The offered work is devoted to the interrelated problems (*see slide 2*) of extracting knowledge units from a set of subject-oriented texts (the so-called corpus), selecting texts to the corpus by analyzing the relevance to the initial phrase and completeness of reflection of revealed actual knowledge in initial phrases. The problems are of importance when constructing systems for processing, analysis, estimation and understanding of information, in particular, for knowledge testing by means of open-form test assignments. The most natural knowledge source here will be the scientific papers of highest rank scholars in appropriated topical area. The main practical goal here is finding the most rational variant to transfer the meaning in a knowledge unit defined by a set of subject-oriented natural language (NL) phrases equivalent-by-sense (i.e. semantically equivalent, SE). This decision means the following expert tasks to be automated (*see slide 3*):

- to search SE-forms for description of reality fragment in the given NL;
- to compare the knowledge of given expert with the closest knowledge fragments of another experts.

It is necessary to note that text significance here, as a rule, is unrelated to the image representing the initial phrase in analyzed texts. The requirements to interrelations of constituents of an image revealed in text are can be formulated as follows (*see slide 4*):

- in analyzed text a fragment, which corresponds to image component, can be identified with some semantic relation of words in initial phrase;
- the coupling strength of words of each such fragment is always greater than between any word from given fragment and a word not related to it;
- loosely coupled words of initial phrase cannot be related to the same fragment according to definition. It is obviously, that for terms prevailing in corpus, a combinations with a general vocabulary can be related to the extracted image component only at presence of fragments with a greater coupling strength of words;
- the links of words of different phrases from the set of initial mutually equivalent or complementary in sense and related to the same image are allowable.

In addition, generally not be required the presence of strictly predetermined part of components of initial phrase's image in text. To correctly extract this image it is necessary to analyze of occurrence both for separate words and their combinations with the estimation of coupling strength of words relatively to text and corpus. Itself, the initial phrase only in a few cases meet the standard for comparison. Herewith in some cases it is advisable to increase the quantity of initial phrases to two or more for more accurate description of the represented knowledge fragment expressed in conceptual relationships. In addition, the consideration of word combinations only within bigrams and relatively to NL-syntax is unwanted here if the percentage of general vocabulary and terms of topical area are comparable. It is actual, for example, for texts related to the fields of philosophy of science and techniques close to artificial intelligence.

To solve the given range of problems in current work *n*-grams on sequences of pairs of words related either syntactically or by sense are entered into consideration with simultaneous elimination of requirement of strict orientation on NL-syntax at word relations revelations (*see slide 5*).

«Classic» *n*-grams (*L*-grams, according to C. Shannon) as the sequences of *n* elements are widespread in mathematical investigations, biology and in information retrieval. The closest to the problematic considered in current work are syntactic *n*-grams that are determined not by linear structure of text but by routs in dependency trees or constituent trees. Let's note that at consideration of coupling strength of words of initial phrase relatively to text as a basis of its relevance estimation here such routs should be measured not

from tree root but from the word combinations with greatest values of coupling strength. Unlike the chunking of sentences in Russian based on conditional random fields, chunks here can contain prepositions and conjunctions, what is important for search of Russian paraphrasing tools for initial phrase in corpus texts.

As an estimation of «coupling strength» of words in current work the *estimation (1) represented on the slide 5* is taken. Among the estimations applied in Distributive-Statistical Method of Thesaurus Construction this estimation being close to Tanimoto coefficient is the most evident from one side, and respects the individually occurrence of each word – from the other. The mentioned method is substantially close to considered problem of revelation in analyzed texts the image represented by initial phrase. The main hypothesis of the method is the existence of some relation between words which are co-occurred within some text interval, in particular, within the same phrase. Herewith there are no any restrictions for applied estimations of co-occurrence of words.

As the basis of revelation of links of words in current work the splitting of words of initial phrase according to their values of TF-IDF metrics as an alternative and in addition to syntactic dependences is taken. In text analysis and informational retrieval TF-IDF is a numerical statistic that is intended to reflect how important a given word is to some document being a member of some corpus. According to classic definition mentioned on the *slide 6*, TF-IDF is the product of two statistics: term frequency (TF) and inverse document frequency (IDF). Term frequency is the quotient of number of times that the word occurs in document by total number of words in this document. The inverse document frequency is a measure of how much information the word provides, that is, whether the designated term is common or rare in corpus.

It is necessary to note (see *slide 7*) that with the growth of word's occurrence frequency in corpus documents the value of IDF metrics for this word tends to zero. It is true both for general vocabulary (for example, function words) and for those terms which are prevail in corpus. At the same time, for example, the words from general vocabulary which are define the conversive replacements, like *«приводить ⇔ являться следствием»* (in Russian), will have the higher values of IDF.

