Научный отчет по проекту 17-29-09159 офи_м

Значительная часть исследований проекта была направлена на разработку грамматики предложных конструкций современного русского языка, который позволит осуществлять квантитативную оценку ее компонентов применительно к сбалансированным корпусам современных текстов, а также различным жанровым вариациям. Была разработана квантитативная онтология русских предложных конструкций. Сложность построения онтологии связана с тем, что предложные конструкции достаточно часто интерпретируются как нелексические или не вполне лексические языковые элементы, содержащие существенный грамматический компонент значений. М.И.Стеблин-Каменский [Стеблин-Каменский М.И. Спорное в языкознании. Л.: Издательство Ленинградского университета, 1974] указывал на то, что первым существенным противопоставлением грамматических и лексических значений является степень их детализации: лексические значения могут содержать очень большое число конкретных, трудно формулируемых визуальных или функциональных компонентов, которые образуют целые лексико-семантические поля, описывающие все необходимые характеристики объектов и признаков, используемые в номинации конкретной национально-специфической лексической системы языка. В грамматической же системе сходные компоненты отображаются довольно схематично, без развернутой конкретизации. Второе существенное различие лексических и грамматических значений — это неполная осознанность грамматических значений, которая выражается, например, в приписывании одушевленности таким словам, как покойник, мертвец или кукла и тому подобное. Довольно часто грамматические значения вообще слабо соотносятся с какой-либо осмысленной характеристикой, например, значение русских падежей. Такие грамматические категории, объединяющие группы значений, называют формальными. Однако важно, что даже в таких случаях возможно выделение некоторых оппозиций, как это было продемонстрировано Р.Якобсоном [Якобсон Р.О. Избранные работы. М., 1985.] применительно как раз к значениям русских падежей и глагольным категориям. В описании оппозиций Р.Якобсон следовал собственной логике возможности нейтрализации некоторого маркированного грамматического значения, так называемой признаковой категории. То есть оппозиция не противопоставляет признак «А» и «не А», как это обычно трактуют многие исследователи, а признаку «А» противопоставлен признак «не А или А». Для нас это дает довольно четкий принцип выделения грамматических оппозиций на основе корпусной статистики реализации однотипных грамматических значений: их соотношение будет стремиться к пропорции «один» к «полутора», т. е. признаковая категория встречается реже в текстовых корпусах, поскольку это же значение может быть выражено без прояснения вторым членом оппозиции. Описанные Р.Якобсоном оппозиции падежных значений довольно четко реализуются в корпусной статистике, при этом их содержательная трактовка, данная в работах Р.Якобсон, никак не связаны с частотными соотношениями реализации значений. Мы будем использовать идею «признаковых категорий» при анализе значений предложно-падежных конструкций, поскольку чисто логические сопоставления значений невозможны в силу грамматического характера этих конструкций, что приводит к так называемой «непоследовательности» в употреблении предложно-падежных конструкций.

Поскольку невозможно интерпретировать предложно-падежную конструкцию как композиционную («значение предлога» + «значение падежной формы»), мы выбрали подход, предложенный Г.А.Золотовой [Золотова Г.А. Синтаксический словарь: Репертуар элементарных единиц русского синтаксиса. 4-е изд. М.: Едиториал УРСС, 2011.], при котором предложно-падежная конструкция рассматривается как синтаксема —

минимальная грамматическая конструкция, выражающая определенное значение. Синтаксемы могут быть независимыми, т. е. относительно самостоятельными, но чаще зависимыми, когда они присоединяются в виде целых блоков к знаменательным словам, чаще всего глаголам. Количество синтаксем в словаре Г.А.Золотовой составляет несколько десятков. Одна и та же синтаксема может быть выражена разными предлогами в сочетании с разными падежными формами, которые могут обеспечивать синонимичную замену при некоторых глаголах, но в общем случае они не синонимичны.

Для создания квантитативного корпусного описания предложных конструкций нам было необходимо переопределить необходимый и достаточный набор синтаксем как центральное звено грамматической онтологии. Как показали наши пилотные исследования реализации темпоративных и локативных синтаксем [Азарова И.В., Захаров В.П. Корпусное исследование значений русских предложно-падежных конструкций // Структурная и прикладная лингвистика: Межвузовский сборник, вып. 13. СПб.: Издательство Санкт-Петербургского университета, 2019. С. 157–172.], сходные по значению синтаксемы образуют группы, так называемые семантические рубрики, которые составляют верхний уровень оппозиций онтологии. Рубрика, как правило, содержит несколько синтаксем в духе описания Г.А.Золотовой, которые в свою очередь могут подразделяться на подтипы. Таким образом обозначается круг типовых оппозиций, представленных на трех уровнях абстрагирования предложных грамматических значений. Каждый из элементов онтологии на определенном уровне будет задавать структуру возможных маркированных значений, характеризующихся количественными параметрами, которые соотносятся с ранговыми квантитативными позициями в системе противопоставлений на данном уровне онтологии и в заданном кластере значений.

