Разработка базы данных по семантике русских предлогов

А.Д. Москвина, Е.В. Еникеева, В.П. Захаров

Санкт-Петербургский государственный университет

moskvina.anya@gmail.com, protoev@yandex.ru, v.zakharov@spbu.ru

Аннотация

В данной статье рассматривается разработка базы данных, содержащей информацию о предложных конструкциях русского языка. Работа проводится в рамках проекта по созданию квантитативной грамматики русских предложных конструкций. Исходной информацией является материал из нескольких корпусов (Araneum Russicum Minus, подкорпуса НКРЯ) и результаты его семантической обработки. На данный момент размечено 4200 употреблений двадцати пяти первообразных предлогов, на выборке из этих данных построена экспериментальная база данных.

Представление накопленного материала в виде базы данных позволяет оптимизировать процесс получения статистической информации об употреблении предлогов, а также упростит изучение связей между семантическими и структурными свойствами конструкций. Так, для каждой конструкции пользователь сможет извлечь информацию о ее частоте, реализованных значениях предлога, получить расширенный контекст, сравнить статистику по корпусам разных жанров.

В статье рассматриваются особенности материала, обосновывается структура разрабатываемой базы данных, приводятся примеры запросов. Предварительные результаты говорят о целесообразности дальнейшей разработки.

Ключевые слова: русские предлоги, предложные конструкции, значения предлогов, корпусная лингвистика, база данных

Библиографическая ссылка: Москвина А.Д., Еникеева Е.В., Захаров В.П. Разработка базы данных по семантике русских предлогов // Компьютерная лингвистика и вычислительные онтологии. Выпуск 3 (Труды XXII Международной объединенной научной конференции «Интернет и современное общество», IMS-2019, Санкт-Петербург, 19 – 22 июня 2019 г. Сборник научных трудов). — СПб: Университет ИТМО, 2019. С. 106–115. DOI: 10.17586/2541-9781-2019-3-106-115

1. Введение

Предлоги не только встречаются среди наиболее частотных слов русского (и не только) языка, но и представляют собой связующий, оформляющий отношения между знаменательными словами элемент, способный коренным образом изменить значение синтаксической единицы, в которую он входит. Несмотря на существование разных подходов к описанию семантики предлогов, нельзя полностью отрицать наличие у них и некоторого лексического компонента [1]. Кроме того, в русском языке класс предлогов является открытым и активно пополняется, а имея доступ к обширным корпусам текстов, мы обладаем возможностью наблюдать в динамике процесс трансформации сочетаний, например, первообразного предлога со знаменательным существительным в цельную единицу, функционирующую как предлог. Нам представляется, что качественное изучение данной единицы невозможно без учета фактов непосредственного языка, представленного и доступного сегодня в виде больших корпусов текстов. Данная работа является частью проекта, в рамках которого мы хотим изучить особенности функционирования

предложных конструкций в русском языке. Мы поставили задачу создать максимально полное формальное квантитативное описание поведения предлогов в конструкциях, до сих пор отсутствующее для русского языка, изучить и описать их лексико-грамматические и семантические особенности. Наше исследование является корпусно-ориентированным, то есть в качестве материала мы используем данные корпусов разных жанров, тематики и происхождения. Мы изучаем значения предлогов через особенности их функционирования в составе словосочетаний в реальных текстах, выявляя отношения омонимии между предлогами, управляющими разными падежами и отношения полисемии внутри одного предлога, а также и синонимичности между разными предлогами, имеющими сходные значения и управляющими одинаковым падежом. Значение предлога понимается здесь как обобщенное значение целой предложной конструкции.

В этой статье мы сфокусируемся на том, каким образом и почему разумнее структурировать накопленный в процессе работы материал, исходя из стоящих перед нами задач и теоретических предпосылок, в виде базы данных.

Мы подробно рассмотрим структуру и пользовательские возможности разрабатываемого продукта. Концепция проекта, размышления и содержательные результаты более подробно описаны в статьях [2, 3].