The first step (see *slide 8*) is the calculation of TF-IDF for all words of initial phrase concerning each document in corpus. Each of sequences found here will be sorted descending with splitting into clusters by means of algorithm close to FOREL class taxonomy algorithms. As the mass center of cluster the arithmetic mean of all its elements is taken. For revelation of links the most significant words are related to the first and «middle» clusters of such sequence. To the first cluster will be related the terms which are the most unique in analyzed document. TF-IDF values from the «middle» cluster will be corresponded to terms, which have synonyms at the same document, and to general vocabulary defining the synonymic paraphrases. The estimation of coupling strength for pair of words from initial phrase will be calculated here only if the value of TF-IDF at least for one word of this pair related to either first or «middle» cluster. Let's name further such words as pairwise related by TF-IDF.

The idea of *n*-gram's revelation on a sequence of pairs of initial phrase's words related in depending of method of links revelation either syntactically or by TF-IDF, is represented by *Definition 2* on the *slide 9*. The significance of *n*-gram for document ranking (see *formula (4)* on the *slide 9*) can be defined from geometrical considerations and assumes the maximization of sum value for coupling strength of words in its content at minimum of root-mean-square deviation of mentioned value relatively to all links of words in *n*-gram. Herewith according to agreement assumed by us the links are not nec-

essary cover words exclusively within the same phrase: an acceptable are be links of words from different phrases in a group of initial mutually equivalent or complementary in sense and related to the same image. The rank of the document (according to the *formula (5)* on the *slide 10*) here will be the higher the greater number of *n*-grams from revealed in initial phrase were found in the phrases of analyzed document at the highest possible sum value of coupling strength of words in *n*-gram from one side, and at maximal length of *n*-gram – from the other. Using this estimation we can select those corpus documents in which the constituents of image of initial phrase in *n*-grams are represented most fully. Herewith the documents will be sorted descending values of rank with further clustering by means of the same algorithm that was used for splitting of words of initial phrase according to TF-IDF values. The phrases to annotation will be selected from documents related to the cluster of greatest values of ranking function. Let's name further those documents as the best in *n*-grams. Similarly to documents, according to the values of significance for document ranking the *n*-grams are clustered concerning each of documents related to cluster of the greatest values of ranking function. On the final stage the set of phrases from documents of mentioned cluster is clustered using the same algorithm according to the number of words (or, as a variant, of bigrams) in the most significant *n*-grams. Annotation phrases here form the cluster of greatest values of given estimation.

Let's note, that *n*-gram's revelation by means of offered method allows to estimate the relevance of text corpus to knowledge unit defined by initial phrase or their set using the coverage degree of words of initial phrases by the most significant *n*-grams concerning the documents which are the best in *n*-grams (see *slide 11*).

The experimental material to test the proposed method was selected according to criteria represented on the *slide 12*. It was prepared two variants of text corpus in Russian and, correspondingly, two groups of Russian initial phrases for these variants. The first variant of corpus is presented on the *slide 13*, the initial phrases for it are shown on the *slide 14*. The second variant is presented on the *slides 15* and *16*, the initial phrases for it are shown on the *slide 17*.

The software implementation (in Java) of the offered method and experimental results are presented on the website of Yaroslav-the-Wise Novgorod State University.

Slides 19–23 are show the example of revelation of constituents of images for phrase groups presented on the *slide 18* and formed basing on initial phrases shown on the *slides 14* and *17*. The first group includes the *phrase No.1* from presented on the *slide 17* (together with the synonymic paraphrase), results for which were quite satisfactory both for classifying of its words according to TF-IDF and basing on syntactical relations within bigrams. For the phrases from another two groups shown on the *slide 14* only single results for mentioned experiments were satisfactory. However the phrases within these groups are mutually complementary in sense, what is important for the assumptions about relating them to the same image. For comparison for each considering group of initial phrases the *slides 19* and *22* show the total number of selected phrases from corpus documents (N), the number of selected phrases representing linguistic expressional means (N_1), synonyms (N_2) and concept relations at the topical area (N_3). In order for a comprehensive assessment of the retrieval effectiveness the mentioned data are complemented by the number of linguistic expressional means (N_1^1), number of synonyms (N_2^1) and concept relations from mentioned in initial phrases which were represented in resulted phrases (N_3^1).

As can be seen from experimental results represented on the *slides 20–22* for the same phrases but taken separately, the introduction into consideration the group of initial phrases mutually equivalent or complementary in sense together with *n*-grams allows in a row of cases to describe more precisely the image revealed in texts in a form of combinations of words related by sense.

A good confirmation of this thesis is the result for the *phrase group No.3* on the *slide 18*, where according the number of words within the most significant *n*-grams revealed without attraction of the base of syntactic rules the phrase represented in the top part of the *slide 23* and defines the concept of *heuristics* was selected. The given phrase here is a single selected to annotation, herewith from the words of the most significant *n*-grams the phrase contains the Russian words *эвристика*, *в*, *задача*, *на*, *способ*, *решение*, *мочь (heuristics, in, task, on, method, solving, can)*. Simultaneous presence of these words in selected phrase allows to relate the concepts of *heuristics* and *knowledge* mentioned in initial phrases with the *methods of solving tasks* and to release the variant of Russian expressional means *«в результате ⇔ как результат»* together with synonymic replacements *«способ ⇔ приём»*, *«опираться ⇔ основываться»* and *«практический ⇔ прикладной»*.

