The offered work is devoted to the interrelated problems (see *slide 2*) of *completeness of knowledge extraction* from a set of subject-oriented texts (the so-called corpus) by analyzing the relevance to the initial phrase and *finding the most rational linguistic variant* to express the revealed knowledge fragment. 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 phrases equivalent-by-sense (i.e. semantically equivalent, SE). Herewith *the optimal meaning transfer* is provided by those phrases from initial set of equivalent-by-sense which are *of minimal character length* under *a maximum of words most frequently used* in all initial phrases (with the respect of possible synonyms). Just such phrases represent a sense standard (see *slide 3*), which is defined by the set of textual units and their links necessary and enough for representation of knowledge unit.

It is necessary to note, however, that *the precision of sense standard's revelation* on the set of semantically equivalent phrases herewith is essentially dependent on completeness of description of linguistic expressional forms given by expert for knowledge unit. One of the variants to increase the precision of description for expert knowledge fragment extracted from corpus texts is the introduction into consideration the group of initial phrases mutually equivalent or complimentary in sense and related to the same image. The most effective implementation of this solution assumes to include to the mentioned group the phrases of formed annotation which are equivalent or complementary in sense to initial phrases from the point of view of expert.

Another «bottleneck» of the process of sense standard's revelation is the parsing (full or partial) of initial SE-phrases with the purpose of search the most significant links and calculation the distance statistics for linked words within separate phrases. Herewith primarily estimated are links for pairs of words taking into account the occurrence of each word in analyzed set of SE-phrases both separately and in a pairs with another words. The complete syntactic analysis with a construction of dependency tree assumes a large size of statistical model (if the used parser is based on machine learning as MaltParser, for example) and, as a result, the significant resource intensity at the significantly more percent of errors, than, for example, in a case of shallow parsing with conditional random fields. It is necessary to note, that the mentioned shallow parsing is sensitive to presence of prepositions and conjunctions within chunks what restricts its application in the considered problem for analysis of linguistic expressional means for constructing the paraphrases within a given set of SE-phrases. Another *problem of syntax analyzers* today which eventually impacts to the accuracy of parsing is related to the compatibility of morphological characteristics and their tags used by different programs implemented morphological analysis. This information is required for correctly establishment the relationships of words, but one-to-one correspondence does not always take place here.

The goal of the current paper is to find a compromise between the accuracy of revelation of word relationships most significant for linguistic representation of knowledge unit and the number of initial SE-forms of its description by expert.

For solving the given range of problems the current paper investigates the possibilities of revelation of constituents of image of initial phrase by combined use of estimation of coupling strength of its word combinations occurred in the phrases of analyzed text (including in the *n*-grams) and classifying these words according to their TF-IDF values

relatively to corpus texts. 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 4*, 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 5*) 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.

As an estimation of «coupling strength» of words in current work the *estimation (3) 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. 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.

The first step (see *slide 6*) 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. Further concerning to clustering of any objects in current paper we'll have in mind this algorithm. 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.

Let's consider the variant of how to use *n*-grams to solve the problem of completeness of description of unit of expert knowledge extracted from phrases of subject-oriented text set. 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 7*. The significance of *n*-gram for document ranking (see *formula (4)* on the *slide 7*) 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 necessary 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 8*) 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 maxi-

mal 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. 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.

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 correspond to the cluster of greatest values of ranking function (see *formula (6)* on the *slide 8*).

The experimental material to test the proposed method is represented on the *slides 9–14*. The software implementation (in Java) of the offered solutions is presented on the website of Yaroslav-the-Wise Novgorod State University. The main criterion for selection of phrases to groups shown on the *slide 14* was the mutually complementary in sense.

