МОДЕЛЬНЫЙ ПРИМЕР В АВТОМАТИЧЕСКОМ СЕМАНТИЧЕСКОМ СЛОВАРЕ: К ПОИСКУ ПРОТОТИПА УПОТРЕБЛЕНИЯ ЛЕКСЕМЫ

С.Ю. Семенова

Институт научной информации по общественным наукам РАН,
Российский государственный гуманитарный университет, г. Москва

Обсуждаются вопросы текстового иллюстрирования русского прикладного формализованного семантического словаря РУСЛАН, предназначенного для автоматического анализа текста. В словаре отражается полисемия (с ограничениями прагматического характера), и материал словарных статей, включая иллюстративную зону, нацелен в том числе на дизамбигуацию. Для модельных примеров, вводимых в словарь наряду с цитатами из Национального корпуса русского языка, указанная цель обусловливает стремление показать наиболее общие, прототипические употребления лексем. Кроме того, составление прототипического примера представляется полезным и с когнитивной точки зрения. Для уточнения интуитивного представления о прототипическом контексте рассматривается ряд типов минимальных модельных примеров в Малом академическом словаре (МАС). Показано, что структура и лексический состав модельных примеров зависит не только от свойств вокабулы и денотата, но и от структуры словарной статьи, поэтому не все типы примеров МАС, представляющихся удачными с точки зрения отражения прототипа, могли бы подойти для словаря РУСЛАН с его более детализированным форматом.

Ключевые слова: семантический словарь для автоматического анализа текста, данные для дизамбигуации, модельный пример в словаре, минимальный контекст лексической единицы, прототипическое употребление слова.

В данной работе тема понимания затрагивается в двух аспектах. С одной стороны, широкой прикладной задачей исследования является построение компьютерного словаря для автоматического анализа (в т.ч. понимания) текста. С другой стороны, в процессе текстового иллюстрирования словаря возник вопрос о подборе таких модельных примеров, которые как можно ближе соответствовали бы прототипическому употреблению вокабулы, и в этом отношении речь идет о нацеленности на понимание того, каков может быть прототип употребления слова.

Более конкретно. В настоящее время продолжается разработка русского прикладного семантического словаря РУСЛАН, предназначенного для автоматического анализа текста. Словарные описания, отражающие семантические, а также грамматические, сочетаемостные, тезаурусные и некоторые энциклопедические свойства слова, строятся на формальном метаязыке, предоставляющем лексикографу довольно разнообразные средства кодирования. Создателем метаязыка, как и всей концепции словаря, является Н.Н. Леонтьева. Первые относительно завершенные версии словаря были получены под ее руководством (и при участии ряда других специалистов, в том числе автора) к середине 2000-х гг.; см., например, [Леонтьева 2001, 2006 и др.], [Леонтьева, Семенова, 2002, 2003].

Наряду с другими (формализованными) зонами и полями заполнялось свободное поле иллюстраций; в основном оно содержало модельные примеры. Фактически первые версии словаря создавались еще в «докорпусную» эпоху (по крайней мере, для отечественной лексикографической традиции), и основным средством иллюстрирования служил модельный пример (хотя отчасти в это поле вносились также литературные цитаты и сокращенные цитаты из конкорданса, построенного на определенном массиве деловых официальных текстов) [Семенова 2003].

С 2017 г. под руководством автора начались работы по модернизации словаря, и при этом существенное внимание уделено расширению иллюстративного материала, теперь уже – с существенной опорой на корпусные данные. Зона иллюстраций ныне структурирована, она включает ряд новых полей для корпусных иллюстраций (цитаты подбираются, главным образом, в Национальном корпусе русского языка). В круг новых полей входят: поле для корпусных примеров общего характера; поля для иллюстрирования каждой валентности заглавного слова; поле для показа в тексте фразеологических единиц с заглавным словом, а также поле для нестандартных корпусных примеров, полезных для последующего уточнения словарных описаний. Кроме этого, в формируемой ныне версии сохранено поле модельных иллюстраций (ИЛЛ_МОД) [Семенова 2017].

