СЕМАНТИЧЕСКАЯ ИНТЕРПРЕТАЦИЯ ЧАСТОТНЫХ ПРЕДЛОЖНО-ПАДЕЖНЫХ КОНСТРУКЦИЙ ПО КОРПУСУ РУССКИХ ТЕКСТОВ

SEMANTIC INTERPRETATION OF RUSSIAN PREPOSITION PHRASES BASING ON CORPUS FREQUENCIES

Азарова И.В*. (azic@bsr.spb.ru), Санкт-Петербургский государственный университет*

В докладе рассматриваются параметры семантического описания предложно-падежных конструкций при автоматическом анализе русских текстов. Рассматриваемый вариант описания будет использован в формально-грамматическом парсере Russ4IR, сопряженном с компьютерным тезаурусом RussNet.

1. Структура данных системы анализа текста в проекте Идеограф

Проблема семантической интерпретации предложно-падежных конструкций в русском языке в данном докладе рассматривается применительно к определенной системе анализа текста в проекте Идеограф, в которой используется формально-грамматический парсер Russ4IR[1] и компьютерный тезаурус RussNet[2]. Описываемая система анализа задает структуру данных[3], в рамках которой происходит семантическая интерпретация. В грамматическом плане в тексте идентифицируются группы синтаксически связанных слов, образующих независимое простое предложение или предложение в составе сложного: устанавливаются отношения доминирования между элементами группы, определяется частеречная принадлежность составляющих словосочетания и их канонические формы (леммы), которые интерпретируются по тезаурусу RussNet. Вхождение леммы в структурную единицу тезауруса (синсет) позволяет отождествить вершину семантического дерева, иногда просто вышестоящий синсет, который может использоваться для семантической субкатегоризации реализованного лексического значения. На синсетах заданы базовые семантические отношения: антонимия, пресуппозиция, каузация и проч. Окказиональные слова (отглагольные существительные, относительные прилагательные и др.), построенные по регулярным словообразовательным моделям, при интерпретации отсылают к синсету мотивирующего слова при помощи деривационно-семантической связи: синонимии, гипонимии, ролевой характеристики, транспозиции.

Грамматическая и лексическая неоднозначность частично снимается за счет рамок валентностей[4], входящих в статьи тезауруса. Рамки валентности включают устойчивые грамматические и семантические контекстные маркеры, характеризующие реализацию того или иного значения в выборочной совокупности контекстов в корпусе современных текстов Бокренок кафедры математической лингвистики СПбГУ. Каждая валентность в рамке получает характеристику «устойчивости», которая вычисляется по контекстам корпуса. В тех случаях когда зафиксированные маркеры рамок валентностей не снимают полностью неоднозначность текста, они используются для того, чтобы упорядочить варианты анализа, задавая в качестве первого варианта анализа тот, который отвечает в наибольшей степени «стереотипному» описанию контекстов корпуса.

В процессе фиксации рамок валентностей было обнаружено, что для отдельных групп значений, в частности для глаголов движения, грамматическая (морфо-синтаксическая) спецификация валентности, представленная предложно-падежной конструкцией, имеет тенденцию к варьированию используемого предлога. В отдельных случаях можно было выделить один или несколько «доминирующих» предлогов, например у глагола *направиться: к* или *в*, в других – целый ряд предлогов используется примерно с равной частотностью. В случа-

---

[1] Азарова И.В., Секликов Ю.В., Иванов В.Л. Интерпретация текстовых документов с использованием формальной грамматики AGFL и компьютерного тезауруса RussNet // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной кон-ференции Диалог'2004. Верхневолжский, 2–7 июня 2004 г . М., 2004. С. 1–6.

[2] Азарова И.В., Митрофанова О.А., Синопальникова А.А. Компьютерный тезаурус русского языка типа WordNet // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции Диалог'2003. Протвино, 11–16 июня 2003 г . М., 2003. С . 43–50.

[3] Азарова И.В., Иванов В.Л., Овчинникова Е.А. Семантическая структура пропозиции при извлечении фактов из текстов на рус-ском языке // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции Диалог'2005. Звениго-род, 1–7 июня 2005 г. М., 2005. С. 6–11.

[4] Азарова И.В., Иванов В.Л., Овчинникова Е.А. Использование схемы наследования рамок валентностей в тезаурусе RussNet для автоматического анализа текста // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции Диалог'2006. Бекасово, 31 мая – 4 июня 2006 г. М., 2006. С. 18-25.

