СЛОВАРНАЯ ПОДДЕРЖКА ПРОЦЕДУР СЕМАНТИЧЕСКОЙ ИНТЕРПРЕТАЦИИ ПРЕДЛОЖНЫХ СВЯЗЕЙ[1]

DICTIONARY SUPPORT FOR SEMANTIC INTERPRETATION OF PREPOSITION RELATIONS

В. Ш. Рубашкин

СПбГУКИ, С.-Петербург

vrub@mail.nw.ru

[1] Доклад подготовлен при частичной поддержке РФФИ (проект № 03-06-80109)

Рассматривается вопрос о параметрах словарного описания предлогов в семантическом словаре, ориентированном на поддержку процедур анализа текста. Показано, что достаточно использовать всего несколько словарных характеристик. Кратко о самих процедурах и требованиях к функциональности словаря

.

1. Задачи

О такой сравнительно частной проблеме, как семантика предлогов, написано столько, что, казалось бы, трудно прибавить к этому что-либо достаточно значимое. Тем не менее, мы рискнем предложить точку зрения, отличную от большей части уже высказанных суждений. И дело здесь не в претензиях поставить под сомнение всю массу сделанных десятками, если не сотнями лингвистов наблюдений и выводов. Дело даже не в возможности разных интерпретаций языковых наблюдений. Дело в том, что в контексте инженерной семантики должна быть достаточно радикально изменена точка зрения на задачи словарного описания предлогов. Разъяснению и конкретизации этого тезиса и посвящен предлагаемый доклад.[2]

Описание семантики предлогов на уровне естественного языка в стиле обычных толковых словарей для носителя языка с практической точки зрения заведомо излишне: предложная подсистема вместе с грамматикой составляет ядро языка; без свободного владения тем и другим вообще не может идти речь о свободном владении языком. Попытки максимально подробно инвентаризировать на словарном уровне все оттенки смысла, которые способен выразить тот или иной предлог, как показывает опыт построения таких описаний, ведут в дурную бесконечность. Сами принципы дискретизации смыслового поля предлога являются предметом дискуссий, которые вряд ли могут в такой постановке завершиться конструктивным результатом.[3] Поучительно взглянуть с этой точки зрения также на словарные описания предлогов в двуязычных словарях. Легко убедиться, что описания соответствий, в сущности, оказываются там набором примеров, рассчитанных на способность пользователя словаря самостоятельно произвести их адекватную генерализацию - с опорой на языковую интуицию, сформировавшуюся в процессе овладения родным языком, и, сверх того, собранием фразеологизмов, демонстрирующих случаи нестандартного и идиоматического их употребления.

Словарное описание предлогов практически значимо как раздел машинного словаря; имеется в виду концептуальный словарь ("онтология") - вычислительный ресурс, предназначенный для поддержки технологий автоматической обработки текста, прежде всего процедур семантического анализа текста. И с этой точки зрения состав словарного описания предлогов должен определяться исключительно востребованной алгоритмами анализа функциональностью словаря. В контексте так поставленной задачи всю совокупность имеющихся содержательных описаний семантики предлогов можно, если угодно, считать лишь предварительной систематизацией материала, необходимой для построения такого словаря.

Семантическая интерпретация предлога должна рассматриваться как частный случай более общей задачи - задачи семантической интерпретации синтаксических связей. Если между словами (в общем случае - текстовыми элементами) *W1* и *W2* парсером обнаружена (сочтена возможной) синтаксическая связь, ставится вопрос о ее семантическом эквиваленте. В случае предложной связи вопрос может быть сформулирован так же, но в этом случае речь идет о конструкции вида *W1---> P---> W2*, где *W1* – (возможный) синтаксический хозяин, *W2* - синтаксический слуга, а предлог *P* маркирует связь между *W1* и *W2*, которая и является объектом семантической интерпретации. Подчеркнем, что интерпретируется не связь между знаменательным словом и предлогом, а связь двух знаменательных слов, уточняемая и конкретизируемая с помощью соединяющего их предлога.

2. Целевой язык

Задача машинного понимания текста сводится к переводу с естественного языка на язык представления знаний (ЯПЗ), в котором точно описаны правила построения и правила вывода. Поэтому выбор параметров словарного описания предлогов должен соотноситься с двумя вопросами.

(1) Какой ЯПЗ используется и какие элементы его выражений могут быть сопоставлены предлогам.

(2) В каких процедурах анализа естественно-языковых выражений и как именно будут использоваться предлагаемые словарные описания.