В словаре (и в первых его версиях, и сейчас) отражается полисемия; описания отдельных лексем полисемичных слов располагаются в отдельных словарных статьях. В силу потенциальных сложностей дизамбигуации разделение слов на лексемы является укрупненным (по сравнению с традиционными современными толковыми словарями): принято эмпирическое ограничение — не более 5 лексем на слово; при этом в целях дизамбигуации словарные статьи строятся так, чтобы по возможности снабдить описания разных лексем различающимися формальными контекстными признаками [Леонтьева, Семенова 2002].

Иллюстративная работа (как подбор корпусных, так и составление модельных примеров) также в существенной мере нацелена на применение словарных данных (включая иллюстративный материал) при автоматическом распознавании лексем полисемичных слов. Для модельных примеров (обычно кратких, часто минимальных) это связано со стремлением отразить прототип (или ряд прототипов) употребления лексемы-вокабулы.

Определенным ориентиром при осмыслении того, какой контекст лексической единицы следует считать прототипическим, могут служить модельные примеры в Малом академическом словаре (МАС; Словарь русского языка: В 4-х т. / АН СССР, Ин-т рус. яз.; Под ред. А. П. Евгеньевой. – 2-е изд., испр. и доп. М.: Русский язык, 1981—1984).), который является одним из наиболее авторитетных русских толковых словарей. В этом словаре представлены два сорта примеров — модельные и литературные.

Данная работа написана в развитие [Семенова 2021], где приведены некоторые различающиеся по содержательным и формальным признакам типы минимальных модельных примеров в МАС, а также отмечен ряд

случаев, когда модельные примеры отсутствуют, и их отсутствие, по-видимому, мотивировано затруднениями с определением прототипа.

Остановимся на модельных примерах в МАС более подробно. В самом деле, среди искусственных минимальных контекстов (а в этом словаре модельные примеры имеют форму именно минимальных контекстов) можно выделить некоторые черты, характерные для прототипических примеров как на содержательном, так и на собственно языковом уровнях (последнее связано с выбором структурных и лексических средств).

Так, с точки зрения содержания одной из стратегий иллюстрирования (на наш взгляд, вполне оправданной) является стремление к охвату наиболее общих, базисных аспектов обозначаемой сущности. Показательным примером (отмеченным также в [Семенова 2021]) может служить небольшая группа модельных контекстов исходной лексемы существительного *слово* (в значении «основная единица языка»): *Значение слова. Порядок слов в предложении. Русские и иностранные слова.* Эти примеры затрагивают сферы, важные для денотата – семантику, синтаксис и этнический аспект языка. С позиции обобщенного, схематичного отражения «картины мира» интересна, в частности, цепочка примеров в МАС для исходного значения глагола *работать*: *Работать на заводе. Работать в школе. Работать в колхозе*. Эта последовательность соответствует официальной (хрестоматийной для отечественного дискурса середины XX века) схеме советского общества: рабочий класс — интеллигенция («прослойка») — колхозное крестьянство. Также раскрытие базовых свойств денотата может сводиться, например, к указанию основных типов вещества или материала, из которого он может состоять (в частности, быть изготовленным); таковы примеры *Газовый шарф. Мохеровый шарф* для имени артефакта *шарф* или *Лист бумаги. Лист картона. Стальной лист* для соответствующей «артефактной» лексемы слова *лист.*

С точки зрения синтаксиса естественным путем иллюстрирования минимальными контекстами является «экскурс» по валентностям слова; ср. *Прививка плодовых растений. Предохранительные прививки против брюшного тифа* (для слова *прививка*). Часто иллюстрирование по валентностям сводится к показу заполнения главной, информационно наиболее значимой валентности. Для целого ряда существительных это валентность объекта, насыщаемая генитивной группой: *Директор завода. Директор театра* (статья слова *директор*). В ряде случаев осуществляется показ грамматически различающихся способов насыщения «главной» валентности: *Дать книгу. Дать денег на дорогу. Дать хлеба* для глагола *дать* (в значении «вручить»); т.е. здесь приводятся варианты модели управления с аккузативом и партитивом. Иногда для иллюстрирования используется сочинительная конструкция с клишированным лексическим составом: *Инженеры и техники* в статье слова *техник.* Среди модельных примеров встречаются дефисные формы: *техник-электрик* (также для имени *техник*) или *слесарь-инструментальщик* (для имени *слесарь*).