ях доминирующих предлогов было очевидно, что варианты выражают разные «фокусы» представления ситуации: более обобщенное описание атрибутов действия (чаще) или более подробное описание их (реже). И главное – конкретные наборы доминирующих предлогов не наследовались в семантическом дереве глаголов движения. На основании этого был сделано следующее предположение: возможно, грамматическую спецификацию следует задавать в более обобщенной форме – групп предложно-падежных конструкций.

Кроме того, семантический модуль Russ4IR предусматривает набор стереотипных интерпретаций для грамматических правил, которые определяются по частотности их реализации в нашем корпусе современных текстов. Были рассмотрены интерпретации присубстантивных падежных форм[5] в терминах семантических деревьев RussNet. Следующий шаг в этом направлении будет сделан в данном докладе: будут описаны основные параметры семантического описания предложно-падежных конструкций, которые частотно реализуются в упомянутом корпусе современных текстов.

2. Семантическая природа предложных конструкций

Значение предлогов, особенно первообразных, было объектом рассмотрения в большом числе исследований[6], однако остается неясным ряд вопросов. Является ли пространственное значение ядром (прототипом в определенной мере) для остальных значений, и как интерпретировать значения первообразных предлогов, которые не имеют регулярных пространственных употреблений? В какой степени значение предлога зависит от значения присоединяемого имени, а также тех слов (глаголов, существительных и проч.), которые «предсказывают» появление предложной конструкции? Даже элементарная проблема «номенклатуры» –исчисления предлогов – и та требует определенного решения. Рассматривать ли в качестве отдельной единицы предложно-падежную форму (сочетание предлога с падежной формой имени)? Как ограничить перечень производных предлогов (предложных сочетаний)? Дополнительной проблемой, которая связана с типом представления семантической информации в проекте Идеограф, является то, что в тезаурусе RussNet. хранится информация о лексических значениях основных частей речи: существительных, глаголов, прилагательных и наречий. Местоименные слова «проецируются» на структуры знаменательных слов (то есть вхождения в текст местоимений приравниваются к вхождению соответствующих знаменательных слов), например, личные местоимения 1-го и 2-го лица имеют проекцию на дерево существительных «человек». Каким образом задавать семантическую интерпретацию значений предлогов: в виде отдельных семантических правил или через проекцию на тезаурус?

Поскольку задачей данного исследования является выработка принципов представления информации о значениях предлогов в системе автоматического анализа текста, нам в первую очередь требуется данные о частых, регулярных явлениях, связанных с интерпретацией значений предлогов. Для этой цели мы исследовали употребление предлогов в двух совокупностях из 1000 контекстов, отобранных случайным образом из корпуса Бокренок. Одна совокупность характеризовала приглагольные, а вторая – присубстантивное употребления предложно-падежных групп.

2.1. Приглагольные употребления предложных конструкций

Разметка приглагольных предложных конструкций в выборочной совокупности показала, что примерно половина (54%) используемых предлогов являются первообразными, остальные – производные. Первообразные предлоги покрывают 93% контекстов. На схеме 1 приведено распределение частот предлогов в обследованной совокупности. Очевидно, что наиболее частотные предлоги *в, на* и *с* устойчиво занимают доминирующие позиции в частотных списках[7].

Контексты употребления предложных конструкций были размечены в отношении следующих обобщенных типов значений: пространственные, объектные, обстоятельственные, отдельно временные, переносные или фразеологизированные. Доли контекстов в выборке для перечисленных значений составили: пространственные – 33%, объектные – 26%, обстоятельственные – 18%, временные – 9%, переносные – 14%, что достаточно четко указывает на то, что пространственные и объектные употребления являются первичными, составляют более половины от общего числа употреблений предлогов.

---

[5] Азарова И.В., Овчинникова Е.А. Семантическая интерпретация именных конструкций по корпусу русских текстов // Труды ме-ждународной конференции «Корпусная лингвистика – 2006». СПб., 2006. С. 25–33.

[6] См. напр.: Исследования по семантике предлогов: Сборник статей. М., 2000; Пекар В.И. Семантика предлогов вертикальной со-положенности в когнитивном аспекте: Автореф. дис. … канд. филол. наук. Уфа, 2000.

