The offered work is devoted to *interrelated problems* of estimation of text affinity to the most rational (i.e., standard) form for transfer of its sense and forming the representative (i.e., reference) text collection, concerning which the estimating itself is carried out.

It's necessary to note, that in practice it's required quite often to collect publications concerning the given topic. In addition to scientific research, such task is actual when are e-learning materials being prepared. The most significant here, as a rule, are papers, for which at maximal disclosure of topic interesting for the end user a maximum of an average number of the most significant terms per a one simple spread sentence at minimum of its length measured in words, is typical. Substantially, this corresponds to the most brief, but succinct narration, that satisfy to the standard variant of sense transfer in a given natural language (see *slide 2*). The given requirement can be reformulated as follows: texts in a collection should be relevant as much as possible to the given topical area from the point of expert view both in vocabulary, and in internal relations (syntactic, semantic, etc.). The text collection itself here is called *representative (or reference)*. For a more detailed analysis collection texts can be labeled according to some determined rules (syntactically, by application of a base of role dependencies, etc.). In such case we have a reference corpus, which can be an initial data for machine learning of recognition of dependencies in texts related to a given scope.

However, the problem of minimization of the handwork of expert at preparation of such collections is actual here. The most advisable for expert here is to use short texts which are comparable in vocabulary and (possible) in relationships between words with documents being added to the reference collection. In a role of such texts abstracts of scientific articles or other texts that are resume facts significant from the point of expert view relatively the given topical area, can entirely be. Here we have the task inverse to the abstractive summarization: to find a text, in which general ideas described in an abstract (or in a collection of abstracts) are reflected the most fully.

In this study, the variant to solve this problem using of shares of non-zero values of word frequency in an analyzed document, calculated by phrases of abstracts, is represented.

In the «classic» problem statement (see *slide 3*) the task of measuring the text cognitive complexity was solved on the basis of quantiles of empirical distributions of frequency of tokens over the reference corpus. In such statement for each linguistic level, its own alphabet of tokens is defined. For example, words or terms for the lexical level, types and lengths of syntactic links for syntactic one. Herewith the occurrence frequency for token is considered as abnormally high if it is greater than 95th percentile of its frequency in a reference corpus of texts which are not complicated for the implied readership. The considered task of document selection to the reference corpus on the base of collection of abstracts is the task diametrically opposed to the mentioned measuring the text cognitive complexity – namely, tokens from abstracts should be also represented in the analyzed document as much as possible. For intuitive clarity, further in the reasoning we will restrict our consideration to the lexical level, for other linguistic levels (i.e., phonetic, morphological, syntactic, discursive) reasoning is carried out similarly. In such problem statement, we are talking about the minimum necessary representation of words (terms) from abstracts in a document. It's reasonable to assume that the 5^{th} percentile (i.e., 5% quantile) of frequency characteristic of word relatively to the given document here we should consider.

The next step is the choice of frequency characteristic itself (see *slide 4*). The main requirement here is independence from the number of document words. Let's calculate for each phrase in each abstract the share of non-zero values of TF-measure for words of the phrase relatively to the analyzed document. TF-measure (*term frequency*) is the ratio of the number of occurrences of a word in a document to the total number of document words. One phrase here corresponds to the simple spread natural-language sentence (according to the terminology of «Meaning⇔Text» approach). Since the share of complex sentences in real abstracts is minimal, it is quite acceptable to apply this term to sentences as part of abstracts. In order to reflect the content of the articles as much as possible, their abstracts will be considered together with the titles. It is admissible here, that the same phrase may appear in more than one abstract of the collection (for example, if these articles denoted to the same author). In any case each phrase is accepted to consideration only once.

Note that using exactly the share of non-zero values of TF-measure, and not the *term frequency* values themselves for words of a phrase, allows solving the problem of dependence of significance estimation value for a document from the number of words in it. Indeed, only the presence of the maximum number of words from the abstracts in the analyzed document is important, while the frequency of individual words is not principled here.

