Представляемая работа посвящена взаимосвязанным проблемам (*плакат 2*) *полноты выделения знаний* из множества (корпуса) тематических текстов анализом релевантности исходной фразе и *поиска наиболее рационального языкового варианта* описания выделяемого фрагмента знаний. Данные проблемы *актуальны* для построения систем обработки, анализа, оценивания и понимания информации, в частности, тестирования знаний на основе открытых тестов. Естественным источником знаний при этом будут публикации ведущих научных школ по соответствующей проблематике. Конечной *практической целью* здесь является поиск наиболее рационального варианта передачи смысла в единице знаний, определяемой множеством семантически эквивалентных (СЭ) фраз предметно-ограниченного естественного языка (ЕЯ). *Оптимальная передача смысла* при этом обеспечивается теми фразами, которые *при минимальной символьной длине* имеют *максимум слов, наиболее употребимых* во всех СЭ-фразах заданного множества (с учётом возможных синонимов). Именно такие фразы представляют смысловой эталон (*плакат 3*), которому отвечает набор единиц текста и их связей, необходимый и достаточный для представления и передачи единицы знаний.

Следует отметить, однако, что *точность выделения смыслового эталона* на множестве СЭ-фраз при этом в значительной степени *зависит* от полноты описания экспертом ЕЯ-форм выражения соответствующей единицы знаний. Один из вариантов повышения точности описания выделяемого в текстах корпуса фрагмента экспертного знания – использование совокупности исходных фраз, взаимно эквивалентных либо дополняющих друг друга по смыслу и представляющих единый образ. Максимально эффективное использование данного решения предполагает включение в число исходных фраз из формируемой аннотации, эквивалентных исходным либо дополняющих их по смыслу с точки зрения эксперта.

Ещё одним «узким местом» процесса выделения смыслового эталона является синтаксический разбор исходных СЭ-фраз с целью определения наиболее значимых связей слов и статистики расстояний между словами в составе связи в рамках отдельных фраз. При этом в первую очередь оцениваются связи пар слов, принимается во внимание встречаемость каждого слова в анализируемом множестве СЭ-фраз как по отдельности, так и в составе паросочетаний. Полный синтаксический разбор с построением дерева зависимостей здесь предполагает большой объём статистической модели (если анализатор основан на машинном обучении, пример – MaltParser) и, как следствие, существенную ресурсоёмкость при значительно большем числе ошибок, чем, допустим, в случае частичного синтаксического разбора с помощью условных случайных полей. Последний, следует отметить, критичен к наличию внутри связанных групп соседних слов предлогов и союзов, что ограничивает его применение в рассматриваемой задаче для анализа языковых выразительных средств конструирования перифраз в рамках заданного множества СЭ-фраз. Другая *проблема* современных *синтаксических анализаторов*, в конечном итоге влияющая на точность разбора, связана с различиями в используемых морфологических характеристиках слов и формата пометок (тегов) для указанных характеристик разными программами морфологического анализа. Данная информация требуется для правильного установления связей слов, а взаимно-однозначное соответствие здесь имеет место не всегда.

Цель настоящей работы – найти компромисс между точностью выделения связей слов, наиболее значимых для языкового представления единицы знаний и числом исходных СЭ-форм её описания экспертом.

Для решения данного круга проблем в настоящей работе исследуются возможности механизма выделения составляющих образа исходной фразы совместным использованием оценки силы связи сочетаний её слов, встречающихся во фразах анализируемого текста (в том числе в составе *n*-грамм), и разбиения этих слов на классы по значению меры TF-IDF относительно текстов корпуса. В задачах анализа текстов и информационного поиска TF-IDF есть статистическая мера, используемая для оценки важности слова в контексте документа, входящего в некоторый текстовый корпус. Согласно классическому определению (*плакат 4*), данная мера есть произведение TF-меры (отношения числа вхождений некоторого слова к общему числу слов документа) и инверсии частоты встречаемости слова в документах корпуса (IDF). Следует отметить (*плакат 5*), что чем чаще слово встречается в документах корпуса, тем ближе к нулю будет для него значение меры IDF. Это относится как к словам общей лексики (глаголы-связки, служебные части речи), так и к словам-терминам, преобладающим в корпусе. В то же время, к примеру, слова из общей лексики, задающие конверсивные замены (*«приводить ⇔ являться следствием»*) будут иметь более высокие значения меры IDF.