[7] Ср. Шаров С.А. Частотный словарь // URL: http://www.artint.ru/projects/frqlist.asp

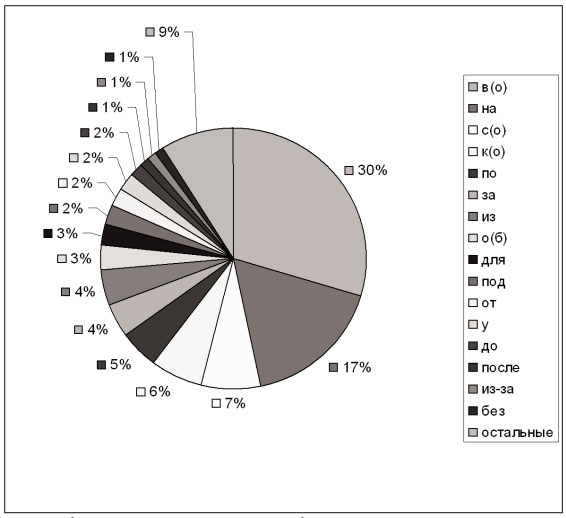

Схема 1. Распределение частот предлогов, используемых при глаголах

Распределение значений для каждого из предлогов неравномерно, поэтому можно выделить «характерные» значения для предлогов на основании оценки «неслучайности»[8]

$$MI(pr, z) = \log_2(\frac{fr_{pr|z}}{fr_{pr} \times fr_z})$$

где:

$fr_{pr|z}$ – относительная частота контекстов предлога, в которых он используется в значении z;

fr_{pr} – относительная частота предлога в выборочной совокупности;

fr_z – относительная частота значения предлога z во всей совокупности.

Характерные значения MI, большие единицы, имеют следующие предложные значения: *для* – обстоятельственное; *до* – временное; *за* – временное; *к(о)* – объектное; *о(б)* – объектное; *по* – обстоятельственное; *под* – переносное; *после* – временное; *при* – обстоятельственное; *с(о)* – обстоятельственное.

Частотные предлоги (*в, на, с*) имеют в выборке все виды значений, остальные – лишь часть, хотя, возможно, это связано с недостаточной представительностью выборки для менее частотных предлогов.

Среди семантических групп глаголов, к которым присоединяются предложно-падежные формы, особо выделяются глаголы движения (18% контекстов) и глаголы общения (8%). Вычисляя оценку «неслучайности», предлоги можно связать с определенными группами глаголов: *для* – глаголы использования; *за* – глаголы восприятия, глаголы обладания; *к(о)* – глаголы изменения, глаголы общения; *в(о)* – глаголы принятия положения в пространстве; *по* – глаголы движения; *с(о)* – глаголы состояния, глаголы социального взаимодействия, глаголы общения; *о(б)* – глаголы общения; *из* – глаголы движения.

2.2. Присубстантивное употребления предложных конструкций

В совокупности контекстов, иллюстрирующих присубстантивное употребление предложно-падежных групп, доли частотных предлогов отчасти похожи на приглагольные употребление: *в(во)* – 21%, *с(со)* – 17,5%, *на* – 12%, *о(об)* – 6%, *из* – 6%, *к(ко)* – 6%, *по* – 5%, *от* – 5%. Общая доля высокочастотных предлогов (*в+с*) почти в 1,5 раза меньше, чем у приглагольных предлогов.

Доли обобщенных значений предлогов в присубстантивных контекстах существенно отличаются от рассмотренных выше: пространственные – 8%, объектные – 46%, обстоятельственные – 37%, временные – 3%, переносные – 6%. Кроме того, анализ семантических типов существительных, присоединяющих частотные предлоги, показывает, что значительная часть (80%) предложных употреблений «унаследованы» от глаголов; семанти-

[8] Азарова И.В., Синопальникова А.А., Смрж П. Представление устойчивых лексических сочетаний в компьютерном тезаурусе RussNet // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции Диалог'2005. Звенигород, 1-7 июня 2005 г. М., 2005. С. 11-17.

ческие группы существительных, часто присоединяющих предлоги: *в(во)* – обозначение процесса, *с(со)* – артефакта, *на* – действия или процесса, *о(об)* – сообщения, *из* – артефакта, *к(ко)* – состояния, *по* – сообщения, *от* – процесса. В случае объектных значений наиболее часто используется предлог *в*, который присоединяется к существительным, обозначающим процессы, для обстоятельственных (предлог *с*) – артефакты, для пространственных (предлог *в*) – артефакты и совокупности.

3. Параметры семантического описания предложных конструкций

Проанализированные контексты употребления предлогов показали, что нет общей для всех предлогов схемы соотношения с доминирующими словами, в роли которых чаще всего выступают глаголы и существительные. Те предлоги, которые используются приглагольно, как правило, имеют значительную часть (около 40%) «унаследованных» предложно-падежных употреблений у отглагольных существительных, поэтому в этом смысле приглагольная модель является статистически доминирующей. В этой модели регулярно присутствуют две основные части: объектные или в более широком смысле аргументные употребления предлогов и «сирконстантые» (собственно обстоятельственные) употребления, для которых характерна связь со значением присоединяемого предлогом имени. Соотношение аргументных и обстоятельственных употреблений довольно сильно варьируется: некоторые предлоги скорее используются для оформления аргументных позиций (например, предлог *о* или предлог *в*, управляющий винительным), другие в подавляющем количестве случаев передают обстоятельственные значения (например, предлог *в*, управляющий предложным/ местным падежом, или предлог *из-за*).