Универсальным по своим выразительным возможностям языком представления знаний является язык логики. Другие корректно определенные ЯПЗ (семантические сети, фреймовые и продукционные языки, языки реляционных БД) - даже если они первоначально строились совершенно независимо от

---

[2] Мы вынуждены ограничиться здесь минимумом ссылок на литературу. Достаточно представительный список литературы можно найти в работе [1], посвященной специально русским предлогам; там же дан, на наш взгляд, весьма убедительный анализ общей ситуации в этой области исследований. См. также, например, обзорную работу [2].

[3] См. [1, с. 11-35].

логических языков - переводимы на язык логики: любые выражения этих языков могут быть представлены некоторым ограниченным классом формул логического языка.

Очевидно, в языке логики нет прямого аналога предложной системы ЕЯ. Как будет видно из дальнейшего, в разных языковых ситуациях (в семантически разных типах контекстов) интерпретация предлога - даже одного и того же предлога - может быть принципиально разной. Тем не менее, **задача семантической интерпретации предложных связей всегда состоит в том, чтобы указать их коррелят в логической формуле, представляющей смысл рассматриваемого словосочетания**. Именно поэтому отправным пунктом при определении необходимого набора параметров для описания семантики предлогов должен быть не естественный язык - со всеми его неисчерпаемыми возможностями выражения самых тонких оттенков смысла, - а целевой язык, т. е. язык логического представления.

Поскольку предложная связь определена в пределах простого предложения, а аналогом простого предложения считается предикатная формула

$$R\,(x_1, x_2, ..., x_n),$$

то интерпретацию предлога следует соотносить с элементами этой формулы.

Однако эта стандартная схема требует существенных уточнений и дальнейшей детализации в нескольких направлениях. Во-первых, в отношении способов спецификации ролевых позиций индивидов в предикатной записи; во-вторых, в отношении способа представления в ней самих актантов; в третьих, в отношении способов представления в логической нотации числовой информации наряду с собственно концептуальной; в четвертых, в отношении способов формализации (семантической) сочетаемости имен актантов с именем предиката. В рамках нашей темы наиболее существенны два первых пункта.

В отношении способов спецификации ролевых позиций существенно следующее. В математической логике сложилась традиция различать смысловые роли индивидных термов относительно терма отношения ***R*** порядком их следования: «x_1 - первый объект при предикате ***R***», «x_2 - второй объект при предикате ***R***» и т.д. Это означает, что содержательная характеризация роли просто выносится за пределы ЯПЗ - в неформальный комментарий, где, скажем, говорится: «первая позиция при предикате *покупать* - это *покупатель* (*кто купил*); вторая позиция - *продавец* (*у кого купил*); третья позиция - *товар* (*что куплено*)». Такой стиль формализации исторически был обусловлен уже сложившейся к моменту изобретения Расселом и Уайтхедом современной логической нотации традицией представления функций; с другой стороны - и это более существенно - тем, что набор ролевых отношений в математическом языке достаточно беден, так что содержательная характеризация смысловых ролей здесь не актуальна - ролевые позиции либо равнозначны и не требуют спецификации (как в записи $a = b$), либо недвусмысленно определяются самим предикатным символом (как в записи $a > b$). Иначе обстоит дело в естественном языке и формируемых на его базе профессиональных подъязыках эмпирических наук. Здесь имеет место большое разнообразие смысловых ролей и невозможность однозначно мотивировать порядок их следования в логической формуле смыслом и символизацией предиката. Поэтому здесь нужна нотация, позволяющая ввести указание смысловых позиций **внутрь** ЯПЗ, сделать это частью самого формализма. Для этого:

a) в ЯПЗ должна быть заранее определена номенклатура различаемых ролей; соответственно, в словарь языка должен быть добавлен **список ролевых указателей:** $\rho_1, \rho_2, \rho_3, ...$ [4]

b) должна быть введена маркировка ролей в самой предикатной записи:

$$R\,(\rho_1{:}\, x_1, \rho_2{:}\, x_2, ..., \rho_n{:}\, x_n);$$ [5]

Другое уточнение касается использования индивидных констант. Основной тезис здесь состоит в том, что за пределами языка математики профессиональные тексты оперируют не *собственными именами*, непосредственно указывающими (или даже, можно сказать, предъявляющими) соответствующий индивидуальный объект, а *дескрипциями*. В соответствии с этим описание ситуации, например, *покупки* должно быть представлено не в виде записи с индивидными константами

$$КУПИЛ\,(кто{:}\,a_1,\ у\ кого{:}\,a_2\ \ что{:}\,a_3),$$

а посредством присоединения дескрипций всех актантов:

$$КУПИЛ\,(Agent{:}\,x_1,\ Experiencer{:}\,x_2,\ Theme{:}\,x_3)\ \&\ A_1(x_1)\&A_2(x_2)\&A_3(x_3)$$ [6].