В качестве оценки «силы» связи слов в настоящей работе берётся представленная на *плакате 5 оценка (3)*, содержательно близкая коэффициенту Танимото. Из оценок силы связи слов в дистрибутивно-статистическом методе построения тезаурусов данная оценка наиболее наглядна, но в то же время учитывает встречаемость каждого слова в отдельности. За основу выделения самих связей в настоящей работе наряду с синтаксическими зависимостями в качестве альтернативы берётся разбиение слов исходной фразы по значению меры TF-IDF.

Первым шагом (*плакат 6*) относительно каждого документа корпуса вычисляются значения меры TF-IDF для всех слов исходной фразы. Каждая из полученных при этом последовательностей сортируется по убыванию с последующим разбиением на кластеры алгоритмом, содержательно близким алгоритмам класса FOREL. Далее применительно к разбиению последовательности на кластеры будем подразумевать именно этот алгоритм. В качестве центра масс кластера здесь берётся среднее арифметическое всех его элементов. Для выделения связей здесь важны слова первого (термины из исходной фразы, наиболее уникальные для анализируемого текстового документа) и «серединного» (общая лексика, обеспечивающая синонимические перифразы, и термины-синонимы) кластеров последовательности, сформированной для исходной фразы на основе TF-IDF её слов. При этом оценка силы связи для пары слов исходной фразы вычисляется только в том случае, если значение TF-IDF минимум одного из слов пары принадлежит либо первому, либо «серединному» кластеру. Назовём далее такие слова связанными в паре по TF-IDF.

Рассмотрим вариант использования *n*-грамм для решения проблемы полноты описания единицы экспертного знания, выделяемого во фразах множества тематических текстов. Порядок выделения *n*-граммы на последовательности пар слов исходной фразы, связанных в зависимости от метода выделения связей синтаксически либо по TF-IDF, представлен *Определением 2* на *плакате 7*. Значимость *n*-граммы для ранжирования документов (*формула (4)* на *плакате 7*) оценивается из геометрических соображений и подразумевает максимизацию суммы силы связи слов в её составе при минимуме среднеквадратического отклонения указанной величины по всем связям слов в составе *n*-граммы. При этом в соответствии с принятым нами соглашением связи не обязательно охватывают слова исключительно внутри одной фразы: допускаются связи слов из различных фраз в группе исход-

ных, взаимно эквивалентных либо дополняющих друг друга по смыслу и представляющих единый образ. Ранг документа (*формула (5)* на *плакате 8*) будет тем выше, чем большее число *n*-грамм из исходной фразы найдено во фразах анализируемого документа при максимально возможном значении суммарной силы связи слов в составе *n*-граммы с одной стороны, а с другой стороны – максимуме длины *n*-граммы. Содержательно данная оценка позволяет выделить те документы, в которых составляющие образа исходной фразы в *n*-граммах представлены наиболее полно. При этом документы сортируются по убыванию значения ранга с последующим разбиением на классы тем же самым алгоритмом, который используется для разделения слов исходной фразы по TF-IDF. Аналогично документам, но по оценке значимости для ранжирования, кластеризуются сами *n*-граммы относительно каждого из документов кластера наибольших значений функции ранжирования.

Отметим, что выделение *n*-грамм предложенным методом позволяет оценить релевантность текстового корпуса единице знаний, определяемой исходной фразой (их совокупностью), по степени охвата слов исходных фраз наиболее значимыми *n*-граммами относительно документов, отвечающих кластеру наибольших значений функции ранжирования (*формула (6)* на *плакате 8*).

Экспериментальный материал для апробации метода представлен на *плакатах 9–14*. Программная реализация предложенных решений на языке Java представлена на портале Новгородского университета. Основным критерием при отборе фраз в группы на *плакате 14* была взаимная дополняемость по смыслу.