2. Этапы разработки базы данных

2.1. Особенности материала

Разрабатывая базу данных, прежде всего нужно понять с какой целью мы ее создаем и как собираемся ее использовать, и, исходя из этого, продумать ее структуру. В нашем случае мы попытаемся понять, как можно структурировать имеющиеся данные и какую часть процесса мы сможем автоматизировать. Поэтому первая часть статьи будет посвящена описанию процесса сбора статистической информации о предложных конструкциях.

В основе интересующих нас особенностей предлогов лежит такое явление, как полисемия предлогов. В русском языке большинство предлогов обладают несколькими значениями. Однозначными являются некоторые производные составные предлоги (*в связи с, при помощи*), их значение не зависит от конкретного случая использования. Значение же многозначных предлогов (пространственное, временное, объектное) реализуется в контексте, к этой группе относится большинство первообразных предлогов. Класс первообразных (примитивных) предлогов является закрытым, зато частотным, и значительная часть данного исследования посвящена работе именно с ним. Перед нами не стояла задача использовать результаты аннотации конструкций для лингвистической обработки текста более высокого уровня, предлоги и предложные конструкции сами по себе являются объектом нашего исследования. Одна из наших гипотез заключается в том, что как функциональную единицу следует рассматривать не сам предлог, но его сочетание с тем падежом, которым он управляет. Наш подход к семантическому описанию системы предложных конструкций русского языка требует ручной разметки значительного количества материала. Несмотря на существование нескольких словарных грамматических описаний полисемии русских предлогов и различных подходов к описанию семантики предлогов в целом, квантитативный анализ и исчерпывающее описание функционирования русских предложных конструкций в реальном языке до сих пор отсутствует. В нашей работе мы используем несколько готовых классификаций значений предлогов: по Г. А. Золотовой [4], по МАС [5], и wiki словарю [6].

В процессе работы с различными источниками, как теоретическими, так и с живым материалом, возникали дополнительные вопросы и задачи. Тем не менее, важной частью процедуры стала работа разметчиков, для которых были разработаны пошаговые инструкции работы с подкорпусами НКРЯ (основной, газетный, устный и поэтический) и

вебкорпусом Araneum Russicum Minus [7]. На первом этапе для каждого предлога составляется шаблон-запрос в формате CQL, формируется конкорданс (см. рис. 1), из которого сохраняется случайная выборка указанного объема, которая представляет собой список предложных конструкций. Отдельно сохраняется список более широких контекстов. CQL-запрос представился наиболее удобной формой обращения к корпусам, предусмотренной в корпус-менеджере NoSketchEngine [8]. Он позволяет сформулировать запрос на основании линейной последовательности языковых единиц заданных частей речи (atag="Nn") и других грамматических категорий, а также конкретных лексических единиц (lemma="для"), при помощи набора морфосинтаксических тегов, операторов конъюнкции и дизъюнкции, указании контекстного окна. Используемый в NoSketchEngine тегсет основан на [9], что делает запрос читаемым и универсальным при работе с другими корпусами-источниками.

На первых этапах работы при составлении запроса мы пытались учитывать конструкции разнообразной структуры - глагольное и именное управление, усложненное определениями или нет, – в рамках одного шаблона.

Рис. 1. Пример упорядоченного конкорданса Aranea

Пример такого запроса для предлога "в":

```
[atag != "Zz"] [atag = "Vb"] [atag != "Zz"& atag != "Vb"] {0,2}
[lemma="в" | lemma="во"] [atag = "Aj"|atag = "Pn"]{0,2} [atag="Nn" &
word!="соответствии" & word!="результате" & word!="области" &
word!="связи" & word!="сфере" & word!="силу" & word!="пределах"]
[atag="Nn"]
```

Следует отметить, что подобные запросы носят чисто формальных характер и не учитывают семантических особенностей и даже внутреннего согласования единиц; далеко не все сочетания с предлогами, попадающие в список на выходе, являются цельными конструкциями или синтаксемами, и требуют дополнительной обработки. Несмотря на эти издержки, вызванные грубой обработкой большого количества материала, мы пытались найти баланс между точностью и сложностью запроса, настраивая его таким образом, чтобы полученный список было впоследствии удобно упорядочить по "хозяину" или "слуге".