Let's note, that the definition of *heuristics* being alternative to the first phrase of *group No.3* on the *slide 18* and represented in the bottom part of the *slide 23* was also among phrases most relevant to *initial phrase No.6* on the *slide 14* according to the number of links of words by TF-IDF related to «most strong» by estimation (1) from the *slide 5* at maximal sum of this estimation for all links found in initial phrase. From the «most strong» word pairs which became a basis of phrase selection here contains only *«искусственный интеллект»* that decreases significantly the precision of revelation of constituents of image of initial phrase. Actually the found phrase only relates the concept of *artificial intelligence* mentioned in initial phrase with the concept of *heuristics*.

As an illustrative example of advantages of search of constituents of image of initial phrase on the basis of *n*-grams jointly with revelation of links of words based on TF-IDF can be experiments with the *phrase No.3* from the *slide 14*. It's necessary to note, that for this phrase a satisfactory decision was not found by means of contextual annotation applying the «most strong» links of words. The best here were results for *n*-grams revealed on the links of words related by TF-IDF with the selection of phrases to annotation by the number of words in the most significant *n*-grams, where to the number of four output phrases the phrase represented in the top part of the *slide 24* was related. Associating the conception of *knowledge* from initial phrase with the *knowledge model*, the given phrase allows to construct Russian paraphrases like *«определяется как ⇔ понимается как»* by means of pronominal adverb «*как» (as)*. The same phrase also was among the resulted in experiment with the phrase selection on the base of the «most strong» links by TF-IDF of initial phrase's words. Herewith in addition to the results for search by *n*-grams it was succeeded to reveal a number of conceptual relationships from initial phrase (primarily for the concept of *information*) with other concepts of the same topical area. Note (see *slide 25*), that at greater relevance of text corpus herewith we have the best result of search the constituents of image of initial phrase with application of *n*-grams.

The precision of revelation of constituents of image of initial phrase can be clearly estimated by word-by-word comparison of the most significant links and *n*-grams relatively to documents from the number of the most relevant simultaneously

by *n*-grams and by the number of the «most strong» links at maximal sum of coupling strength for all links found in initial phrase.

So, for experiment, which results are shown on the *slides 26* and *27*, the mentioned links and *n*-grams are fully matched word-by-word for *initial phrase No.1* from represented on the *slide 14*. At the same time for the *phrases No.1*, *7* and *9* from the shown on the *slide 17* the documents simultaneously relevant by two above-mentioned criterions were not found. But experiment presented on the *slides 28* and *29* gave fully coincidence of vocabulary in considered bigrams and n-grams simultaneously on the *phrase No.4* from shown on the *slide 14* and on the *phrase No.1* from shown on the *slide 17*. Nevertheless, the documents simultaneously relevant to initial phrase by *n*-grams and «most strong» links were not found there for more *phrases*: *No.3*, *6*, *7* and *No.9* on the *slide 14*, and also for the *phrases No.7* and *8* on the *slide 17*.

As can be seen from example, the introduction into consideration of links of words by TF-IDF as an alternative to syntax in considered document ranking allows to respect terms more, what is important for topical areas where the percentage of general vocabulary and terms are comparable.

It's necessary to note, that unlike the proposed method, the search of phrases close to initial in described knowledge fragment on a ready syntactical marked text corpus covered all given natural language, requires pre-revelation by expert the words and their combinations representing terms of topical area in initial phrase. As an example on the *slide 30* for initial phrases from the *slides 14* and *17* the words and their combinations belonged at least to one phrase from documents of Russian National Corpus are shown. As can be seen from the table on the *slide 31*, actually the found phrases does not concern the synonymy at the expressed conceptual relations. Moreover, the effectiveness of search here depends from representation of corresponding thematic in texts of corpus.

Thus, along with the decision of its main task, being applied together with the selection of phrases on the basis of the «most strong» links of words of initial phrase, the method offered in current work allows to automate the revelation by expert the required words and their combinations to organize the search a non-fiction texts of given thematic in a syntactically marked text corpus. Moreover, itself the text selection to topical corpus on the base of ranking with application of *n*-grams additionally to the most significant word links allows to precisely define its topic by a complex of special terms co-occurred in text documents. Herewith the output of phrases which are not relevant to initial ones neither at the described knowledge fragment nor at its linguistic expressional forms can be reduced, on average, by 17 times.

The topics for separate consideration are the speed and precision of morphological analysis which is used for revelation of links of words. Here, in particular, of interest is the *Python*-implementation of the offered method with attraction of *NLTK (Natural Language Toolkit)* library and applying the morphological analyzer *Pymorphy* as an alternative to solution implemented in current work and based on *Russian morphology framework*.