Как видно из приведённых на *плакате 15* результатов, большей релевантности корпуса отвечает и лучший результат поиска составляющих образа исходной фразы по *n*-граммам (соответствующие строки таблицы выделены зелёным цветом). Результат по группе фраз №3 на *плакате 14* позволяет сделать важный практический вывод о возможности итеративного целенаправленного отбора фраз, эквивалентных исходным либо дополняющих их по смыслу, сопровождаемого ростом оценки релевантности текстового корпуса (*формула (6)* на *плакате 8*) на каждой итерации. При этом на очередном шаге из сформированной аннотации эксперт отбирает фразу, максимально релевантную совокупности исходных фраз, добавляя новую фразу к совокупности исходных и сравнивая оценку по текущей и предшествующей итерации. Её уменьшение говорит о завершении поиска – результирующий набор фраз, определяющих рассматриваемую единицу знаний, будут составлять исходные фразы предыдущей итерации. Для более точного выделения контекста терминов на множестве ЕЯ-форм представления единицы знаний связи слов в составе *n*-грамм здесь следует рассматривать без учёта предлогов и союзов.

Пример итерационного добавления фраз аннотации в число исходных представлен на *плакате 16*. Более предсказуемое изменение релевантности в случае без учёта предлогов и союзов обусловлено большим удельным весом терминов в составе выделяемых *n*-грамм. Таким образом, оценка силы связи слов без учёта предлогов и союзов позволяет найти все значимые связи понятий в рамках единицы знаний относительно множества тематических текстов. Отдельная задача – анализ близости эталону контекстов общей лексики в составе языковых выразительных средств заданного фрагмента знаний по отдельным фразам (*плакат 17*).

Содержательно смысловой эталон задают те СЭ-фразы из множества описывающих единицу знаний, которые при минимуме символьной длины имеют максимум слов, наиболее употребимых в различных фразах указанного множества (с учётом возможных синонимов). Основные эмпирические соображения относитель-

но численных оценок близости фразы эталону представлены на *плакате 17*. При этом для большей точности выделения контекстов общей лексики оценку силы связи слов следует вычислять относительно не отдельных текстов, а всего рассматриваемого корпуса. Учитывая требование минимизации длины фразы, актуальным является рассмотрение только тех связей, которые имеют синтаксическую природу.

Оценка близости фразы смысловому эталону на основе меры TF-IDF строится из следующих эмпирических соображений. Во-первых, разделение на общую лексику и термины здесь должно быть выражено как можно в большей степени. Помимо первого и «серединного» кластеров последовательности, сформированной для исходной фразы на основе TF-IDF её слов, содержательный интерес при этом также представляет последний кластер, которому соответствуют слова-термины, преобладающие в корпусе. Сказанное позволяет сделать вывод о значениях сумм величин TF-IDF для слов трёх указанных кластеров как основы оценки близости фразы эталону, представляемой *формулой (7)* на *плакате 18*.

Другой немаловажный момент – слова в кластерах, сформированных по TF-IDF слов исходной фразы, должны быть распределены более или менее равномерно. Содержательно это можно представить как максимизацию величины *оценки (8)* на *плакате 18*. Значение близости исходной фразы эталону на основе *оценок (7)* и *(8)* из геометрических соображений можно представить как площадь прямоугольника со сторонами, равными найденным значениям указанных оценок относительно заданного документа. Далее документы корпуса сортируются по убыванию произведения *оценок (7)* и *(8)*. Для взаимной оценки близости эталону отдельных фраз в группе исходных, взаимно эквивалентных либо дополняющих друг друга по смыслу, для каждой фразы берётся пара указанных оценок по документу, получившему наибольшее значение их произведения. Далее в указанной паре *оценки (7)* и *(8)* делятся на свои максимумы по всем фразам группы исходных и приводятся к диапазону [0,1]. Сами исходные фразы сортируются по убыванию произведения нормированных *оценок (7)* и *(8)* с последующей разбивкой на кластеры.

Для оценки близости фразы смысловому эталону анализом связей её слов модифицируем *оценку (3)* на *плакате 5* следующим образом (*плакат 19*). Во-первых, введём требование наличия синтаксической связи анализируемых слов. Во-вторых, значения из используемых в числителе и знаменателе *формулы (3)* будем вычислять относительно всего корпуса. В-третьих, сама оценка вычисляется только тогда, когда слова *связаны в паре по TF-IDF*. Таким образом, применительно к конкретному документу в составе корпуса предлагаемый модифицированный вариант оценки силы связи слов зависит исключительно от значения TF-IDF слов анализируемой пары. В целях анализа близости «эталонным» языковых средств конструирования перифраз из задействованных в исходных фразах воспользуемся двумя вариантами применения оценки – с учётом предлогов и союзов и без таковых.