Синтаксемы частотных предлогов явно носят грамматический характер, исходя из соотношения их частот с частотами лексических единиц (2500 против 300 IPM, это число оцениваемых примеров на миллион токенов корпуса). Кроме того, синтаксемы могут быть выражены так называемыми производными предлогами, которые могут иметь гораздо меньшую частотность, при этом они более ясно выражают значение синтаксемы и обладают текстовой вариативностью, характерной для лексических словосочетаний. Например, темпоративная синтаксема выражается при помощи предлога “в” в сочетании с предложным или местным падежом существительных, обозначающих такие интервалы времени, как месяцы или годы: в 1999 году, в августе, но для существительных, обозначающих время суток или дни недели, используется форма винительного: в пятницу, в 10 часов. На базе этого предлога формируется целый ряд производных предлогов: во время войны, в период беременности, в момент опасности, во времена крестовых походов. Вопрос о «грамматичности» таких конструкций наталкивается на ряд соображений за или против, однако как только такая конструкция преодолевает некоторый «порог» частотности, она становится полноправным членом грамматических способов выражения синтаксемы. Результат грамматикализации четко виден на примере некоторых производных предлогов, например, форма творительного падежа "посредством" используется для выражения медиативной синтаксемы с корпусной частотой 19 IPM в сопоставлении с практически минимальной лексической частотой лексемы «посредство» 0,01 IPM. Известно, что заключительная стадия грамматикализации может сопровождаться орфографическим подтверждением: в течение года, несмотря на непогоду.

Опираясь на предложенные выше параметры реализации грамматических значений посредством предложных конструкций, можно представить конфигурацию онтологического ядра предложных конструкций, если рассмотреть типовые значения для частотной группы первообразных предлогов. В эту группу входят 2 суперчастотных

предлога "в" и "на", которые регулярно занимают первую и вторую позицию в частотном списке предлогов, причем как в корпусах общей тематики, так и в функционально-стилистических подкорпусах. К ним примыкает предлог "с", который стабильно занимает третью позицию в частотном списке, но с существенно меньшей частотой. Для прояснения грамматических оппозиций мы взяли еще 7 частотных предлогов из корпусной статистики общей тематики [Ляшевская О.Н., Шаров С.А. Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М.: Азбуковник, 2009], при этом было очевидно, что ранговый порядок предлогов частично зависит от стилистического и/или тематического баланса корпусов современного русского языка.

Чтобы оценить наборы семантических рубрик, реализуемых в группе частотных предлогов, мы использовали идею признаковых категорий как основание для противопоставления грамматических значений в сбалансированном корпусе текстов кафедры математической лингвистики СПбГУ: локализация [8090 IPM], темпоратив [5090 IPM], объектив [3240 IPM], дериватив, т. е. образование производных предлогов и устойчивых выражений [2080 IPM], квалитатив [1160 IPM], партитив [690 IPM], квантитатив [430 IPM]. Долевое соотношение выделенных типов рубрик: локализация [35%],темпоратив [22%], объектив [14%], дериватив [9%], квалитатив [5%], партитив [3%], квантитатив [2%]. Существуют небольшие пересечения между рубриками при их синкретичном выражении: лежать в нескольких метрах (локализация + квантитатив + дериватив / устойчивое выражение).

В процессе работы по описанию предложно-падежных конструкций в терминах синтаксем классификация Г.А.Золотовой была переработана и переориентирована на применение в задачах, связанных с автоматическим семантико-синтаксическим анализом. Чёткая разграниченность сформированных семантических классов, одномерность классификации, уменьшение количества классов способствовали успешному использованию классификации как при ручной разметке, так и в автоматической классификации. В дополнение к Синтаксическому словарю Г.А.Золотовой классификация включает не только первообразные предлоги, но и составные производные предлоги, недостаточно исследованные и статус которых в русской грамматике и лексикографии однозначно не определен.

