Представляемая работа посвящена взаимосвязанным проблемам (*плакат 2*) выделения единиц знаний из множества (корпуса) тематических текстов, отбора текстов в корпус анализом релевантности исходной фразе и полноты отражения в исходных фразах выделяемого фактического знания. 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 отечественных и зарубежных 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). При этом в круг задач эксперта, требующих автоматизации, входит (*плакат 3*):

- поиск СЭ-форм выражения отдельного фрагмента фактического знания в заданном ЕЯ;
- сопоставление фрагментов собственных знаний эксперта с наиболее близкими фрагментами знаний других экспертов.

Следует отметить, однако, что значимость текста здесь, как правило, безотносительна к образу, представляемому исходной фразой и выделяемому в анализируемых текстах. Требования же к соотношениям составляющих выделяемого в тексте образа можно сформулировать следующим образом (*плакат 4*):

- фрагмент анализируемого текста, отвечающий составляющей образа, отождествим с некоторой смысловой связью слов в исходной фразе;
- сила связи слов каждого такого фрагмента всегда больше силы связи любого слова данного фрагмента и слова, не принадлежащего ему;
- слабосвязанные слова исходной фразы не могут отождествляться (по определению) с одним фрагментом. Очевидно, что сочетания общей лексики и терминов исходной фразы, преобладающих в корпусе, в анализируемом тексте можно отнести к составляющим искомого образа только по присутствию фрагментов с большей силой связи слов;
- допускаются связи слов из фраз в группе исходных, взаимно эквивалентных либо дополняющих друг друга по смыслу и представляющих единый образ.

Кроме того, в общем случае не выдвигается требование наличия в тексте строго заданной части составляющих образа исходной фразы (ОИФ). Корректное выделение этого образа предполагает исследование встречаемости и отдельных слов, и их сочетаний с оценкой «силы» связи слов относительно текста и корпуса. Сама же исходная фраза лишь в единичных случаях соответствует эталону для сопоставления. При этом число исходных фраз в ряде случаев целесообразно увеличивать до двух и более для более точного описания представляемого фрагмента знаний (понятий и их связей). Кроме того, ограничение рассмотрения связей слов биграммами и рамками синтаксиса ЕЯ здесь критично для случаев, когда доля общей лексики сравнима с долей слов-терминов (например, в текстах по близким искусственному интеллекту разделам философии науки и техники).

Для решения данного круга проблем в настоящей работе вводятся в рассмотрение *n*-граммы на последовательностях пар слов, связанных либо синтаксически, либо по смыслу, с одновременным уходом от жёсткой ориентации на синтаксис ЕЯ (*плакат 5*).

«Классические» *n* -граммы (по К. Шеннону – *L*-граммы) как последовательности из *n* элементов нашли широкое применение в математических исследованиях,

биологии, а также информационном поиске. Наиболее близкими рассматриваемой в настоящей работе проблематике являются синтаксические *n*-граммы, которые определяются не линейной структурой текста, а путями в деревьях синтаксических зависимостей либо деревьях составляющих. Отметим, что при использовании силы связи слов исходной фразы относительно текста как основы оценки его релевантности последней такие пути следует отсчитывать не от вершины дерева, а от сочетаний слов с наибольшими значениями силы связи. В отличие от поиска синтаксически связанных групп соседних слов с помощью условных случайных полей, наличие внутри связанных фрагментов текста предлогов и союзов здесь не является критичным, что немаловажно для поиска в текстах языковых выразительных средств конструирования перифраз исходной фразы.

В качестве оценки «силы» связи слов в настоящей работе берётся представленная на *плакате 5 оценка (1)*, 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. Сам же метод содержательно близок рассматриваемой задаче выделения в анализируемых текстах образа, представляемого исходной фразой. Основная гипотеза метода заключается в наличии некоторой связи между словами, совместно встречающимися в пределах некоторого текстового интервала, в частности, в пределах одной фразы. При этом каких-либо ограничений на применяемые оценки совместной встречаемости слов не накладывается.

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. 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 6*), 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 7*), 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.

Первым шагом (*плакат 8*) 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. 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.