Также в некоторых случаях в запрос добавлялись лексикализированные стоп-слова, предупреждающие появление составных предлогов. Так, в приведенном выше примере исключаются наиболее частотные составные предлоги (предложные сочетания), включающие в себя предлог «в»: *в соответствии*, *в результате*, *в области* и т.д. Частность была рассчитана на основе встречаемости этих сочетаний в корпусе Araneum Russicum Minus.

Один разметчик работал, как правило, с одним предлогом; данные сохранялись в виде таблиц Excel, в которых далее проводилась ручная разметка. Каждой конструкции ставилось в соответствие значение, реализованное предлогом в конкретном контексте, затем для каждого значения подсчитывалась процент, суммарная частота и ipm (см. табл.1).

Таблица 1. Размеченные конструкции с предлогом «в»

Lemma	Frequency	Значение	Суммарная частота	Процент	Ipm
храниться в музей	2	Локатив	5	4	601
домен в зона	2	Локатив	32	4	601
установить в Дайтоне	1	Локатив	32	4	601

Несмотря на наличие общего шаблона образца, разметчик мог и порой был вынуждены модифицировать и уточнять запрос. Например, как было описано выше, для предлога «в» показалось эффективней исключить составные предложные сочетания на этапе поиска по корпусу. Аналогичная статистика собиралась по нескольким корпусам. Поиск конструкций осуществляется по корпусам разных функциональных стилей, например, устный, газетный, поэтический подкорпуса НКРЯ (см. табл. 2).

Таблица 2. Сводная таблица значения предлога «через» по корпусам, % от числа отобранных

Корпус НКРЯ	Темпоратив	Транзитив	Медиатив	Дименсив	Фразеологизм
Основной	51,0	30,4	17,6	2,0	0,5
Газетный	49,0	16,5	35,2	0	0,4
Устный	48,0	27,5	23,2	0,5	0,9
Поэтический	23,2	70,4	5,5	0	1,2

Средние показатели ранговой частоты предлогов, с одной стороны, указывают на особенности функционирования предлогов в зависимости от функциональных характеристик рассматриваемого корпуса. С другой стороны, накопление статистических данных по разнообразным корпусам (а точнее, по выборкам из них) позволит в последствии оценить качество и сбалансированность других корпусов, по которым вычисляются квантитативные характеристики, в частности веб-корпусов типа Aranea.

Отдельный этап работы связан с синонимией предложных значений. Для исследуемого предлога на основе словарей и собственных наблюдений создавался список синонимов, для каждого из которых производились аналогичные подсчеты. Все данные вносились в сводную таблицу, имеющую следующий вид:

Таблица 3. Фрагмент сводной таблицы синонимичных значений предлога «под»

предлог	значение	Araneum Russicum веб-корпус 120 млн. (200 случ. контекстов/200 частотных контекстов/)	
		употр.	ipm
под	директив	5 (2,51%)/180 (2,11%)	19,9 26,9
близ	директив	1013	8,4
возле	директив	5 (2,5%)	1,1
возле	ДРУГОЕ ЗНАЧЕНИЕ не из списка значений для «под»	9 (4,75%)	2,1
около	директив	73 (36,5%)	92,1
вокруг	директив	9800	81,7
рядом с	директив	27 (13,5%)	7,9

