh i evangel'skikh kodeksov) [Experience of applying quantitative and statistical methods to find meaningful words in the historical corpus (based on the medieval Slavic hymnographic and evangelical codes).* Germany, Bonn (in press), available at: https://goo.gl/yX4Udd (accessed 31.12.2017).

Liashevskaia, O.N., Sharov, S.A. (2009). Vvedenie k chastotnomu slovariu sovremennogo russkogo iazyka [Preface in the frequency dictionary of the modern Russian language], In *Chastotnyi slovar' sovremennogo russkogo iazyka (na materialakh Natsional'nogo korpusa russkogo iazyka) [Frequency dictionary of the modern Russian language (on the materials of the National Corpus of the Russian language)]*, 8. Moscow, "Azbukovnik", available at: http://dict.ruslang.ru/freq.pdf (accessed 31.12.2017).

Rayson, P., Garside, R. (2000). Comparing corpora using frequency profiling, In *Proceedings of the Comparing Corpora Workshop at ACL 2000*, Hong Kong, 1–6, available at: http://ucrel.lancs.ac.uk/people/paul/publications/rg_acl2000.pdf

Robertson, S. (2004). Understanding inverse document frequency: on theoretical arguments for idf, In *Journal of Documentation*, 60, 503–520, available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.438.2284&rep=rep1&type=pdf

Roelleke, T. (2013). Information Retrieval Models: Foundations and Relationships, In *Synthesis Lectures on Information Concepts, Retrieval, and Services*, July 2013,

Vol. 5, No. 3, 1–163, available at: https://www.morganclaypool.com/doi/abs/10.2200/S00494ED1V01Y201304ICR027

Salton, G., Yang, C.S. (1973). On the specification of term values in automatic indexing, In *Journal of Documentation*, 29, 351–372.

Sparck, K.J. (1972). A statistical interpretation of term specificity and its application in retrieval, In *Journal of Documentation*, 28, 11–21.

Zholobov, O.F. (2007a). Poucheniia Efrema Sirinav intertekstual'nykh i kompozitsionnykh otzvukakh original'noi drevnerusskoi pis'mennosti [The Ephrem the Syrian's teachings in the intertextual and compositional echoes of the original Old East Slavic writings], In *Vestnik PSTGU. Seriia "Filologiia" [Bulletin of the Saint Tikhon's Orthodox University. Series "Philology"],* 3 (9), 7–13.

Zholobov, O.F. (2007b). Korpus drevnerusskikh spiskov Parenesisa Efrema Sirina, I, RGADA, Sin. 38 [The corpus of Old East Slavic lists of Ephrem the Syrian's *Paraenesis*, I, Russian State Archive of Ancient Acts, Syn. 38*], In Russian Linguistics,* 31(1), 31–59.

Zholobov, O.F. (2008). Drevneslavianskie spiski Parenesisa Efrema Sirina: novye dannye i novye aspekty issledovaniia [The Old East Slavic lists of Ephrem the Syrian's Paraenesis: new data and new aspects of the research], In *Pis'mo, literatura i fol'klor slavianskikh narodov, XIV Mezhdunarodnyi s'ezd slavistov, doklady rossiiskoi delegatsii [Writing, literature and folklore of the Slavic peoples, XIV International Congress of Slavicists, Talks of the Russian delegation].* (Ohrid, September 10-16, 2008), 51–75. Moscow, Publishing House "Indrik".

Appendix 1. The list of the 10 most frequent lemmas from three lists of the Ephrem the Syrian's *Paraenesis*, their number, relative frequency, TF-ICTF′ and LL values, ranks

№	Lemma	Number	Frequency	TF*ICTF'		LL	
				Rank	Weight	Rank	Weight
1	бꙑти	7016	0,02816	10	0,86179	1	422,180
2	богъ	1329	0,00533	8	1,18908	8	44,542
3	глаголати	1164	0,00467	7	1,22836	7	58,739
4	имѣти	967	0,00388	4	1,32878	4	185,103
5	съ творити	755	0,00303	3	1,36496	6	110,111
6	дьнь	624	0,00250	1	1,46265	3	199,660
7	слово	604	0,00242	5	1,31663	10	0,337
8	иднъ	597	0,00240	6	1,29977	9	1,912
9	рєщи	575	0,00231	9	1,15824	2	352,072
10	доуша	570	0,00229	2	1,46073	5	142,038

Appendix 2. The list of the most frequent lemmas (11-28) from three lists of Ephrem the Syrian's *Paraenesis*, their number, relative frequency, TF-ICTF′ and LL values, ranks

№	Lemma	Number	Frequency	TF*ICTF'	LL
11	слава	536	0,00215	1,42534	60,934
12	видѣти	529	0,00212	1,31845	4,054
13	зємля	499	0,00200	1,42378	38,332
14	мъногъ	491	0,00197	1,41048	22,863
15	вѣкъ	483	0,00194	1,61588	340,016
16	приити	468	0,00188	1,33315	8,889
17	хотѣти	460	0,00185	1,58464	244,849
18	чловѣкъ	440	0,00177	1,46518	47,412
19	дроугъ	414	0,00166	1,46925	36,353
20	прияти	407	0,00163	1,36000	9,379
21	творити	393	0,00158	1,40476	0,016
22	мощи	388	0,00156	1,52623	76,440
23	миръ	380	0,00153	1,36019	17,860
24	страхъ	374	0,00150	1,81866	501,671
25	жити	360	0,00145	1,67776	251,553
26	помꙑслъ	353	0,00142	2,22050	913,759
27	братия	246	0,00099	1,77887	191,890
28	братъ	235	0,00094	1,76097	155,666