Для выделения предложных конструкций было решено отказаться от используемого ранее метода контекстного окна (CQL-запросы) в пользу извлечения предложных конструкций на основе структур зависимостей, что позволяет повысить качество выдачи (см. Dirk Hovy, Stephen Tratz, and Eduard Hovy. 2010. What’s in a preposition? Dimensions of sense disambiguation for an interesting word class. In Coling 2010: Posters, pages 454–462, Beijing, China). Для решения задачи был разработан на языке Python инструмент pphrase (https://github.com/merionum/pphrase) для извлечения предложных конструкций на основе зависимостей, получаемых в результате морфосинтаксической разметки корпуса с помощью модели UDPipe (http://ufal.mff.cuni.cz/udpipe). Инструмент позволяет извлекать полную предложную конструкцию вида «управляющее слово — предлог — именная группа зависимого слова» для простых и сложных предлогов, имеющих соответствующую частеречную разметку в используемой модели UDPipe. Несмотря на ориентацию на русские предложные конструкции, инструмент применим для извлечения предложных конструкций из корпусов любого из языков, для которого доступна модель UDPipe. Выдача характеризуется высокими значениями точности и полноты (полнота 0.795, точность 0.771 на тестовом корпусе). Запланировано расширение списка обрабатываемых предлогов, а также фильтрация нерелевантных служебных единиц, включающих в себя предлоги.

С помощью инструмента pphrase был сформирован не имеющий аналогов Банк русских предложных конструкций с семантической разметкой в терминах синтаксем. Обращаясь к нему, пользователь может извлекать информацию о частоте конструкций и их составляющих, значениях конкретного предлога и др.

На основе Банка предложных конструкций с ручной семантической разметкой была произведена автоматическая классификация значений предложных конструкций с использованием алгоритма машинного обучения с учителем (см. Gudkov, V., Golovina, A., Mitrofanova, O., Zakharov, V. Russian prepositional phrase semantic labelling with word embedding-based classifier // (2020) CEUR Workshop Proceedings, 2552, pp. 272-284). Было получено частотное распределение предложных значений в объемном корпусе. Результаты эксперимента показывают, что задача семантической классификации предложных значений может быть с высокой степенью точности решена методами машинного обучения при наличии качественно размеченных обучающих данных в достаточном объёме.

Корпусно-ориентированный конструкционный подход к исследованию предложных структур и их семантики предполагает максимальное приближение к изучению актуальных языковых реалий. Полученные статистические данные указывают, в частности, на широкую распространённость сложных предлогов, до сих пор получающих мало внимания в современных грамматиках русского языка. Планируется более тщательное исследование специфики производных предлогов как объемного и значимого подмножества, изучение которого без корпусов и методов их автоматического анализа невозможно. Достаточно новым и малоизученным остаётся синтаксемный подход к описанию значений предложных конструкций. Отказ от описания предложных значений в терминах предикатно-актантной структуры в пользу синтаксемного подхода способствовал прицельному изучению специфики предлогов как самостоятельного класса единиц, с одной стороны, и предложных конструкций как единств с собственной семантико-синтаксической уникальной структурой — с другой. Адаптация использовавшейся ранее синтаксемной классификации Г.А. Золотовой для задачи автоматической семантической разметки предложных конструкций стала шагом на пути к масштабному исследованию предложной семантики в русском языке. Разработанная классификация отличается от прочих существующих описаний предложных значений системностью описания, небольшим количеством классов, чёткостью их разграничения, широким использованием корпусной статистики, покрытием наиболее частотных случаев, а также направленностью на весь класс единиц предложного типа. Предполагается дальнейшее усовершенствование полученной классификации с целью приведения плоской классификации, принятой для индексации конструкций в Банке предложных конструкции, в соответствие с разработанной предложной онтологией.

Базовый метод описания единиц квантитативной грамматики предложных конструкций представляет собой сбор, группирование и выделение кластеров релевантных единиц методом статистико-комбинаторного анализа в сочетании со смысловой разметкой типов конструкций (синтаксем), лексико-семантических типов главных и зависимых слов в конструкциях на материале корпусов разных типов. Построение онтологии проводилось в несколько этапов.

Прежде всего была создана база данных предложных конструкций, полученная на основе больших корпусов русского языка. Первоначально конструкции извлекались по запросам, представляющим собой шаблоны на языке CQL, опирающиеся на морфологически размеченные корпусы и word-sketch грамматику системы Sketch Engine.

На втором этапе выполнения проекта был разработан инструмент извлечения предложных конструкций pphrase на основе дерева зависимостей, обученный на синтаксически размеченном корпусе SynTagRus с помощью модели UDPipe (http://ufal.mff.cuni.cz/udpipe). Инструмент позволяет извлекать полную предложную конструкцию вида «управляющее слово — предлог — именная группа зависимого слова» для простых и сложных предлогов и затем выполнять разбор этой конструкции в терминах морфологических характеристик главного и зависимого слов, включая леммы. Затем выполнялась как смысловая разметка конструкций лингвистами-операторами с опорой на дефиниции значений предлогов в терминах модифицированных синтаксем Г.А.Золотовой, так и автоматизированная классификация.