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 9*. Значимость *n*-граммы для ранжирования документов (*формула (4)* на *плакате 9*) 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 исходных, взаимно эквивалентных либо дополняющих друг друга по смыслу и представляющих единый образ. Ранг документа (*формула (5)* на *плакате 10*) здесь будет тем выше, чем большее число *n*-грамм из выделенных в исходной фразе 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 исходного текстового множества, 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. Отбор фраз в аннотацию производится из документов, отвечающих кластеру наибольших значений функции ранжирования (далее – документов, лучших по *n*-граммам). Аналогично документам, но по оценке значимости для ранжирования, кластеризуются сами *n*-граммы относительно каждого из документов кластера наибольших значений функции ранжирования. На заключительном этапе множество фраз документов указанного кластера группируется тем же самым методом по числу слов (либо по числу биграмм) в составе наиболее значимых *n*-грамм, а в аннотацию отбираются фразы кластера наибольших значений указанной оценки.

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, лучших по *n*-граммам (*плакат 11*).

Экспериментальный материал для апробации предложенного метода подбирался в соответствии с критериями, представленными на *плакате 12*. Были подготовлены два варианта исходного текстового множества и, соответственно, две группы исходных фраз. Состав первого варианта представлен на *плакате 13*, исходные фразы для него – на *плакате 14*. Второй вариант приведён на *плакатах 15* и *16*, исходные фразы – на *плакате 17*.

Программная реализация метода на языке Java и результаты экспериментов представлены на портале Новгородского университета.

На *плакатах 19–23* показаны примеры выделения составляющих образов для представленных на *плакате 18* групп исходных фраз из приведённых на *плакатах 14* и *17*. Первая группа включает *фразу №1* из представленных на *плакате 17* (вместе с синонимической перифразой), для которой были получены вполне удовлетворительные результаты как разбиением её слов на классы по значению меры TF-IDF, так и на основе синтаксических связей в рамках биграмм. Фразы двух других групп представлены на *плакате 14*, причём удовлетворительными по данным экспериментам оказались лишь отдельные результаты. Вместе с тем, фразы внутри этих групп взаимно дополняют друг друга по смыслу, что немаловажно для пред-

положения о соответствии им единого образа. Для сравнения на *плакатах 19* и *22* для рассматриваемых групп исходных фраз приведено общее число отобранных фраз (N), в том числе представляющих выразительные средства языка (N_1), синонимы (N_2) и связи понятий предметной области (N_3). В целях более всесторонней оценки результативности поиска здесь вводится в рассмотрение, соответственно, число представляемых в найденных фразах выразительных средств языка (N_1^1), синонимов (N_2^1) и связей для понятий из упомянутых в исходных фразах (N_3^1).

Как видно из представленных на *плакатах 20–22* результатов экспериментов с теми же фразами, но по отдельности, введение в рассмотрение совокупности исходных фраз, взаимно эквивалентных либо дополняющих друг друга по смыслу, совместно с *n*-граммами позволяет в ряде случаев более точно описывать выделяемый в текстах образ в виде сочетаний связанных по смыслу слов.

Хорошим подтверждением данного тезиса является результат для группы фраз №3 на *плакате 18*, где по числу слов в составе наиболее значимых *n*-грамм, выделяемых без привлечения базы синтаксических правил, была отобрана фраза, представленная в верхней части *плаката 23* и дающая определение по*нятия эвристики*. Данная фраза – единственная вошедшая здесь в аннотацию, при этом из слов наиболее значимых *n*-грамм во фразе присутствуют *эвристика*, *в*, *задача*, *на*, *способ*, *решение*, *мочь*. Одновременное наличие этих слов в отбираемой фразе позволяет соотнести понятия *эвристика* и *знание*, упоминаемые в исходных фразах, с *приёмами решения задач*, а также реализовать вариант языковых выразительных средств *«в результате ⇔ как результат»* плюс синонимические замены *«способ ⇔ приём»*, *«опираться ⇔ основываться»* и *«практический ⇔ прикладной»*.

Отметим, что определение *эвристики*, альтернативное первой фразе *группы №3* на *плакате 18* и представленное в нижней части *плаката 23*, было и среди фраз, наиболее релевантных *исходной №6* на *плакате 14* по числу связей слов (по TF-IDF), относимых к «наиболее сильным» согласно оценке (1) на *плакате 5* при максимуме суммы её значений по всем связям слов исходной фразы. Из «наиболее сильных» пар слов, служивших основой отбора фраз, здесь содержится только *«искусственный интеллект»*, что существенно снизило точность выделения составляющих образа исходной фразы. Фактически найденная фраза лишь соотносит понятие *искусственный интеллект* из исходной фразы с понятием *эвристика*.