В случае аргументного употребления значение предлога не требует экспликации, поскольку доминирующее слово задает его семантический тип в валентной рамке, хотя повторяющаяся валентная функция может абстрагироваться и использоваться относительно независимо, например, в детерминирующих членах. Тогда можно предположить, что оправданным будет построение обобщенной рамки валентности для семантического дерева, в которой будут перечислены все аргументные позиции, реализованные в данной группе значений, в сочетании со способами предложного/ предложно-падежного оформления этих позиций, которые упорядочены в соответствии с частотностью употребления морфологических форм. Например, для дерева глаголов движения[9] валентность «конечная точка» будет иметь варианты *в*+В.п., *на*+В.п., *к*+Д.п. ... *в сторону*+Р.п., *в направлении*+Р.п., наречие направления (*вперед, назад, влево* и проч.). Мы уже указывали на то, что мотивированные предлоги используются довольно редко в сравнении с первообразными: частота встречаемости наиболее частотных из них (например, *в течение*) ниже на порядок в сравнении со «средними» значениями первообразных (например, *в*+В.п.), что приводит к тому, что они не попадают в описание рамки валентности, поскольку не обладают статистической регулярностью. Обобщенная рамка валентности позволит задать класс «условной эквивалентности» предложно-падежных форм в рамках заданного семантического дерева (лексико-семантической группы глаголов), которые в строгом смысле не являются синонимами, поскольку могут передавать внефокусное или фокусное представление действия.

Обстоятельственные употребления предлогов регулярно связаны с семантическим типом присоединяемого имени. Например, конструкция *в*+В.п. в сочетании с именем, принадлежащим к семантическому дереву «место; местоположение», будет передавать пространственное значение (*в стране, долине*), а в сочетании с именем из дерева «время» – временное значение (*в детстве, истории*). Осложняет ситуацию то, что многие имена могут иметь множественное подчинение, то есть принадлежат к разным деревьям, даже если речь идет об одном значении, например *страна* обозначает также жителей и относится к дереву «совокупность», *история* является коммуникативным объектом, а *детство* – состоянием человека. Как интерпретировать конструкцию *в*+В.п. при сочетании с «нехарактерными» именами? Например сочетание с названиями естественных или искусственных объектов (*в реке, сугробах, комнате, лодке*), частями тела человека (*во рту, кишечнике*), совокупностями (*в группе, листве*) будет иметь пространственное значение, осложненное дополнительными оттенками, а сочетания с именами-процессами (*в деятельности, выборах*) – временное значение. Потенциальная возможность «распространения» некоторого значения с наиболее типичного семантического дерева на другие, примыкающие к нему, позволяет предположить наличие характерных группировок семантических деревьев, сходных с объединениям деревьев, которые используются при семантической спецификации валентности. Отсюда возникает также и идея о способе описания значений предлогов: их следует рассматривать как своеобразные «проекции» на глагольные деревья, например, принятия положения в пространстве или событий.

[9] Азарова И.В., Иванов В.Л., Овчинникова Е.А. Использование схемы наследования рамок валентностей в тезаурусе RussNet для автоматического анализа текста.

4. Выводы и перспективы исследования

Были рассмотрены регулярные, частотно реализующиеся в контекстах корпуса современных текстов, характеристики предложно-падежных конструкций. Полученные данные требуют дальнейшего осмысления и указывают на значительное варьирование базовых параметров этих конструкций: от типа управляющего (или грамматически доминирующего) слова до смысловой характеристики, выражаемой предлогом в сочетании с определенной падежной формой. В описываемой модели предполагается, что интерпретация обстоятельственных предложно-падежных сочетаний однотипна представленным в тезаурусе лексическим значениям основных частей речи (в первую очередь, глаголов) и может проецироваться на соответствующие узлы семантических деревьев с использованием элементов валентных рамок – через задание группировок деревьев, имеющих одинаковое или сходное значение. Характер распределения проекций позволит представить возможное градуирование предложных значений и выделить лексические значения, являющиеся прототипами обстоятельственных.

Аргументные употребления предложно-падежных конструкций следует фиксировать в обобщенных рамках валентностей для семантических деревьев, что позволит подключить к их описанию менее частотных мотивированных предлогов.