С точки зрения лексики минимального контекста — для примеров МАС в целом характерен выбор некоторых привычных (в том числе, в рамках дискурса своего времени), широкоупотребительных, немаркированных лексических средств (что следует, на наш взгляд, считать целесообразным в плане отражения прототипа): *Победить врага. Победить в бою* для первого значения глагола *победить* («нанести поражение…»). Для минимального контекста выбирается частотная, стилистически нейтральная, в определенной мере даже банальная лексика (отметим, что лексическое наполнение корпусных примеров, наоборот, часто бывает «экзотическим» и весьма далеким от прототипа [Семенова 2021]). У имени *слово* лексический состав приведенных выше примеров из МАС соответствуют общепонятному, «школьному» варианту метаязыка лингвистики (при этом можно было бы предложить и некоторое редактирование: быть может, сочетание *Значение слова* стоило бы заменить сочетанием *Смысл слова*, так как слово *значение* в большей мере полисемично, чем слово *смысл*). Очевидно, из стремления к отвлеченному прототипу иллюстраторы МАС стараются избегать ономастики, но примеры с собственными именами все же встречаются (когда такие имена ассоциируются с самим образом денотата заглавного слова): *Доминиканский орден* для имени *орден* в значении «религиозная община» или *Орден Ленина* для того же слова *орден* в значении «награда».

Не все способы иллюстрирования, применяемые в МАС, приемлемы для прикладного словаря РУСЛАН – в том числе, в силу более разветвленной структуры последнего. Так, ряд иллюстраций МАС «укладывается» в поля основной части словарной статьи РУСЛАНа. Таковы, в частности, контекстные лексические функции, а также термины и иные устойчивые словосочетания с заглавным словом. В самом деле, зона лексической сочетаемости РУСЛАНа включает: поле ЛФ (контекстных лексических функций, соответствующих, в основном, Модели «Смысл-Текст»); поле ТЕРМ (терминологических словосочетаний с заглавным словом); поле СЛСЧ (устойчивых словосочетаний, не удовлетворяющих или в меньшей степени, чем явные термины, удовлетворяющих условиям терминологичности). И ряд прототипических контекстов МАС в словаре РУСЛАН мог бы быть помещен в эти поля, то есть в «тело» словарной статьи, а не в иллюстративную зону. Например, контексты МАС для первого значения слова *лес* («деревья, стоящие на корню») *Дремучий лес. Хвойный лес* могли быть описаны в РУСЛАНе, соответственно, как лексическая функция Magn и как термин:

ЛФ (*лес 1*) = Magn (*дремучий*);
ТЕРМ (*лес 1*) = *хвойный лес*.

Отметим, что пока в РУСЛАНе нет специального поля для термина-гипонима (сочетание *хвойный лес* естественно трактовать как гипоним термина *лес*), но, возможно, такое поле будет добавлено. Вообще, гипонимы в качестве модельных контекстов представлены в МАС довольно широко, например, *Электрическая лампа. Керосиновая лампа. Настольная лампа* для лексемы *лампа* в значении «осветительный прибор».