Переход от метода CQL-запросов к автоматическому выделению предложных конструкций на основе синтаксических зависимостей позволил значительно уменьшить трудоемкость и при этом повысить точность извлечения описываемых единиц. Важным преимуществом подхода является также возможность извлечения предложных конструкций с осложненной структурой (предшествованием предложной именной группы управляющему слову, наличием подчиненной именной группе клаузы и т.д.), что способствует извлечению и исследованию количественно и качественно более объемного материала. В проведённом исследовании впервые для русского языка были использованы для семантической классификации предложных конструкций методы машинного обучения с учителем, реализованные на корпусе предложных конструкций, семантически размеченных вручную. Эксперимент показал, что предложные значения, наиболее репрезентативно представленные в размеченном вручную корпусе, успешно определяются на тестовом материале (полнота 0.795, точность 0.771). Удовлетворительная точность результатов автоматической классификации показывает, что задача семантической классификации предложных значений может быть с достаточной степенью эффективности решена методами машинного обучения при наличии в достаточном объеме качественно размеченных обучающих данных.

Для выполнения задач проекта был разработан метод, комбинирующий структурный лингвистический анализ и модифицированный анализ корпусных данных. Результат экспертной разметки конструкций показал, что для предложных конструкций согласованность экспертных мнений по поводу «детализированной» схемы анализа (Fine-Grained) довольно низкая, что совпадает с аналогичными исследованиями по разметке лексических значений [Erk K., McCarthy, D., Gaylord N., (2013) Measuring Word Meaning in Context. In: Computational Linguistics September 2013, Vol. 39, No. 3: 511–554], которые больше подходят для экспертного анализа. Использование нейронных сетей, которые часто рассматриваются как универсальный метод анализа, также показывает, что не только в лингвистической области, но и в сфере перцептивного восприятия довольно «детализированная» схема классификации ухудшает распознавание образов [Chen, Zh., Ding, R., Chin Ting-Wu, Marculescu D., Understanding the Impact of Label Granularity on CNN-based Image Classification. In: 2018 IEEE international conference on data mining workshops (ICDMW), 895-904]. Второй осложняющий момент состоял в том, что корпусная статистика давала неустойчивые результаты.

Комбинированный метод корпусного анализа состоял в выработке методики получения и анализа случайной выборочной совокупности контекстов из корпусов русских текстов общей тематики. В рамках выделенной совокупности производится дифференциация исследуемых параметров. Важным является случайный характер отбора контекстов в выборку, что не всегда реализовано в больших корпусах, например в НКРЯ. Эксперименты с объемом выборочной совокупности показали, что для высоко- и

среднечастотных явлений подходят выборки от 50 контекстов, для низкочастотных явлений объем выборки необходимо повышать до 100 или даже 1000 контекстов. Если доли грамматических и лексических значений в рамках случайной выборки превосходят 5%, то они хорошо коррелируют с данными, которые получаются при увеличении выборки на порядок. Доля частотности какого-либо параметра менее 5% в выборке является неустойчивой, не масштабируемой, для таких значений необходимо расширение выборочной совокупности, более того необходимо понимать, что эти значения будут отличаться от корпуса к корпусу. Важно подчеркнуть, что конкретное значение корпусной частотности является менее значимым, чем соотношение частот, поскольку в оценке частотности присутствует потенциальная ошибка экстраполяции.

Дифференциация значений предложных конструкций позволила выработать трехуровневую онтологическую схему анализа предложных конструкций.

(1) Обобщенный уровень систематизации значений по наиболее общим параметрам семантических параметров предложения позволил выделить шесть базовых семантических рубрик: Локализация, Темпоратив, Объектив, Квалификатив, Квантификатив, Партитив и седьмую рубрику Дериватив, в которой предлог в сочетании с другими знаменательными лексемами (существительными, наречиями и глагольными деепричастиями) формирует новые предложные единицы или входит в структуру других составных единиц, таких как наречия или фразеологизмы. Последняя рубрика отчасти выпадает из собственно смысловой последовательности, однако игнорирование единиц такого типа приведет к неполноте даже обобщенной классификации, кроме того ограничит схему анализа закрытым перечнем базовых позиций, что абсолютно не соответствует природе языка, в котором одновременно присутствует «результат» и «процесс», т. е. эта область включает предложные конструкции, которые в перспективе могут менять внутреннюю конфигурацию семантических рубрик. Некоторые компоненты рубрики Дериватив уже грамматически вошли в структуру собственно семантических рубрик, поэтому между этой рубрикой и другими, например Темпоративом, есть зона пересечения в корпусной статистике. Корпусная статистика выделенных рубрик позволила сформулировать понимание статистики собственно грамматической реализации: от 14 тысяч до полутора тысяч ipm (instances per million — число примеров на миллион токенов в корпусе). Важным является то, что корпусная частота рубрик создает ряд геометрически убывающей прогрессии, хотя и с определенными нарушениями за счет совмещений некоторых реализаций: 14k, 12k, 9k, 7k, 6k, 3k, 2k ipm.