Basic ideas of document significance estimation for adding to the reference collection are represented on *slides 5* and *6*. Herewith for each word of each phrase of each abstract from the collection formed by an expert the value of TF-measure relatively to estimated document, is calculated, and for a separate phrase the share of non-zero TF values is evaluated according to the formula (2) on the *slide 5*. Further, the 5^{th} percentile of empirical distribution of estimation (2) value concerning the analyzed document for the given collection of abstracts, is entered into consideration. We'll associate the collection of abstracts here with the combining of sets of phrases for separate abstracts. Let's sort documents that are candidates for adding to the reference corpus, by decreasing of the value of the mentioned quantile. Herewith for each of them the vector of quantiles values is entered into consideration, Into this vector deciles together with the first and third quartiles will be included in addition to above-mentioned 5^{th} and 95^{th} percentiles. For each of obtained vectors (see *slide 6*) the Euclidean distance to document with the maximal value of the 5^{th} percentile of empirical distribution of the share of non-zero TF values for a given collection of abstracts. The sequence of vectors for documents that are reference corpus candidates, is splitted into clusters according to the value of mentioned distance. Herewith (see *Statement 1* on the *slide 6*) the most significant for the target collection will be a document having the maximal value of the 5^{th} percentile together with documents related to the cluster of the least distances to it.

To improve the *recall* of search of significant documents for reference collection the above-mentioned classification of documents that are candidates for adding should be implemented independently for several collections of abstracts of articles devoted to close scopes. The recall of search is estimated here by the ratio of the number of documents that meet the condition of *Statement 1* and classified as significant by an expert, to the total number of documents from recognized as significant by the same expert.

The search accuracy for significant documents in the problem considered here largely depends on the content of the used collection of abstracts. Substantially, here it is required to maximize the estimations proposed by us earlier for affinity to the semantic standard, for each of the abstracts. To estimate the abstract significance at selection of

documents to the target collection the value of the 5th percentile of empirical distribution corresponding to the array of shares of non-zero TF values, is considered relatively to the analyzed abstract and compared with the value of the same percentile concerning the united set of phrases from all abstracts of the collection (see *slide 7*). Herewith among abstracts of the collection the five groups can be distinguished according to the *Statement 2*, where the highest precision for search of significant documents is reached with abstracts of groups from first to third. Using the ranking of abstracts inside the groups according the value of mentioned percentile, we obtain an alternative variant of estimation of closeness of short texts to the semantic standard – namely, *1st group* abstracts will be closest to the sense standard. Recall that the assessing variant proposed by us earlier for proximity of the text to the sense standard is based on the division of words of each of its phrases into classes according to the value of the TF-IDF measure.

The experimental material to test the proposed approach is represented on *slides 8–10*. The software implementation (in Python 2.7) of the offered solutions and experimental results are presented on the website of Yaroslav-the-Wise Novgorod State University. For the more accurate revelation of lexical context for terms the calculation of *term frequency* values was made without taking into account of prepositions and conjunctions.

Further, *Table 1* on the *slide 11* represents documents that meeting the condition of *Statement 1* and recognized by an expert as significant from the experiments with four collections of abstracts mentioned on the *slide 10*. For each document the table represents the number of phrases (N_1), the total number of words with the respect of all occurrences of each word (N_2), the number of collections of abstracts, where the document meets the *Statement 1* condition (N_3). Note, that among the documents being candidates for adding and recognized by an expert as significant for reference corpus formation, in the considering series of experiments only for one of them the condition of *Statement 1* was not fulfilled, what demonstrates the recall value approximately equal to 5/6. As for the accuracy of the search, that in experiment on the collection for the «Methods and Models of Pattern Recognition and Forecasting» section of the proceedings of the 14th All-Russian Conference «Mathematical Methods for Pattern Recognition» (2009), in addition to the documents presented in *Table 1*, the scientific report mentioned on *slide 8*, which was not recognized by the expert as significant in the problem being solved, was identified as meeting the condition of *Statement 1*. For this document we have $N_1 = 65$, $N_2 = 1626$, $N_3 = 1$, which, in comparison with other candidate documents, indicates a significant dependence of the accuracy of the search by the method proposed in our study both on the structure of the collection of abstracts and on the number of phrases in the document. For comparison, *Table 2* on the *slide 11* represents candidate documents that not meet the condition of *Statement 1*, but concerning which using the estimation variant offered by us earlier the fact of maximal affinity to standard was possible to establish at least in one phrase in experiments with the collection of abstracts for «Statistical Learning Theory» section of the proceedings of the MMPR-15 Conference (2011).