As can be seen from experimental results represented on the *slide 15*, 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 corresponding table rows are highlighted in green). The result for the *phrase group No.3* on the *slide 14* allows us to make an important practical conclusion about the possibility of iterative purposeful selection of phrases equivalent to initial ones or complementary them in sense, accompanied by an increase in relevance estimation for text corpus (see *formula (6)* on the *slide 8*) on each iteration. Herewith on the current step from formed annotation the expert selects a phrase maximally relevant to the set of initial ones and adds new phrase to initial ones with comparing the estimation for current and next iterations. Its decrease indicates that the search is complete, and the resulted set of phrase defining the considering knowledge unit will be consists of the initial phrases of previous iteration. For more accurate revelation of context for terms on the set of natural language forms of knowledge unit representation the word links within *n*-grams here should be considered without taking into account of prepositions and conjunctions.

An example of iterative adding annotation phrases to the set of initial ones is represented on the *slide 16*. More predictable changing of relevance in a case without taking into account of prepositions and conjunctions is conditioned by a greater specific share of terms within revealed *n*-grams. Thus, the estimation of coupling strength of words without taking into account of prepositions and conjunctions allows find all significant conceptual relationships for knowledge unit concerning the set of subject-oriented texts. The separate problem is to estimate the affinity to the sense standard the contexts of general vocabulary within linguistic expressional means for given knowledge fragment relatively to separate phrases (see *slide 17*).

Meaningfully, the sense standard is defined by those SE-phrases from the set of describing the knowledge unit which are of minimal character length under a maximum of words most frequently used in different phrases of mentioned set (with the respect of possible synonyms). The basic empirical considerations concerning the numerical estimations for affinity of phrase to sense standard are represented on the *slide 17*. Herewith for more accuracy of revelation of contexts for general vocabulary the estimation of coupling strength of words should be calculated concerning not separate texts,

but all considered topical text set (corpus). Respecting the requirement of minimization of phrase length, actual here is to consider only those links that are syntactical in nature.

The estimation for affinity of phrase to sense standard basing on TF-IDF metrics may be constructed from the following empirical considerations. First of all, the division of words of initial phrase into general vocabulary and terms according to their TF-IDF should be expressed here as much as possible. Besides the first and «middle» clusters of sequence formed for the initial phrase on the base of TF-IDF values for its words, the meaningful interest herewith also represents the last cluster, to which the terms prevailing in corpus are correspond. The told allows make a conclusion about the sums of TF-IDF values for words of three mentioned clusters as a base for estimation represented by the *formula (7)* on the *slide 18* for the phrase affinity to standard.

Another important moment is that in clusters were formed for words of initial phrase according to their TF-IDF relatively to some corpus document the words must be distributed more or less evenly. This requirement can be represented as the maximization of value of *estimation (8)* on the *slide 18*. The affinity to standard for initial phrase on the base of *estimations (7)* and *(8)* from geometrical considerations can be represented numerically as an area of rectangle with the sides equal to the found values for mentioned estimations concerning given document. Further corpus documents are sorted descending values of product of *estimations (7)* and *(8)*. For mutual estimation of affinity to standard for separate phrases in a group of initial ones mutual equivalent or complementary in sense, for each phrase the pair of mentioned estimations is taken for the document with a greatest value of their product. Further in mentioned pair the *estimations (7)* and *(8)* are divided by their maximums over all phrases of the group of initial ones and transform to [0, 1] range. The initial phrases themselves are sorted descending the values of product of normalized *estimations (7)* and *(8)* with subsequent clustering.

For estimating the phrase affinity to sense standard by analysis of links of its words let's modify the *estimation (3)* shown on the *slide 5* as follows (see *slide 19*). First of all, let's enter the requirement that analyzed words must be linked syntactically. Secondly, the values used in numerator and denominator of *formula (3)* will be calculated relatively to all corpus. Thirdly, the estimation itself will be calculated only when the words *are pairwise related by TF-IDF*. Thus, in relation to the certain document of corpus the offered modified variant of estimation for coupling strength of words depends exclusively on TF-IDF values of words of analyzed pair. In order to analyze the affinity to standard for linguistic expressional means for constructing the paraphrases from employed in initial phrases let's use the two variants of applying the estimation: with and without taking into account of prepositions and conjunctions.