Сама оценка близости эталону, представляемая *формулой (9)* на *плакате 19*, подразумевает для максимально близких эталону фраз относительно заданного документа максимум связей с наибольшими значениями «силы» при максимальной суммарной величине оценки «силы» для всех найденных во фразе связей. Аналогично оценкам на основе меры TF-IDF, для каждой исходной фразы берётся максимальное значение *оценки (9)* по документам корпуса, которое далее приводятся к диапазону [0,1] посредством деления на свой максимум по всем фразам группы исходных. Сами исходные фразы сортируются по убыванию полученного нормированного значения *оценки (9)* с последующим разделением на кластеры.

Отнесение фразы к кластеру наибольших значений произведения нормированных *оценок (7)* и *(8)* либо нормированной *оценки (9)* является необходимым, но не достаточным условием заключения о принадлежности к «эталонным». Более точное ранжирование требует анализа расхождений рассматриваемых оценок при отнесении одной и той же фразы к кластерам наибольших/наименьших значений.

Для решения указанной задачи применительно к тройке значений: произведения нормированных *оценок (7)* и *(8)*, а также нормированной *оценки (9)* с учётом и без учёта предлогов/союзов по отдельной фразе вводится в рассмотрение среднеквадратическое отклонение (СКО), а также разность и частное наибольшего и наименьшего значения (далее – СКО-оценки). При этом «неэталонные» фразы определяются на основе введённых СКО-оценок в соответствии с правилами, представленными на *плакате 20*. В конечном итоге определять эталон будут фразы из числа отнесённых к кластерам наибольших значений величин произведения нормированных *оценок (7)* и *(8)*, а также нормированной *оценки (9)*, при этом из рассмотрения исключаются попадающие под одно из вышеуказанных правил.

В качестве примера рассмотрим представленное на *плакате 21* множество фраз, эквивалентных по смыслу (с точки зрения эксперта) фразам №1–4 на *плакате 13*. Как видно из результатов на *плакатах 22* и *23*, к числу определяющих смысловой эталон здесь будут отнесены фразы №2–6, 8–10 из представленных на этом плакате, что вполне сопоставимо по точности с ранее предложенным решением на основе всех возможных перифраз исходной фразы.

Вполне предсказуемый результат работы предлагаемого здесь метода выделения смыслового эталона был получен и для фраз, не в точности эквивалентных, но взаимно дополняющих друг друга по смыслу (*плакат 24*) относительно предметной области, где доля общей лексики сравнима с долей слов-терминов.

Как видно из результатов на *плакате 25*, для указанного множества фраз к числу определяющих смысловой эталон следует отнести фразу №3. Действительно, по рассматриваемой группе данная фраза характеризуется максимумом числа отражаемых понятий и их связей при минимуме используемой общей лексики. Более того, в указанной фразе *конкретизируется представление* об *интерпретации* составляющих *фрейма* (*выражение*, *входящее во фрейм*, … *знак в нём*), вводимое фразой №2 из той же группы, через *программную часть вычислительной (информационной) системы*, которая здесь выступает как инструмент передачи *человеком смысла на соответствующем языке представления знаний*.

Сокращение текстовой информации, необходимой для представления единицы знаний множеством взаимно эквивалентных либо дополняющих друг друга по смыслу фраз, получаемое при введении смыслового эталона, можно оценить посредством представленного в нижней части *плаката 25* соотношения. Так, для рассмотренных примеров указанный объём информации сокращается минимум вдвое.

Основной *результат* данной работы *– метод оценки близости ЕЯ-фразы смысловому эталону относительно представляемой ей единицы знаний*. Преимущество предложенного метода – отсутствие необходимости описания как можно большего числа СЭ-форм выражения соответствующей единицы знаний в языке. С другой стороны, результаты работы метода существенно зависят от подбора корпуса экспертом. Здесь учитывается и уровень сложности текста, и его значимость в решаемой задаче (например, с позиции тематического моделирования). В этом плане представляет интерес исследование динамики изменения оценки силы связи по документам корпуса для синтаксически связанных слов исходной фразы.