Преимущества поиска составляющих ОИФ на основе *n*-грамм совместно с выделением связей слов по TF-IDF наиболее наглядно иллюстрируются экспериментами с *фразой №3* на *плакате 14*, для которой контекстно-зависимым аннотированием по «наиболее сильным» связям слов удовлетворительного решения найдено не было. Наилучшими здесь оказались результаты для *n*-грамм на связях слов по TF-IDF с отбором фраз в аннотацию по числу слов в составе наиболее значимых *n*-грамм, где в число четырёх результирующих вошла фраза, представленная в верхней части *плаката 24*. Соотнося представление о *знании* из исходной фразы с *моделью знания*, данная фраза посредством местоимённого наречия «*как*» также позволяет строить перифразы вида *«определяется как ⇔ понимается как»*. Рассматриваемая фраза была в числе результирующих и в эксперименте с отбором фраз на основе «наиболее сильных» связей слов исходной фразы по TF-IDF. В дополнение к результатам аннотирования по *n*-граммам здесь удалось выделить ряд связей понятий из исходной фразы (в первую очередь – для понятия *информация*) с другими понятиями той же предметной области. Отметим (*плакат*

25), что большей релевантности текстового корпуса при этом отвечает и лучший результат поиска составляющих ОИФ по *n*-граммам.

Точность выделения составляющих ОИФ наглядно оценивается пословным сравнением наиболее значимых связей и *n*-грамм относительно документов из числа максимально релевантных одновременно и по *n*-граммам, и по числу «наиболее сильных» связей при максимальном суммарном значении силы для всех найденных в исходной фразе связей.

В эксперименте, результаты которого приведены на *плакатах 26* и *27*, для исходной *фразы №1* из представленных на *плакате 14* указанные связи и *n*-граммы по составу слов полностью совпадают. В то же время для *фраз №1, 7* и *9* из приведённых на *плакате 17* не нашлось документов, релевантных одновременно по двум вышеназванным критериям. Представленный же на *плакатах 28* и *29* эксперимент дал полное совпадение состава слов рассматриваемых биграмм и *n*-грамм одновременно по *фразе №4* на *плакате 14* и *фразе №1* на *плакате 17*. Тем не менее, одновременно релевантных документов по *n*-граммам и «наиболее сильным» связям здесь не нашлось для большего числа *фраз*: *№3, 6, 7, 9* на *плакате 14*, а также *фраз №7* и *8* на *плакате 17*.

Как видно из примера, введение связей слов по TF-IDF как альтернативы синтаксическим отношениям в рассматриваемом ранжировании документов позволяет в большей степени учитывать термины, что немаловажно для предметных областей, где их доля сравнима с долей общей лексики в текстах.

Следует отметить, что в отличие от предложенного метода, поиск фраз, близких исходной по описываемому фрагменту знания, на синтаксически размеченном текстовом корпусе, охватывающем весь заданный ЕЯ, требует предварительного выделения экспертом в исходной фразе слов и их сочетаний, представляющих термины предметной области. В качестве примера на *плакате 30* приведены слова и их сочетания для исходных фраз на *плакатах 14* и *17*, входящие минимум в одну фразу из документов Национального корпуса русского языка. Как видно из таблицы на *плакате 31*, найденные при этом фразы из понятийных связей практически не отражают синонимии. Кроме того, результативность поиска здесь зависит от представленности соответствующей тематики в текстах корпуса.

Таким образом, наряду с решением своей основной задачи, будучи совместно используемым с отбором фраз на основе «наиболее сильных» связей слов исходной фразы, предложенный в настоящей работе метод позволяет автоматизировать выделение экспертом требуемых слов и их сочетаний для организации поиска в синтаксически размеченном текстовом корпусе нехудожественных текстов по заданной тематике. Кроме того, сам отбор текстов в тематический корпус на основе ранжирования по *n*-граммам и наиболее значимым связям слов позволяет точно задать его тему совокупностью специальных терминов предметной области, совместно встречающихся в текстовых документах. При этом в среднем в 17 раз сокращается выход фраз, не релевантных исходной фразе ни по описываемому фрагменту знания, ни по языковым формам его выражения.

Тема отдельного рассмотрения – скорость и точность морфологического анализа, необходимого для выделения связей слов. Здесь, в частности, представляет интерес реализация предложенного в работе метода на языке *Python* с привлечением библиотеки *NLTK (Natural Language Toolkit)* и морфологического анализатора *Pymorphy* как альтернатива реализованному авторами решению *на базе библиотеки русской морфологии*.