Appendix 3. The ranks 10, 12, 16, 20, 24, 28 of the most frequent lemmas from three lists of Ephrem the Syrian's *Paraenesis* in accordance with the TF-ICTF′ and LL values

№	Lemma	TF-ICTF'						LL					
		10	12	16	20	24	28	10	12	16	20	24	28
1	бꙑтн	10	12	16	20	24	28	1	1	1	1	2	3
2	богъ	8	10	14	18	22	26	8	9	10	12	14	18
3	глаголати	7	9	13	17	21	25	7	8	9	10	12	16
4	имѣти	4	5	9	13	17	21	4	4	5	6	7	10
5	сътворити	3	4	7	10	13	17	6	6	7	8	9	13
6	дьнь	1	1	2	5	7	11	3	3	4	5	6	8
7	слово	5	7	11	15	19	23	10	12	16	20	23	27
8	иднъ	6	8	12	16	20	24	9	11	15	19	22	26
9	рєщи	9	11	15	19	23	27	2	2	2	2	3	4
10	доуша	2	2	3	6	8	12	5	5	6	7	8	12
11	слава	–	3	4	7	9	13	–	7	8	9	11	15
12	видѣти	–	6	10	14	18	22	–	10	14	18	21	25
13	ꙁємлꙗ	–	–	5	8	10	14	–	–	11	13	15	19
14	мъногъ	–	–	6	9	11	15	–	–	12	15	17	21
15	вѣкъ	–	–	1	1	2	6	–	–	3	3	4	5
16	приити	–	–	8	12	16	20	–	–	13	17	20	24
17	хотѣти	–	–	–	2	3	7	–	–	–	4	5	7
18	оуловѣкъ	–	–	–	4	6	10	–	–	–	11	13	17
19	дроугъ	–	–	–	3	5	9	–	–	–	14	16	20
20	прияти	–	–	–	11	15	19	–	–	–	16	19	23
21	творити	–	–	–	–	12	16	–	–	–	–	24	28
22	мощи	–	–	–	–	4	8	–	–	–	–	10	14
23	миръ	–	–	–	–	14	18	–	–	–	–	18	22
24	страхъ	–	–	–	–	1	2	–	–	–	–	1	2
25	житиє	–	–	–	–	–	5	–	–	–	–	–	6
26	помꙑслъ	–	–	–	–	–	1	–	–	–	–	–	1
27	братиꙗ	–	–	–	–	–	3	–	–	–	–	–	9
28	братъ	–	–	–	–	–	4	–	–	–	–	–	11

Опыт статистического анализа славянского Паренесиса Ефрема Сирина (на материале электронной коллекции трех списков XIII–XIV вв. корпуса «Манускрипт»)

В.А. Баранов
Ижевский государственный технический университет имени М.Т. Калашникова
Россия, 426069, Ижевск, ул. Студенческая, 7

Представлен опыт использования статистических мер для поиска тематически значимых слов в трех древнерусских списках Паренесиса Ефрема Сирина.

Количественные данные получены с помощью поисковых форм исторического корпуса «Манускрипт» (manuscripts.ru) – многотекстового модуля и модуля n-грамм. Базовым корпусом для анализа 28-ми наиболее частотных лемм знаменательных слов Паренесиса (объем коллекции – более 100 тыс. словоформ) стали пять разножанровых коллекций корпуса – списков майской минеи, служебных миней на другие месяцы года, стихирарей, Апостола, Евангелий (общий объем – более 1 млн словоформ).

Для оценки значимости лемм, полученных с помощью автоматического морфологического анализатора системы, использованы статистические меры TF-ICTF' (вариант меры TF-IDF) и Log-Likelihood. Увеличение количества анализируемых лемм с 10-ти до 28-ми позволило продемонстрировать больший статистический вес лемм, которые используются реже, чем максимально частотные.

Для устранения расхождений в статистической оценке лемм осуществлено сравнение рангов лемм и построены диаграммы. Анализ диаграмм позволил выявить ядро и периферию перечней и определить леммы с наибольшим усредненным статистически весом – ПОМЫСЛЪ и СТРАХЪ, а также леммы ближайшей периферии ВЕКЪ, ЖИТИ, ХОТЕТИ, БРАТИЯ, которые репрезентируют направленность текстов Паренесиса на духовный поиск и мотивы наказания и конца света.

Сделан вывод о результативности и эффективности применения статистических методик к оценке лингвистических данных исторических корпусов, которые сегодня в силу объективных причин существенно уступают современным по объему, и о необходимости расширения материалов для анализа за счет привлечения данных всего корпуса славянских письменных памятников «Манускрипт» и всего списка словоформ (лемм) анализируемой рукописи (подкорпуса).

Ключевые слова: корпусная лингвистика, лингвистическая статистика, средневековые славянские рукописи, Паренесис Ефрема Сирина, ключевые леммы.

Статья написана при поддержке Российского фонда фундаментальных исследований (РФФИ) в рамках проекта «Лингвостатистический анализ однокомпонентных и многокомпонентных лексических единиц исторического корпуса "Манускрипт"» (грант № 18-012-00463).

Научная специальность: 10.02.00 – компьютерная лингвистика.