---

[4] Список ролей должен устанавливаться исходя прежде всего из принципа **минимальной достаточности** выразительных возможностей языка: большое количество ролей скорее вредит алгоритмизации, чем улучшает ее качество. Практический опыт работы с языковым материалом позволяет утверждать, что для описания актантной структуры основной части предикатной лексики достаточно использовать всего 6 ролевых указателей. В соответствии с установившейся в англоязычной литературе традицией их можно определить, например, следующим списком: *Agent, Experiencer, Source, Goal, Theme, Instrument (Method/By_means_of)*. См., например, [3].

[5] Для одноместных предикатов ролевой указатель, разумеется, излишен, поскольку он всегда один и тот же - $P\,(x)$ всегда будет читаться одинаково: *объекту* ***x*** *приписывается свойство* ***P.***

3. Логическая интерпретация

Синтаксические связи принято подразделять на «сильные» и «слабые». **Сильные связи** обычно интерпретируются как связи между предикатным термом - именем ситуации (синтаксический хозяин) и термом - актантом, специфицирующим определенную ролевую позицию для указанной ситуации (синтаксический слуга). Функция предлога в такой конструкции (для языков с развитой падежной системой, к каковым, в частности, принадлежит русский язык, - предлога совместно с падежом управляемого слова) - указание ролевой позиции актанта. В языках, не имеющих падежных форм, вся нагрузка по реализации этой функции ложится на предложную систему (и, возможно, порядок слов). Так, фразы *прибыл **в** Москву* и *прибыл **из** Москвы* радикально меняют смысл исключительно благодаря замене предлога (и падежа).

С точки зрения принятой выше схемы интерпретацией собственно предлога (для русского языка – предлога вместе с падежом зависимого слова) будет ролевой указатель ρ_i, а вся предложная синтагма получит интерпретацию вида

$$R\,(\ldots\ \rho_i{:}\ x_i\ \ldots\ )\ \&\ A_i\,(x_i),$$

где слову, управляющему предлогом, сопоставлен предикатный терм ***R,*** а слову, управляемому предлогом – терм A_i, дающий описание актанта x_i. Так что в приведенных выше примерах предлогу *в* будет соответствовать выбор ролевого указателя *Goal,* а предлогу *из* - выбор ролевого указателя *Source.*

Информация о том, какую ролевую позицию **может** маркировать тот или иной предлог, должна, очевидно, присутствовать в семантическом словаре. Там должно быть указано, например, что ролевую позицию *Source* могут маркировать русские предлоги *ИЗ, ОТ,* C_{gen} , английские *FROM, OUT, OUT OF*; ролевую позицию *GOAL* могут маркировать русские предлоги *К,* B_{acc}, HA_{acc}, английские *TO, INTO, TOWARDS, IN, AT, ON, UP, UPON, FOR*; ролевую позицию *INSTR* - русские предлоги HA_{loc}, B_{loc}, *ПОСРЕДСТВОМ,* английский *BY,* и т. д.

Отсутствие такой словарной информации у предлогов *ДЛЯ, У, ПРИ, ПО, ПОД, ЧЕРЕЗ* и др. укажет процедурам анализа, что эти предлоги данным способом интерпретированы быть не могут.

Для слабых связей - независимо от наличия или отсутствия предлога - нужна другая интерпретация, выражаясь точнее, другие интерпретации. Чаще всего это интерпретация, требующая лексикализации синтаксической связи, оформляемой предлогом.

При такой интерпретации различимы следующие смысловые составляющие:

(1) дескрипция ***B(x),*** соответствующая синтаксическому хозяину;

(2) дескрипция ***A(y),*** соответствующая синтаксическому слуге;

(3) подразумеваемое (не имеющее лексического выражения в тексте) отношение ***R***, устанавливаемое между сущностями, указанными референциальными индексами ***x*** и ***y***.

Соответственно, получаем следующую логическую схему интерпретации:

$$A\,(x)\ \&\ B\,(y)\ \&\ R\,(x, y)$$

Выбор «предметного» отношения при такой интерпретации может быть мотивирован поразному. Для связей, маркируемых предлогом, одна из возможных мотивировок - указание отношения самим предлогом. Так, в словосочетаниях

рукопись на столе →
рукопись находится_на столе;

рукопись в столе →
рукопись находится_внутри стола;

рукопись под столом →
рукопись находится_под столом;

именно предлог (для русского - взятый вместе с падежом управляемого слова) определяет выбор подразумеваемого отношения.