Предложные значения описываются в рамках синтаксем — минимальных семантико-синтаксических единиц, выступающих как носитель элементарного смысла и компонент более сложных синтаксических конструкций [10]. В рамках нашего материала исследуемые синтаксисы представляют собой предложные конструкции, в которых фиксируются грамматические характеристики главного и зависимого компонентов («хозяина» и «слуги») и падежное управление, то есть выбор падежа зависимого, определяемый вершиной словосочетания и предлогом [1]. Предполагается также, что значение предлога в рамках заданной таким образом конструкции может определяться лексико-семантическим классом «хозяина» или «слуги» (например, описанных в классификации проекта «Лексикограф», которая продолжает использоваться в разметке НКРЯ [11]. Одной из задач, таким образом, стала попытка выработать наиболее показательные и удобные грамматические шаблоны. Как уже было отмечено выше, в ходе работ подходы к составлению таких шаблонов менялись: были попытки в одном шаблоне учесть максимальное количество возможных конструкций с исследуемым предлогов (управление глагола, существительного, расположение "хозяина" и "слуги" справа или слева от предлога и т. д.); напротив, были попытки составления максимально простых шаблонов типа предлог + существительное в определенном падеже. Во втором случае, для исследования поведения одного предлога в одном корпусе требуется совершить целый ряд запросов. На основе эмпирических данных менялся также объем сохраняемой случайной выборки. Частоты, проценты и ipm подсчитывались в каждом отдельном случае практически вручную. Следует отметить, что инструкции тоже не раз менялись и уточнялись на основании анализа полученных данных, что привело к неоднородности накопленного материала.

Таким образом, с одной стороны, имеется уже достаточно большое количество потенциально информативного материала в неудобном для восприятия и автоматизированного анализа виде. С другой стороны, возникает потребность оптимизировать дальнейшую работа по сбору и структурированию данных о предложных конструкциях.

2.2. Структура базы данных

Создание базы данных позволяет решить две основные задачи. Во-первых, такой вид представления данных упростит и ускорит процедуру обработки и квантитативного анализа информации, во-вторых, позволит разработать интерфейс для разметчиков.

База данных по семантике русских предложных конструкций разрабатывается в MySQL Workbench для эксплуатации в СУБД MySQL. База данных состоит из таблиц, хранящих информацию о предлогах и их возможных значениях, в виде конструкций, извлеченных из корпусов на разных этапах работы, количественной информации об используемых корпусах и их типах, шаблонов, использующихся для поиска. Например, предлогу «на» будет соотнесена информация о всех его допустимых значениях (как локатив, темпоратив, объект и т.д.), несколько шаблонов для поиска по корпусам (универсальных, сформулированных в виде CQL запроса Sketch Engine), списки полученных по этим запросами контекстов (*мы всегда можем договориться. Эта цена которую < вы видите в прайс листе > не последняя.*), списки лемматизированных на этапе предобработки с помощью нашего инструмента конструкций (*вы видеть в прайс лист*), частоты данных конструкций и их значений. Благодаря установленным связям для каждой конструкции пользователь сможет извлечь информацию о ее частоте, значениях предлога, получить расширенный контекст. С другой стороны, на каждое значение можно получить список найденных в том или ином корпусе конструкций. Таким образом, работа с SQL-запросами позволит оптимизировать анализ накопленной за время работы проекта информации и подтвердить или опровергнуть рабочие гипотезы.

Наличие адекватного интерфейса упростит работу для разметчиков и уменьшит вероятность совершения ошибки (например, выбор несуществующего значения). Для

просмотра, редактирования и пополнения базы данных разрабатывается веб-приложение. Предполагается реализовать следующие возможности: просмотр агрегированной статистики употребления предлогов, поиск определённой конструкции; добавление и редактирование результатов ручной обработки. Благодаря наличию строгой схемы, данные разметчиков валидируются при загрузке, что уменьшает вероятность ошибки (например, выбор несуществующего значения).