В связи с сопоставлением форматов двух рассматриваемых словарей можно назвать и отдельные неудачные примеры в МАС (которые для данного академического словаря являются редкостью). Это примеры со словом *акустика*: *Хорошая акустика* и *Плохая акустика*. В самом деле, как выразители аксиологических оценок прилагательные *хороший* и *плохой* слишком тривиальны. В словаре РУСЛАН их можно было бы описать как значения контекстных ЛФ Bon и AntiBon, но размещение столь неконкретных и обладающих столь широкой сочетаемостью контекстов в РУСЛАНе не практикуется. (Напротив, удачным примером аксиологической оценки в МАС представляется сочетание *Правильный прикус* для имени *прикус*: прилагательное *правильный* характеризуется более избирательной сочетаемостью, чем *хороший*). Можно привести и еще один не вполне удачный модельный пример МАС: *Работник прилавка* в статье существительного *прилавок*. Ведь сочетание *работник прилавка* являет собой «чистый» пример фразеологизма, обладающего вполне определенным значением («продавец»), и как в МАС, так и в словаре РУСЛАН оно могло бы войти во фразеологическую зону; в РУСЛАНе оно могло быть представлено как: СЛСЧ (*прилавок*) = *работник прилавка.*

В целом представляется, что модельный пример, отражающий прототип, должен содержать конкретику, притом более свободную, чем, скажем, во фразеологизмах или синтагматических лексических функциях, но в то же время не выходящую из центральной части некоего обобщенного образа денотата в сознании носителя языка.

В МАС не предусмотрено отдельное пространство для лексических функций, и потому контексты такого рода часто получают статус всего лишь модельных примеров. В РУСЛАНе такое пространство есть, но имеются и свои проблемы с лексическими функциями: назрела необходимость в уточнении их состава и правил идентификации. Что касается соотношения минимального прототипического модельного примера и фразеологизма, то грань между ними (как в теории, так и в практической лексикографической работе) не всегда вырисовывается четко. Можно говорить о конкуренции некоторых полей (или зон) словарной статьи (для РУСЛАНа это конкуренция полей ЛФ vs ИЛЛ_МОД; ТЕРМ vs ИЛЛ_МОД; СЛСЧ vs ИЛЛ_МОД). И поле ИЛЛ_МОД обычно проигрывает другим указанным полям; как неструктурированное, оно играет роль остаточного.

Необходимо отметить, что модельные примеры в МАС, при всей их когнитивной и стилистической ценности, все же факультативны (при практической обязательности примеров литературных). В этой связи интересно посмотреть случаи, когда модельный пример отсутствует. Представляется, что отказ от модельного примера как раз связан с трудностью мотивированного определения, очерчивания прототипа.

Так, модельные примеры не приводятся для ряда глаголов с валентностью содержания *мечтать, нравиться* и нек. др.; в самом деле, мечтать можно много о чем, и нравиться может многое, и лексикографы МАС вполне оправданно предпочли для таких глаголов отказаться от

детализаций. Также не снабжены модельными примерами многие предметные существительные, тяготеющие к номенклатурной лексике (и в силу этого ассоциирующиеся с отраслевой конкретикой, от которой МАС, очевидно, стремится дистанцироваться). Это, скажем, имена «рядовых» профессий: *программист*, *портниха*, *шахтер* и др.; большинство названий растений: *клубника*, *шалфей* и др.; названия ряда устройств и приспособлений: *лафет*, *штурвал*, *клюшка* и др. (Правда, среди указанных семантических классов есть и исключения, когда слово, наоборот, снабжено модельным примером, т.е. когда в силу разной мотивировки тому или иному возможному отраслевому контексту дается приоритет: *Оператор прокатного стана* в статье слова *оператор*, *Черная смородина* в статье слова *смородина*, *Мотор самолета* в статье *мотор* и др.).