Информация о потенциальных возможностях предлога выражать **в определенных контекстах** то или иное предметное отношение также должна присутствовать в словаре.

Помимо рассмотренной интерпретации, где слуге и хозяину соответствует два **разных** референта, иногда возможна и другая интерпретация, где слуге и хозяину соответствует **один и тот же** референт, как, скажем, в словосочетании

банка темного стекла → *БАНКА (x) &*
СТЕКЛЯННЫЙ (x) & ТЕМНОГО_ЦВЕТА (x).

Здесь синтаксическая связь интерпретируется **отношением кореференции** имен - участников синтаксической связи. Как правило, связь, получающая такую интерпретацию, в русском языке беспредложная. Но в некоторых случаях возможно и предложное оформление такого типа связи. Так, смысл указанного выше сочетания может быть выражен и несколько иначе: *банка **из** темного стекла.* Существенно, что далеко не всякий предлог может быть таким образом употреблен: скажем, в конструкциях *банка **за** темным стеклом, банка **под** темным стеклом* о кореференции уже речи быть заведомо не может. Таким образом, следует знать (словарно), какие предлоги потенциально могут, а какие в принципе не могут выступать в этой функции.

---

[6] Для нарративного текста подразумевается еще и общий для всего текста кванторный префикс, состоящий из кванторов существования по всем индивидным переменным.

Этому может соответствовать словарная помета «допускает кореференцию».[7]

Кроме того, следует отдельно рассмотреть и описать такую специальную, но весьма важную для профессиональных текстов функцию предлогов, как оформление и модификация **числовых групп**. Правильная формализация таких конструкций должна опираться на словарную информацию о том, какие аспекты числовых значений и как именно уточняются с помощью предлогов.[8]

И, наконец, возможен случай, когда предлог никак не влияет на интерпретацию связи между соединяемыми им знаменательными словами. Соответственно, можно говорить о **нулевой интерпретации** предлога.

Таким образом, если нас заботит правильная интерпретация предложных связей в деловом тексте, следует предусмотреть при описании предлогов в семантическом словаре ответы на следующие вопросы:

(1) какие роли при предикатном терме может маркировать данный предлог;

(2) может ли он маркировать связь кореференции;

(3) какие лексические (предметные) отношения он может выражать;

(4) на какие ограничения или функции числовых величин он может указывать.

Семантическое описание предлогов, конечно, является языково-зависимым, однако оно должно, по нашему мнению присутствовать именно в концептуальном словаре.

4. Анализ

Рассмотрим очень кратко, как может использоваться эта информация в работе семантического интерпретатора.[9]

Предполагается, что на вход интерпретатора поступает синтаксически размеченный текст, причем в разметке сохраняются все найденные парсером варианты синтаксических связей, в частности, все варианты подчинения предложных групп. В синтаксической разметке также должны быть представлены все отражаемые словарем лексические варианты (концепты) для каждого знаменательного слова. Интерпретатор выполняет перебор и оценку предлагаемых вариантов, выбирая наиболее приемлемый (приемлемые). Таким способом в ходе интерпретации реализуется процесс разрешения лексической и синтаксической неоднозначности. Детальное обсуждение этой проблемы выходит за рамки нашей темы.

Алгоритм интерпретации предложных связей должен опираться как на словарную информацию о самом предлоге, так и – в большей степени - на словарную информацию при отце (*F*) и сыне (*S*) предлога. В рамках нашей темы здесь существенно то, что непосредственным объектом интерпретации являются не слова, а **концепты**, т.е. интерпретируется тройка вида $D^i(W_F)$ --> P --> $D^j(W_S)$, где $D^i(W)$ есть *i*-ый лексический вариант слова *W*.

Работа процедуры интерпретации должна управляться информацией о семантических характеристиках отца и сына (их текущих лексических вариантов). Точнее, возможные способы интерпретации определяются парой

$$< SC(D^i(W_F)), SC(D^j(W_S)) >,$$

где *SC* - семантический класс текущего лексического варианта указанного слова; назовем это **семантическим контекстом** интерпретации предложной связи. Далее – в зависимости от предполагаемого способа интерпретации - может быть затребована более детальная словарная информация об этих же элементах текста. Что касается информации о самом предлоге, то она привлекается скорее для проверки и уточнения гипотез, выдвигаемых исходя из указанной семантической конфигурации; предлог чаще уточняет, чем предопределяет выбор подходящей интерпретации. Если говорить коротко, процедура семантической интерпретации – это разбор случаев, управляемый семантическим контекстом и завершающийся принятием одного из решений, перечисленных в п. 4.