(2) Следующий уровень — более детализированных единиц в пределах рубрики, для которых мы использовали термин Г.А.Золотовой "синтаксема". В нашем понимании это предложная конструкция, соотносимая с семантической рубрикой и представляющая определенное единство морфосинтаксической и лексической спецификации главного слов, присоединяющего предложную конструкцию, группы предлогов, каждый из которых присоединяет конкретную падежную форму существительного определенного семантического класса. Таким образом синтаксема включает несколько предложных конструкций, синонимичных или частично синонимичных, но не антонимичных. Лексические спецификации главных и зависимых слов в синтаксеме частично пересекаются. Всего нами выделено в семи рубриках 23 синтаксемы (от 3 до 4 для каждой рубрики). Среднее значение корпусной частоты для синтаксем равно 2267 ipm. Практически все синтаксемы по корпусной статистике находится в зоне грамматических значений от 8k до 1k ipm, за исключением трех, которые попадают в зону лексико-грамматических значений. Синтаксемы в рамках рубрики также образуют ряд убывающей прогрессии.

(3) Практически каждая из синтаксем за исключением лексико-грамматических имеет ряд более детализированных вариантов — субсинтаксем, которые в свою очередь могут иметь один или несколько вариантов реализации предложной конструкцией.

Приведем наиболее очевидный примеры субсинтаксем для самой частотной рубрики Локализация. Самая частотная в рубрике синтаксема Локатив (7465 ipm). Она описывается в словаре Г.А.Золотовой перечнем предложных конструкций «от + род, между + род, против + род, среди + род, у + род, за + твор, между + твор, над + твор, перед + твор, под + твор, в + предл, на + предл, при + предл». При корпусном анализе наиболее частотной является конструкция предлог “в/во” в сочетании с формой предложного или местного падежа> [3700 IPM]: сидеть в саду, гулять в лесу. Вторую позицию по частотности занимает конструкция с предлогом “на” в сочетании с формой предложного или местного падежа [1800 IPM]: сидеть на стуле, дышать воздухом на веранде. В [Herskovits A. 1985. Semantics and Pragmatics of Locative Expressions. Cognitive Science, vol. 9, no. 3, pp. 341–378] разница в значениях связывается с определенной идеей «включения объектов» для первого предлога и «опоры» и «смежности объектов» для второго. Однако различие в значениях скорее имеет лингвистический характер, поскольку в отношении многих пространственных объектов неясно, есть ли там включение. Например, «веранда» является трехмерным объектом, и человек находится внутри нее (сидеть на веранде), аналогично стоять на улице, висеть на дереве и т. п., причем в других индоевропейских языках и даже славянских для этих объектов используется конструкция с предлогом “в”. Это означает, что перед нами собственно внутриязыковая классификация пространственных объектов, не всегда логичная. Эти конструкции могут сочетаться в одном контексте: сидеть на веранде в кресле, висеть в бильярдной на стене, что обычно трактуется как доказательство того, что перед нами «разные» семантико-синтаксические роли. Очевидно, что локализация объектов носит «телескопический» характер типа настраиваемого фокуса: более грубо или более точно. То есть первая конструкция с предлогом “в” является субсинтаксемой «локатив_1» в силу большей корпусной частотности, а конструкция с предлогом “на” «локатив_2», поскольку она уступает первой по частотности. Если в приведенных примерах поменять местами предложные конструкции (висеть на стене в бильярдной), то второй локатив становится атрибутом (партитивом) первого. В качестве синтаксических «хозяев» для синтаксем локатив1 и 2 являются разнообразные типы глаголов: существования, принятия положения в пространстве, восприятия и проч. Проще указать на семантические типы, плохо сочетающиеся с этими синтаксемами, это глаголы направленного движения. Полные описания всех введенных синтаксем и субсинтаксем приведены в приложении.