На данный момент создана модель базы в SQL Workbench (см. рис. 2) и начат процесс ее наполнения. База содержит двадцать пять первообразных предлогов, которым соотнесены их значения, и около 300 размеченных употреблений из разных корпусов. Рассмотрим структуру подробнее. Центральная таблица базы (условно названная constructions) представляет собой список предложных конструкций (словосочетаний с предлогом) в качестве первичных ключей, и полный набор интересующих нас характеристик, в виде собой ссылок на другие таблицы. Для каждой конструкции содержится следующая информация: предлог и падеж, значение, реализованное в данном употреблении, корпус, из которого извлечена конструкция, шаблон, по которому производился запрос, контекст, информация о семантических классах участников конструкции, дата обращения к корпусу. Таблица с предлогами (preps) содержит список предлогов и их характеристики (простой/составной, первообразный/производный). Каждому предлогу при помощи отношений многие-ко-многим (таблица preps_has_meanings) заданы допустимые для него значения (таблица meanings). В свою очередь каждому корпусу и подкорпусу (таблица corpora_info) приписаны его характеристики и информация о его объеме. Список значений, использующийся при разметке, основан на классификации Г. А. Золотовой [4], переработанном и дополненном отсутствующими в нем значениями из МАСа и вики-словаря [5, 6]. Эти три источника с разной степенью детализации описывают значения предлогов, для дальнейшего анализа в базе содержатся в виде таблицы связи между значениями в разных классификациях. Под контекстом понимается более широкий нелемматизированный фрагмент текста, иногда необходимый для уточнения смысла и значения предложной группы. Шаблоны, формулирующие запросы поиска по корпусам хранятся в таблице templates, которая связана с таблицей корпусов с помощью промежуточной таблицы template_corpus_freq, содержащей статистическую информацию по данному запросы в данном корпусе. Эта связь делает нас независимыми от числа корпусов, оставляя их список открытым. Таблица semantics содержит информацию о семантическом классе «хозяев» и «слуг».

Все использующиеся нами в качестве источника материалы (сводные таблицы со значениями, списки падежей, наборы семантических классов, списки предлогов с характеристиками и их синонимов) поэтапно заносятся в базу данных в виде таблиц с установленными отношениями-связями, позволяя пользователям и разработчикам формировать необходимые для исследования SQL-запросы.

2.3. Возможности базы данных

Предварительные эксперименты по разметке и квантитативному анализу корпусных данных производились без использования специального интерфейса разметки. Поэтому при подготовке инструментов работы с базой данных предложных значений мы решаем две основные задачи:

– обработка размеченных данных, их валидация;
– собственно разработка инструментов редактирования базы данных и их просмотра.