Общая схема анализа может выглядеть следующим образом.

1) Если в позиции W_S обнаружено **число**, и для предлога в словаре указана как возможная функция модификации числовых групп – уточняется тип числового значения (как указано выше.)

2) Если концепту $D^i(W_F)$ в словаре приписана семантическая модель управления (хозяин квалифицируется как предикатный терм), и предлог может маркировать роль ρ_i – проверяется семантическое согласование концепта слуги $D^j(W_S)$ с указанным в модели управления семантическим условием заполнения данной ролевой позиции предиката. Если все

Не всякое употребление предлогов, получивших такую помету, указывает на кореференцию – сравним, напр., *посуда из стекла* и *посуда из Чехии.* Ранее нами была предложена модель, позволяющая алгоритмически различать эти случаи – см. [4] и [5].

[8] Здесь следует различать, по меньшей мере, следующие случаи: (а) предлог указывает, что речь идет о точном числовом значении (*вода кипит при 100 градусах);* (б) предлог - ограничитель числового диапазона сверху (*продажа партиями до 1000 штук*); (в) предлог - ограничитель числового диапазона снизу (*продажа партиями от 1000 штук*); (г) предлог указывает, что речь идет о приблизительно известном числовом значении (*на расстоянии около 100 км);* (д) предлог - возможный указатель кратности изменения значения (*мощность выросла в 2 раза*); (е) предлог - возможный указатель величины изменения значения (*вес уменьшен на 20 кг).*

[9] Изложение в этом разделе опирается на опыт разработки семантического интерпретатора, выполненной под руководством автора.

условия выполнены – устанавливается связь предикат - актант.

3) Если $SC(D^i(W_F)) =$ *'Объект'* и $SC(D^j(W_S)) =$ *'Объект'*, проверяется объемная совместимость концептов отца и сына. Если концепты совместимы – устанавливается связь кореференции.[10]

4) Если в словаре с парой концептов

$$< D^i(W_F), D^j(W_S) >$$

ассоциировано предметное отношение R_{tez}, предложная связь интерпретируется этим отношением – по схеме, указанной в п. 3.

5) Если в словаре с предлогом *P* и семантическим контекстом $<SC_F, SC_S>$ связано предметное отношение R_{prep}, предложная связь интерпретируется отношением R_{prep} по той же схеме.

6) Если ни одна из указанных интерпретаций оказывается невозможной – результат квалифицируется как неудача.

С точки зрения интерпретатора требуемая им функциональность словаря может быть представлена следующим набором функций.

а) Функции, характеризующие знаменательные слова – члены предложной синтагмы:

SC(D) – возвращает семантический класс концепта ***D***;

Extensional_Rel (D_1, D_2) - возвращает одно из значений, характеризующих соотношение объемов для заданных концептов.

Ass_R (D_1, D_2) - возвращает отношение, ассоциированное в словаре с указанной парой концептов

.Функции, характеризующие предлог:

PrepNumValTyp (D_{prep}) – возвращает возможный способ модификации числового значения предлогом; ***PrepRole*** (D_{prep}, Rf) – возвращает ролевой указатель[11];

PrepCoref_IsPossible (D_{prep}) – возвращает {***Yes/No***} («*допускает/не допускает кореференцию*»;

JoinPrepRs $(D_{prep}, SC_S, SC_F, Rf)$ – возвращает список отношений, ассоциированных с указанным предлогом, указанной семантической конфигурацией отца и сына и, возможно, с заданным падежом управляемого слова.

Список литературы:

1) Солоницкий А. В. Проблемы семантики русских первообразных предлогов. – Владивосток: Изд-во Дальневост. Ун-та, 2003. – 126 с.

2) Филипенко М. В. Проблемы описания предлогов в современных лингвистических теориях (обзор) // Исследования по семантике предлогов. Сб. статей. – М.: Русские словари, 2000. – С. 12 – 54.

3) Grimshaw J. Argument Structure. - The MIT Press, Cambridge, Massachusetts, London, England, 1992.

4) Рубашкин В. Ш. О методах анализа связного текста // Вопросы информационной теории и практики. - Вып. 49. - М.: ВИНИТИ. - 1983. - С. 58-73.

5) Рубашкин В. Ш. Представление и анализ смысла в интеллектуальных информационных системах. - М.: Наука, 1989. - 190 с.

[10] Это решение опирается на предложенную нами модель вычисления кореференции – см. [5].

[11] Аргумент ***Rf*** –представляет имя отпредложной синтаксической связи, т.е., падеж управляемого предлогом слова.