Как уже отмечено, в прикладном словаре, нацеленном на автоматическое понимание текста (каковым является РУСЛАН), актуальна задача обеспечить материал для дизамбигуации. В этой связи в МАС интересно обратить внимание на примеры, когда разные лексемы многозначного слова снабжены ощутимо, в том числе тематически различающимися модельными контекстами. Например, слову *орган* в значении «часть живого организма» поставлены в соответствие примеры *Органы слуха. Органы кровообращения. Органы речи,* а тому же слову в значении «учреждение» — *Органы здравоохранения, Финансовые органы, Органы народного образования*. Или, к статье существительного *образование* предложены примеры: *Образование государства. Образование водяных паров* (для деривата S_0 глаголов *образовать/образоваться)*; *Горные образования. Жировые образования* (для лексемы в предметном, результативном значении S_{res}); *Право на образование. Народное образование* для лексемы, обозначающей сферу педагогики. Необходимо при этом заметить, что МАС, как традиционный толковый словарь, адресованный читателю, не нацелен на автоматическое распознавание значений, и нередки случаи, когда модельные примеры есть не у всех лексем многозначных слов. В прикладном РУСЛАНе при описании многозначных (полисемичных) слов модельные примеры желательны для всех лексем таких слов, и потому в нем допустимы не только действительно удачные прототипические примеры, но и примеры более аморфные — желательно, чтобы поле ИЛЛ_МОД не было пустым, даже если есть затруднения с отражением прототипа для той или иной лексемы.

В МАС, как уже отмечено, модельные примеры имеют форму минимальных контекстов. В словаре РУСЛАН требования к синтаксису примеров менее жесткие: допустимы и целостные высказывания (в том числе, некоторые, унаследованные от прежних версий данного словаря). Правда, составление (а также редактирование прежних) целостных высказываний требует и дополнительных методических решений, связанных со стилистикой, бытийной сферой, неизбежным увеличением балластной составляющей таких высказываний.

ЛИТЕРАТУРА

Леонтьева Н.Н. К теории автоматического понимания текста. Ч. 2. Семантические словари: состав, структура, методика создания. М.: Изд. Моск. ун-та, 2001.

Леонтьева Н.Н. Автоматическое понимание текстов: системы, модели, ресурсы: Учебное пособие. М.: Академия, 2006.

Леонтьева Н.Н., Семенова С.Ю. Об отражении полисемии в прикладном семантическом словаре. Компьютерная лингвистика и интеллектуальные технологии. Труды международного семинара ДИАЛОГ-2002. М., 2002. Т.2. С. 489-496.

Леонтьева Н.Н., Семенова С.Ю. Семантический словарь РУСЛАН как инструмент компьютерного понимания // Понимание в Коммуникации. Материалы научно-практической конференции. 5—6 марта 2003 г. М., МГГИИ, 2003.С.41-46.

Семенова С.Ю. Примеры в компьютерном семантическом словаре: некоторые наблюдения над процессом подбора // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции Диалог'2003. М, 2003. С. 593-598.

Семенова С.Ю. Об использовании данных Национального корпуса русского языка для иллюстрирования статей компьютерного семантического словаря // Труды международной конференции «Корпусная лингвистика — 2017». СПб., 2017. С. 321-324.

Семенова С.Ю. Модельный пример в словаре как мини-текст (2021, в печати).

Artificial textual example in applied semantic dictionary for NLP: towards the prototypical contexts fixation

Semenova S.Yu

The paper deals with the strategy of artificial textual illustrations compiling in a semantic dictionary for NLP. Namely, the Russian formal semantic dictionary of RUSLAN is considered. The dictionary represents polysemy, and the entries as well as the illustrative fields are aimed at disambiguation. That's why the artificial examples should correspond to the typical use of the lexical unit to the greatest extent. The preliminary notion of prototypical context is discussed. A number of minimal artificial examples in Minor Academic dictionary («Maly'j Akademicheskij Slovar'») are considered from the point of view of the prototype representation. It is found out that the proper artificial examples in this famous explanatory dictionary have some special syntactic, lexical, and ontological features. A number of the example types could be placed in the main part of the RUSLAN entry because of the more detailed structure of the latter.

Keywords: semantic dictionary for NLP, artificial textual illustrations in a dictionary, choice of proper minimal context of a word, data for disambiguation.

Сведения об авторе

Семенова Софья Юльевна, кандидат филологических наук, Институт научной информации (ИНИОН) РАН, старший научный сотрудник, Российский государственный гуманитарный университет (РГГУ), доцент, электронный адрес: sonya_sem@mail.ru