Table 3 on the *slide 12* represents the result of ranking of abstracts according to conditions of *Statement 2* for the aforementioned collection for «Statistical Learning Theory» section of the proceedings of the MMPR-15 Conference. For comparison, the document with the maximal value of the 5th percentile of empirical distribution of the share of non-zero values of TF-measure for the given collection of abstracts here has the *serial number 2* by *Table 1* on the *slide 11*. As can be seen from the results in *Table 4* on the *slide 13*, abstracts of the considered collections are related to one cluster according to

the value of the 5th percentile of empirical distribution of the share of non-zero values of TF-measure except for the article whose serial number is 10 in *Table 3* on the *slide 12*. As will be seen later in the experimental results presented on *slides 17* and *18*, according to the variant of estimating the affinity of the text to the sense standard proposed by us earlier, the least value of this estimation was obtained relative to the title of this article, which confirms the compliance of the previously proposed classification by affinity to the standard and the classification of abstracts according to the conditions of *Statement 2*. In addition, when splitting papers from *Table 3* on the *slide 12* into clusters according to the value of the 5th percentile of empirical distribution of the share of non-zero values of TF-measure by phrases of corresponding abstracts with adding the value of mentioned percentile for phrases of abstracts of all collection articles into the sequence for splitting, we'll obtain two clusters: to the first will be related articles with serial numbers 1 and 2, all others will be related to the second. In experiments, results of which are presented on *slides 17* and *18*, the article with the serial number 2 received the maximum value of the both of represented on the *slide 14* estimations of affinity to the sense standard: relative to the article title and phrase with the closest proximity to the standard. The essence of the method we proposed earlier for estimating the proximity of a text to the standard is introduced on *slides 14–16*. The topical corpus documents, concerning which the affinity to the standard is estimated, are sorted descending the values of the product of estimations (4), (5) and (6) presented on the *slide 14*. As the numerical estimation of the closeness of an individual phrase to the sense standard the greatest of the resulting values herewith is taken. An additional confirmation of the compliance of the results in *Tables 3–6* is a negligible difference in the values of the 5th percentile of empirical distribution of the share of non-zero values of TF-measure by phrases of abstracts with serial numbers 1 and 2 from *Table 3*.

The main result of current study is the *proposed method for formation of reference text collection for revelation of dependencies within texts of a given scope*. Dependencies here can be arbitrary and not restricted to the co-occurrence of lexical units and their relationships typical for the most rational (i.e. standard) sense transfer. It's necessary to note, that the higher estimation of significance for reference collection will have those documents, which at greater number of phrases a higher average number of the most significant terms per a one phrase at minimum of its length, contain. Substantially, this corresponds to the most brief, but succinct narration, that satisfy to the «good manner» rule of publications in Physics, Mathematics and Technical Sciences.

It is of interest to develop the alternative variant offered in current study for estimation of affinity of a text to the sense standard based on *Statement 2* on the *slide 7* regard to the usage of different documents that defined by *Statement 1* on the *slide 6* for a given collection of abstracts instead of the document with the maximum of significance estimation. The final estimation of the affinity to the standard here will be defined by the harmonization of the results for classifications of abstracts relatively to different documents from those selected to the corpus, for example, by a mutual comparison of the estimations calculated during the verification of *Statement 2* conditions.

To improve the search accuracy for significant documents, it is of interest to adapt offered estimations to other linguistic levels in addition to lexical. The comparison of classifications relatively to different levels allows making a conclusion about document significance in disputable cases, for example, at non-fulfillment of *Statement 1* condition on one of the levels.