Как видно из предыдущего примера, субсинтаксемы высокочастотной синтаксемы находятся в зоне грамматических значений по корпусной статистике, отсюда не вполне логичная дифференциация пространственных объектов, которые в лексических тезаурусах будут попадать в разные группы. Всего в 23 синтаксемах выделено 72 субсинтаксемы, 173 пары «предлог-субсинтаксема», из них 86 пар для первообразных предлогов, их суммарная частота 35359 ipm, среднее значение 411 ipm; 48 пар с производными предлогами, суммарная частота 2415 ipm, средняя частота 50 ipm; лексически связанных конструкций с первообразными предлогами 15, суммарная частота 641 ipm, среднее значение равно 43 ipm.

Первоначальная группировка полностью или частично синонимичных предложных конструкций используется как база для их объединения в кластеры, для которых вычисляются долевые значения частотности по их реализации в случайных выборках корпусных контекстов. Это позволило перейти к идее трехуровневой онтологии. Полученная квантитативная онтология была использована для построения двухуровневой

аффиксальной порождающей грамматики гибридного типа AGFL (https://www.usenix.org/legacy/publications/library/proceedings/usenix02/tech/freenix/full

_papers/koster/koster.pdf), с включением в генерацию данных о частотных соотношениях предложных конструкций, смысловых типах их синтаксических доминант и зависимых слов.

Принятое в качестве базовой единицы разрабатываемого формализма понятие синтаксемы Г.А.Золотовой [Золотова Г.А. Синтаксический словарь: Репертуар элементарных единиц русского синтаксиса. 4-е изд. М.: Едиториал УРСС, 2011] потребовало уточнения как в плане метода принятия решения о соответствии типу, так и в плане построения связанной системы онтологических параметров. Используя данные корпусной статистики о «долях» частотности предложной синтаксемы, а также сходств и различий при определении семантических типов лексем, занимающих позицию синтаксически главного и зависимого компонентов в предложной конструкции, мы построили иерархию типов: семантические рубрики → синтаксемы → субсинтаксемы.

Поясним структуру на примере наиболее частотной семантической рубрики локализации.

Наиболее частотная синтаксема локатив <“в/во” + форма предложного или местного падежа> [3700 IPM]: сидеть в саду, гулять в лесу. Это же значение может быть выражено предлогом <“на” + форма предложного или местного падежа> [1800 IPM]: сидеть на стуле, дышать воздухом на веранде. В [Herskovits A. 1985. Semantics and Pragmatics of Locative Expressions. Cognitive Science, vol. 9, no. 3, pp. 341–378.] разница в значениях была ассоциирована с идеей «включения объектов» для первого предлога и «опоры» и «смежности объектов» для второго. Однако различие в значениях скорее имеет лингвистический характер, поскольку «веранда» является трехмерным объектом, и человек находится внутри нее. Эти подтипы синтаксем могут сочетаться в одном контексте: сидеть на веранде в кресле, висеть в бильярдной на стене, что обычно трактуется как доказательство того, что перед нами «разные» семантико-синтаксические роли. Очевидно, что локализация объектов носит «телескопический» характер типа настраиваемого фокуса: более грубо или более точно. В таких случаях мы рассматриваем их как подтипы синтаксем (субсинтаксемы): вариант с “в” является локативом1, а с “на” – локативом2, уступающим первому по частотности. Если в приведенных примерах поменять местами предложные конструкции (висеть на стене в бильярдной), то второй локатив становится атрибутом (партитивом) первого. В качестве синтаксических «хозяев» для синтаксем локатив 1 и 2 являются разнообразные типы глаголов: существования, принятия положения в пространстве, восприятия и проч. Проще указать на семантические типы, плохо сочетающиеся с этими синтаксемами - это глаголы направленного движения (см. ниже синтаксему директив). Важным отличием указанных синтаксем является классификация объектов местоположения по типу их оформления с помощью “в” или “на” в зависимой от предлога форме. Часть объектов входят в область пересечения, для них в определенной степени выполняется указанное A.Herskovits противопоставление, однако при потенциальной возможности двух вариантов локативов для объекта в корпусной реализации лишь один представлен как базовый. Подобным образом на реальном корпусном материале были проанализированы и описаны все субсинтаксемы синтаксемы локатив /

Другая синтаксема, директив, в рамках рубрики локализации демонстрирует последовательный параллелизм с сочетаемостью «слуг» локативов для предлогов “в”, “на” и “за”, обозначая конечную точку траектории движения, при этом предлоги присоединяют в форме винительного те же группы зависимых существительных, которые были перечислены выше. В качестве главных в таких конструкциях выступают глаголы направленного движения или перемещения объектов. В зависимости от предлога и семантического класса объекта выделяются субсинтаксемы директива. Еще одна синтаксема, связанная с траекторией движения, — транзитив, обозначает пересекаемое пространство. В группе частотных предлогов она выражается <“по” + форма дательного падежа> [360 IPM]: пройти по коридору, спускаться по лестнице. Другие синтаксемы транзитива также выходят за рамки группы ядра, ближайший по частотности <“через” + форма винительного падежа> [135 IPM]. Более подробно все синтаксемы семантической рубрики локализации описаны в наших публикациях, напр., Zakharov, V., Azarova, I. Grammatical parallelism of Russian prepositional localization and temporal constructions. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2020, 12284 LNAI, с. 122-134.