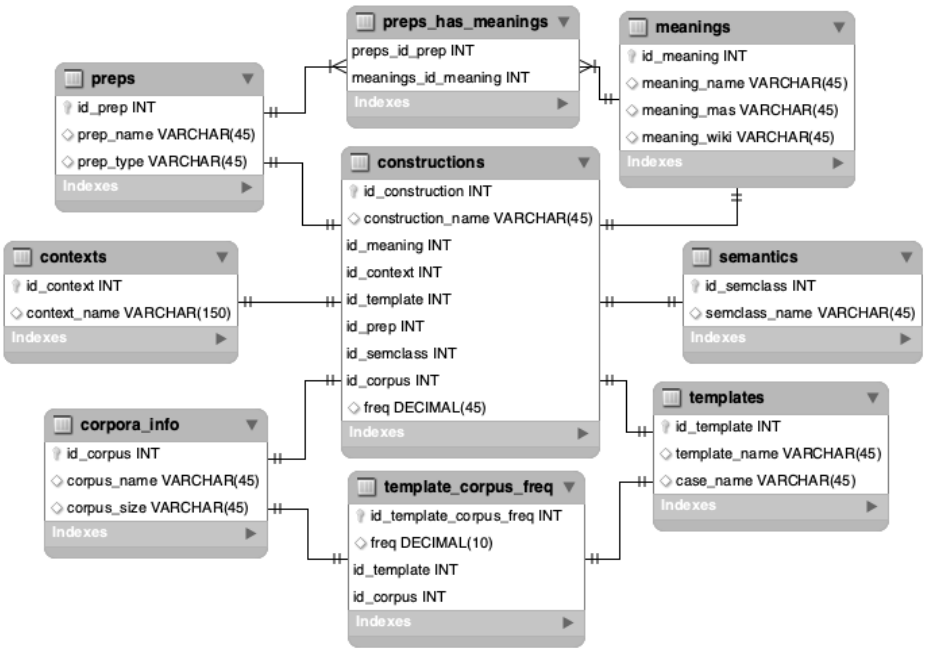

Рис. 2. Схема базы данных по семантике русских предлогов

В размеченных данных в таблицах Excel и отчетах содержится вся информация, которая описана выше: предлог и падеж, значение, реализованное в данном употреблении, корпус, из которого извлечена конструкция, шаблон, по которому производился запрос, контекст. Значительная часть данных представлена в удобном для импорта в базу виде. Однако возникали и некоторые затруднения. Поскольку конструкции (результат выдачи) и контексты извлекаются NoSketch Engine в разные файлы без сохранения соответствий, а конструкции хранятся в лемматизированном виде, то на первом этапе предобработки с помощью морфологического анализатора pymorphy2 [12] восстанавливаются соответствия между конструкциями и контекстами, в которых они встречаются в корпусе. Затем преобразованные данные сопоставляются с имеющейся в базе информацией, сохраняются новые записи и производится валидация: например, проверяется, что указанные разметчиками значения предлогов есть в соответствующей таблице. Редактор в режиме разметки позволяет добавить новые записи, то есть конструкции, для которых разметчик выбирает необходимое значение.

Статистические данные (доля отдельных значений среди размеченных контекстов, ipm и так далее) собираются с помощью запросов на SQL в отдельную сводную таблицу.

Рассмотрим возможности базы данных для исследования и подтверждения или опровержения рабочих гипотез на примере SQL-запроса. Допустим, нас интересует, какие из описанных в словарях значений предлога «в» действительно встречаются в языке регулярно. То есть мы хотим узнать, сколько раз и в каком значении был употрелен предлог «в». В нашем случае запрос будет иметь следующий вид:

```
SELECT COUNT(constructions.id_meaning) AS meanings_freq, meaning_name,
constructions.id_meaning
FROM preps.constructions
JOIN preps.meanings
ON constructions.id_meaning=meanings.id_meaning
WHERE constructions.id_prep = "1"
GROUP BY constructions.id_meaning
```

В результате мы получим таблицу (см. табл. 4) и узнаем, как распределены значения по частотности, какие из них встречаются часто, а какие единично или вовсе отсутствуют. В данном примере использовалась выборка в 100 употреблений - столько контекстов употребления предлога «в» (id = «1») содержится в базе. Уточняя запрос и меняя размер выборки, можно выяснить, какой объем является оптимальным с точки зрения статистических исследований.

Таблица 4. Пример результата SQL-запроса (фрагмент) для предлога «в»

meanings_freq	meanings_name
32	локатив
19	директив
14	объект
16	темпоратив

Итак, с точки зрения пользователя, разрабатываемая база данных позволит посмотреть примеры употребления и статистику, связанные с любым конкретным предлогом, где в качестве ключа поиска могут выступать падеж, тип, значение, корпус, семантический класс хозяев и слуг, а также любые их комбинации. Все поля планируется реализовать как выпадающий список с возможностью выбора значения.

3. Заключение

В основе подхода к выделению предложных конструкций посредством языка запросов (шаблонов на языке регулярных выражений) лежит идея создания грамматики лексико-грамматических шаблонов как инструмента выявления разнообразного реестра конструкций и инструмента их комплексного изучения. База данных является хорошим подспорьем для создания сетевой структуры, состоящей как из первообразных предлогов, так и из производных, элементы которой связаны отношения условной синонимии разного характера. База данных может стать инструментом изучений грамматико-семантических особенностей функционирования предлогов, статистически описывающей функционирование системы русских предлогов в реальном языке, прообразом которого в нашем случае являются корпусы.