The estimation itself for affinity to the sense standard represented by *formula (9)* on the *slide 19* assumes for the phrases which are maximal close to standard concerning the given document, the maximum number of links with the greatest values of «strength» at maximal sum value of «strength» estimation for all links found in phrase. Similarly to the estimations based on TF-IDF metrics, for each initial phrase the maximal value of *estimation (9)* concerning all corpus documents is taken with the following dividing by its maximum over all phrases of the group of initial ones and transforms to [0, 1] range. The initial phrases themselves are sorted descending the values of normalized *estimation (9)* with subsequent clustering.

Relating a phrase to the cluster of greatest values of product of normalized *estimations (7)* and *(8)* or normalized *estimation (9)* is the necessary but not enough condition to make a conclusion about the relating to « standard». The more accurate ranking requires

the analysis of divergence of considering estimations at relating the same phrase to the clusters of greatest/least values.

To solve the mentioned problem the root-mean-square deviation (RMSD), difference and quotient of greatest and least value are entered into consideration concerning the triple of values of: product of normalized *estimations (7)* and *(8)*, normalized *estimation (9)* with and without taking into account of prepositions and conjunctions for separate phrase. Let's name them further as the RMSD-estimations. Herewith the «Non-Standard» phrases are defined on the base of introduced RMSD-estimations according to the rules represented on the *slide 20*. Eventually the sense standard will be defined by the phrases related to the clusters of greatest values of product of normalized *estimations (7)* and *(8)*, and also of the normalized *estimation (9)*, herewith if the phrase satisfies one of the abovementioned rules that it must be excluded from consideration.

As an example let's consider the set of phrases represented on the *slide 21* and equivalent by sense (from the point of view of expert) to the phrases No.1–4 on the *slide 13*. As can be seen from the results shown on the *slides 22* and *23*, to the set of defined the sense standard here will be related the phrases No.2–6 and No.8–10 from represented *on this slide*, what is *quite comparable in accurace* with the previously proposed decision based on considering of all possible paraphrases of initial phrase.

The result for offered method of revelation of sense standard was quite predictable also for phrases not exactly equivalent but mutually complimentary in sense (see *slide 24*) concerning the topical area where the percentage of general vocabulary and terms in texts are comparable.

As can be seen from the results on the *slide 25*, to the set defining the sense standard for the mentioned set of phrases should be referred the phrase No.3. Indeed, concerning the considering group of Russian phrases the given phrase reflects a maximum of concepts and their relationships at minimum of used general vocabulary. Furthermore, the mentioned phrase *concretizes the representation* about *interpretation* for constituents of *frame* (i.e. «*выражение, входящее во фрейм, … знак в нём*» (*expression being the part of frame, … sign in it*)) introduced by the phrase No.2 of the same group via the *software component of computational (informational) system* which performs here as an instrument for *meaning* transfer by *a human on a corresponding language of knowledge representation*.

The reduction in the textual information necessary to represent the knowledge unit by the set of phrases mutually equivalent or complimentary in sense at entering into consideration the sense standard can be estimated by ratio represented in the bottom of *slide 25*. So, for considered examples we have at least a two-fold reduction of mentioned information volume.

The main *result* of current work is the *method for estimation of affinity to sense standard for natural-language phrase relatively to the knowledge unit represented by it*. The evident advantage of the offered method is that there is no need to describe as many equivalent-by-sense expressional forms as possible for corresponding knowledge unit in a given natural language. But from another side, the results given by the proposed solutions significantly depend from the selection of texts into corpus by expert. This takes into account the level of complexity of the text selected to corpus, and its significance in solved task (for example, from the point of view of topic modeling). In this respect it is of interest to study the dynamics of changing the estimation of coupling strength of words concerning the different corpus documents for syntactically related words of initial phrase.