Необходимо указать еще на одно важное свойство предложных синтаксем — их взаимодействие с префиксальными дериватами доминирующих глаголов. Обычная трактовка этой связи формулируется следующим образом: префиксальные глаголы создают конструкции «управляемых» предложно-падежных форм [Скобликова Е.С. Роль лексики в словосочетаниях с управляемым компонентом // Очерки по теории словосочетания и предложения. Куйбышев, 1990. С. 25–46], причем префикс регулярно совпадает с предлогом в такой конструкции. Эти формы рассматриваются как управляемые в силу отношения восполнения (комплетивности) конструкции, поскольку без предложно-падежной формы будет возникать неясность, о чем идет речь. В частности, префиксальные глаголы направленного движения связаны с директивом, департивом и транзитивом.

Разработанная нами классификация предложных значений (онтология) является первой универсальной семантической классификация значений предлогов и предложных конструкций. Она совершенствует лежащую в ее основе классификация синтаксемных значений Г.А. Золотовой (Золотова Г. А. Синтаксический словарь: Репертуар элементарных единиц русского синтаксиса. Изд. 4-е. М.: Едиториал УРСС, 2011) в сторону сокращения инвентаря значений, введения иерархии значений и транспозиции описанных значений на неописанные у Золотовой производные предлоги. Полученная классификация также отличается от существующих словарных описаний предложных значений (Русский Викисловарь, Малый Академический Словарь) системностью и единообразием описания определяемых значений.

Таким образом, проведенное исследование позволяет нам представить в систематизированном виде предложные конструкции современного русского языка, уточнить и дополнить систему семантических типов, сопроводить конструкции характеристиками корпусной статистики, что является абсолютно новым способом представления данных, для которого нет аналогов не только в современной отечественной, но также и в зарубежной лингвистике.

Исследования на тему автоматической семантической классификации предложных конструкций для других языков ведутся уже достаточно давно (см. Litkowski, Kenneth & Hargraves, Orin. (2007). SemEval-2007 Task 06: Word-sense disambiguation of prepositions. 24-29. 10.3115/1621474.1621479; Rudzicz, F., and Mokhov, S. (2010) Towards a Heuristic Categorization of Prepositional Phrases in English with WordNet. Technical Report, Cornell

University, (2003); Бугаков О.В. Создание семантического словаря предложных конструкций на основе Украинского национального лингвистического корпуса. Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2009» (Бекасово, 27-31 мая 2009 г.). Вып. 8 (15).– М.: РГГУ, 2009. с. 51-56), но до сих пор они немногочисленны и отсутствуют аналогичные исследования для русского языка.

Остановимся на методах автоматической классификации значений предложных конструкций. Большинство описанных в научной литературе исследований по проблеме семантической классификации предложных конструкций опирается на наличие готового корпуса с семантической разметкой (напр., PropBank) или использование семантической разметки из лексических баз данных (WordNet, см. Girju, Roxana. (2009). The Syntax and Semantics of Prepositions in the Task of Automatic Interpretation of Nominal Phrases and Compounds: A Cross-Linguistic Study. Computational Linguistics. 35. 185-228. DOI 10.1162/coli.06-77-prep13). Иными словами, подобные исследования традиционно предполагают наличие внешнего готового ресурса, содержащего семантическую информацию о составляющих конструкций, которые затем проходят определённую обработку. Для русского языка корпусов, размеченных в терминах синтаксем, нет, как и вообще нет таких ресурсов. Подход, использованный нами, не требует готового корпуса (с полной семантической разметкой корпуса) и привлечения сторонних ресурсов, работающих с синтаксемами Г.А. Золотовой, а базируется на специально сформированном корпусе семантически размеченных предложных конструкций, полученном непосредственно в проекте.