Промежуточные результаты, как нам кажется, позволяют утверждать, что работа представляет интерес для дальнейшей разработки и принесет научную ценность. На небольшом материале мы разработали и протестировали работу базы данных, продолжаем ее оптимизацию и наполнение, и в скором времени сможем формулировать новые содержательные выводы.

следование выполнено при поддержке Российского фонда фундаментальных исследований, грант № 17-29-09159 «Квантитативная грамматика русских предложных конструкций».

Литература

[1] Русская грамматика / Н. Ю. Шведова (гл. ред.). — М.: Наука, 1980. Т. 2: Синтаксис.
[2] Азарова И.В., Захаров В.П., Москвина А.Д. Семантическая структура русских предложно-падежных конструкций // Компьютерная лингвистика и вычислительные онтологии. Вып. 2 (Труды XXI Международной объединенной конференции «Интернет и современное общество, IMS-2018, Санкт- Петербург, 31 мая — 2 июня 2018 г. Сб. научных статей). СПб., 2018. С. 9–16.
[3] Захаров В.П. О компьютерной онтологии русских предлогов // Российская академическая лексикография: современное состояние и перспективы развития.

Сборник научных статей по материалам Международной научной конференции, посвященной 70-летию выхода первого тома академического «Словаря современного русского литературного языка» / отв. ред. О.Н. Крылова, С.А. Мызников, М.Н. Приёмышева, Е.В. Пурицкая; Ин-т лингв. исслед. РАН. — СПб: Нестор-история, 2018. С. 180-191.
[4] Золотова Г.А. Синтаксический словарь: Репертуар элементарных единиц русского синтаксиса. Изд. 4-е. М., 2011.
[5] Словарь русского языка: в 4 тт. / под ред. А.П. Евгеньевой и Г.А. Разумниковой. Изд. 3-е, стереотип. М., 1988.
[6] Википедия. URL: https://ru.wikipedia.org/ (дата обращения 11.12.2018).
[7] Benko V. Aranea: Yet Another Family of (Comparable) Web Corpora // Petr Sojka, Aleš Horák, Ivan Kopeček and Karel Pala (Eds.). Text, Speech and Dialogue. 17th International Conference, TSD 2014, Brno, Czech Republic, September 8-12, 2014.
[8] Rychlý P. A Modular Corpus Manager // 1st Workshop on Recent Advances in Slavonic Natural Language Processing. Brno: Masaryk University, 2007. p. 65-70.
[9] Sharoff S. Kopotev M., Erjavec, T., Feldman A., Divjak D. Designing and evaluating a Russian tagset // Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008. European Language Resources Association (ELRA). 2008. P. 279-285.
[10] Золотова Г.А. Очерк функционального синтаксиса русского языка. М., 1973.
[11] Кустова Г. И., Ляшевская О. Н., Падучева Е. В., Рахилина Е. В. Семантическая разметка лексики в Национальном корпусе русского языка: принципы, проблемы, перспективы // Национальный корпус русского языка: 2003-2005. Результаты и перспективы. — М., 2005. С. 155—174.
[12] Korobov M. Morphological Analyzer and Generator for Russian and Ukrainian Languages // Analysis of Images, Social Networks and Texts. 2015. P 320-332.

Development of a Database on the Russian Prepositions Semantics

A. Moskvina, E. Enikeeva, V. Zakharov

Saint Petersburg State University

The paper discusses the development of a database, containing information on Russian prepositional phrases. The study is conducted as a part of a project, aimed at the construction of quantitative grammar of Russian prepositional phrases. Source data is based on the material from several corpora (Araneum Russicum Minus, Russian National Corpus subcorpora) and the result of its semantic processing. At this point 4200 cases of prepositional usage of twenty-five primitive prepositions are annotated. We created a database based on a sample of this data.