Реализованные в проекте методы классификации значений предложных конструкций с использованием алгоритма машинного обучения с учителем по сути и по результатам сопоставимы с имеющимся зарубежным опытом. В частности, известно исследование английских предложных значений с применением методов машинного обучения (Schneider, Nathan, Hwang, Jena D., Srikumar, Vivek, Prange, Jakob, Blodgett, Austin, Moeller, Sarah R., Stern, Aviram, Bitan, Adi, and Abend, Omri (2018). Comprehensive supersense disambiguation of English prepositions and possessives. In Proc. of ACL, pages 185–196. Melbourne, Australia). Данный подход также был успешно апробирован для классификации значений предложных конструкций в севернокитайском языке (Peng, Siyao Logan & Liu, Yang Janet & Zhu, Yilun & Blodgett, Austin & Zhao, Yushi. (2020). A Corpus of Adpositional Supersenses for Mandarin Chinese). Эффективность нашего подхода к применению методов машинного обучения с учителем для автоматического определения семантических значений предлогов подтверждается высокими показателями точности и полноты классификации.

В области лингвистической семантики было выполнено несколько основополагающих работ по семантике предлогов, их пространственных [Филипенко 2000] и объектных значениях [Солоницкий 2003]. Для использования данных по значениям предлогов в рамках функциональной грамматики Г.А.Золотова создала словарь синтаксем [Золотова 2011], в котором были описаны предложные конструкции, являющиеся синтаксическими аналогами употребления падежных форм существительных, кроме того в словаре были приведены примеры употребления синтаксем и ограничения на семантические классы слов, присоединяющих предложные конструкции, однако иллюстративный материал не отвечает принципам корпусного анализа. В плане описания грамматической таксономии помимо классического описания в Грамматике современного русского языка [АГ-1980] и работ М.В. Всеволодовой [Всеволодова, М.В. Теория функционально-коммуникативного синтаксиса. М. 2000] была выполнена серия работ совместно в Институте русского языка им. В.В.Виноградова и Национальном

исследовательском университете «Высшая школа экономики» [Ляшевская, О.А., Кашкин, Е.В. Типы информации о лексических конструкциях в системе // Труды института русского языка им. ВВ Виноградова 6, 464-556, 2015.]. Были проведено корпусное исследование функционирования предлогов в текстах различных функциональных стилей [Всеволодова, М.В., Поликарпов, А.А., Русские предлоги и средства предложного типа. Материалы к функционально-грамматическому описанию реального употребления. Книга 1. Введение в объективную грамматику и лексикографию русских предложных единиц. Москва: URSS, 2014]. В современных зарубежных исследованиях по функционированию предлогов основной проблемой является вопрос о присоединении предложно-падежной конструкции к глаголу или существительному для задач синтаксического анализа текста [Agirre, E., Baldwin, T., Martinez D., (2008) Improving Parsing and PP attachment Performance with Sense Information // Proceedings of ACL-08: HLT, pages 317–325, Columbus, Ohio, USA, June 2008], однако проблема автоматического определения значений предлогов как части модуля семантической интерпретации текста является важным компонентом системы автоматического анализа текста [Dahlmeier, D., Hwee Tou Ng, Schultz T., (2009) Joint Learning of Preposition Senses and Semantic Roles of Prepositional Phrases // Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 450–458, Singapore, 6-7 August 2009]. Дифференциация предложных значений на материале английского языка осуществляется на базе размеченных корпусов типа PropBank [O'Hara, T., Wiebe, J., (2009) Exploiting Semantic Role Resources for Preposition Disambiguation // Computational Linguistics June 2009, Vol. 35, No. 2: 151–184]. В последнее время формируются базы данных для контрастного описания функционирования предлогов в европейских языках [Girju, R., (2009) The Syntax and Semantics of Prepositions in the Task of Automatic Interpretation of Nominal Phrases and Compounds: A Cross-Linguistic Study // Computational Linguistics June 2009, Vol. 35, No. 2: 185–228], такие базы данных используются для прояснения значений предлогов [Gonen, H., Goldberg, Y., (2016) Semi Supervised Preposition-Sense Disambiguation using Multilingual Data // Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 2718–2729, Osaka, Japan, December 11-17 2016].

Новизна исследования состоит в том, что разработанная нами методика корпусного анализа значения предложных конструкций на основе дифференциации значений в случайной выборочной совокупности контекстов позволила его выполнить без трудоемкого формирования размеченных семантических корпусов. Полученные данные позволили критически осмыслить представленные в лингвистических и словарных описаниях значения предлогов, сгруппировать их в кластеры значений, имеющих сходные параметры контекстного окружения, сформулировать критерии грамматического использования именных, наречных и глагольных конструкций как производных предлогов. Пилотная проверка автоматической дифференциации предложных конструкций с использованием современных методов показала релевантность разработанной предложной онтологической схемы.

Выполненное исследование позволяет связать воедино работы в области лингвистической семантики, функциональной грамматики, грамматической таксономии, описания корпусной статистики лингвистических параметров текстов для разных функциональных стилей современного русского языка, формального описания.грамматики предложных конструкций, создания размеченных корпусов как базы машинного обучения.