By structuring and integrating the accumulated material into a database we optimize the process of extracting statistical characteristics, as well as information on exercised meanings of a preposition in a wider context. We also have the opportunity to compare the statistics of the texts of different genres.

In the paper we examine the peculiarities of the material of prepositional semantics, justify our views on structure of such database, give the examples of queries. Preliminary results suggest the feasibility of further work.

Keywords: Russian prepositions, prepositional phrases, semantics of prepositions, corpus linguistics, databases

Reference for citation: Moskvina A.D., Enikeeva E.V., Zakharov V.P. Development of a database on the Russian prepositions semantics // Computer Linguistics and Computing Ontologies. Vol. 3 (Proceedings of the XXII International Joint Scientific Conference «Internet

and Modern Society», IMS-2019, St. Petersburg, June 19-22, 2019). - St. Petersburg: ITMO University, 2019. P. 106 – 115. DOI: 10.17586/2541-9781-2019-3-106-115

Reference

[1] Russkaya grammatika / N. Yu. Shvedova (gl. red.). — M.: Nauka, 1980. — T. 2: Sintaksis. (In Russian).
[2] Azarova I.V., Zaharov V.P., Moskvina A.D. Semanticheskaya struktura russkih predlozhno-padezhnyh konstrukciy // Kompyuternaya lingvistika i vychislitelnye ontologii. Vyp. 2). SPb., 2018. P. 9–16. (In Russian).
[3] Zaharov V.P. O komp'yuternoj ontologii russkih predlogov // Rossijskaya akademicheskaya leksikografiya: sovremennoe sostoyanie i perspektivy razvitiya. Sbornik nauchnyh statej po materialam Mezhdunarodnoj nauchnoj konferencii, posvyashchennoj 70-letiyu vyhoda pervogo toma akademicheskogo «Slovarya sovremennogo russkogo literaturnogo yazyka» / otv. red. O.N. Krylova, S.A. Myznikov, M.N. Priyomysheva, E.V. Purickaya; In-t lingv. issled. RAN. — SPb.: Nestor-istoriya, 2018. (In Russian).
[4] Zolotova G.A. Sintaksicheskiy slovar: Repertuar elementarnyh edinic russkogo sintaksisa. Izd. 4-e. M., 2011. (In Russian).
[5] Slovar russkogo yazyka: v 4 tt. / pod red. A.P. Evgenjevoj i G.A. Razumnikovoj. Izd. 3-e, stereotip. M., 1988. (In Russian).
[6] Wikipedia. URL: https://ru.wikipedia.org/ (access date 11.12.2018). (In Russian)
[7] Benko V. Aranea: Yet Another Family of (Comparable) Web Corpora. In Petr Sojka, Aleš Horák, Ivan Kopeček and Karel Pala (Eds.): Text, Speech and Dialogue. 17th International Conference, TSD 2014, Brno, Czech Republic, September 8-12, 2014.
[8] Rychlý P. A Modular Corpus Manager. In 1st Workshop on Recent Advances in Slavonic Natural Language Processing. Brno: Masaryk University. 2007. P. 65-70.
[9] Sharoff S. Kopotev M., Erjavec, T., Feldman A., Divjak D. Designing and evaluating a Russian tagset. Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008. European Language Resources Association (ELRA). 2008. P. 279-285.
[10] Zolotova G.A. Ocherk funkcionalnogo sintaksisa russkogo yazyka. M., 1973. (In Russian).
[11] Kustova G. I., Lyashevskaya O. N., Paducheva E. V., Rahilina E. V. Semanticheskaya razmetka leksiki v Nacionalnom korpuse russkogo yazyka: principy, problemy, perspektivy // Nacionalnyj korpus russkogo yazyka: 2003-2005. Rezultaty i perspektivy. — M., 2005, 155—174. (In Russian).
[12] Korobov M.: Morphological Analyzer and Generator for Russian and Ukrainian Languages // Analysis of Images, Social Networks and Texts. 2015. P 320-332.