ПРИМЕРЫ В КОМПЬЮТЕРНОМ СЕМАНТИЧЕСКОМ СЛОВАРЕ: НЕКОТОРЫЕ НАБЛЮДЕНИЯ НАД ПРОЦЕССОМ ПОДБОРА

С. Ю. Семенова

(Работа выполняется при поддержке Российского гуманитарного научного фонда; проекты РГНФ-01-04-16252а и РГНФ-02-04-00294а)

Естественной частью статей словаря, описывающего семантику лексических единиц, являются текстовые иллюстрации. Роль этого вида словарной информации, казалось бы, простейшего (по сравнению, например, с толкованиями) весьма существенна для итогового качества словаря. Помимо основной функции - сделать

словарные описания более понятными и наглядными для читателя-пользователя, примеры, очевидно, помогают самому лексикографу точнее построить эти описания, дают материал для пополнения словников (т.е. действует так называемая "обратная связь"). Составом и стилевыми характеристиками примеров во многом определяется информационная значимость словарного продукта и его культурное "лицо". Кроме того, если словарь используется в автоматической обработке текста (АОТ), его иллюстративный ресурс может быть задействован при реализации алгоритмов, основанных на аналогиях, например, при автоматическом разрешении многозначности.

При всем этом, процессу наполнения словаря иллюстрациями, весьма трудоемкому (несмотря на кажущуюся простоту) в литературе по теоретической лексикографии не уделяется, насколько нам известно, значительного внимания (например, по сравнению с такими задачами, как подбор словников или выделение и упорядочение лексем). В данной работе на примере конкретного словарного материала делается попытка некоторого осмысления опыта иллюстраторской работы, ее привычных приемов, а также трудностей, "подводных камней". Цель данного рассмотрения - "нащупать" пути для систематизации этой деятельности в дальнейшем. Анализируется опыт работы автора над полями текстовых иллюстраций в двух компьютерных словарных системах: в АОТ-ориентированном семантическом словаре РУСЛАН, содержащем сведения о подробной таксономии, актантах, лексической сочетаемости, тезаурусных связях и нек. др. свойствах слов [4-6] и во фрагментах базы данных "Русский глагол" системы "Лексикограф"; эта база данных предназначена для размещения структурированных элементов толкования слов и для описания параметров их лексического значения [3,7].

Работа над иллюстрациями в этих словарях не является законченной; скорее мы сейчас находимся где-то ближе к ее началу; она совершается и вместе со вводом новой лексики, и при редактировании и дополнении ранее введенных словарных описаний.

С одной стороны, примеры для данных словарей черпаются из текстовых источников, в том числе электронных: нормативные документы федеральных органов власти, изданные в начале 1990-х годов и имеющиеся на машиночитаемом носителе, а также публицистика, художественная литература, материалы в Интернете, известные "бумажные" словари (такие, например, как МАС); с другой стороны, в реальной практике, когда требуется более или менее быстрое наращивание словарного массива, невозможно обойтись и без подручного "школьного" метода - сочинения примеров.

Оба способа - поиск и придумывание - имеют, очевидно, свои достоинства и недостатки.

Ценность примеров, найденных в источниках, обусловлена тем, что, как правило, они свидетельствуют о нормативности контекста, о том, что "так говорят". (Конечно, есть исключения: даже тексты литературных произведений не гарантированы от речевых ошибок, не говоря о современной прессе. Но все же тексты, взятые из источников, вызывают, прежде всего, у самого лексикографа, большее доверие, чем придуманные). Кроме того, примеры из литературных произведений, особенно классических, несомненно, украшают и "облагораживают" словарь.

С другой стороны, даже при использовании современных поисковых средств, в источниках порой трудно бывает найти иллюстрации именно тех свойств, которые лексикограф стремится подчеркнуть в описании; например, в реальных литературных примерах далеко не всегда бывают заполнены все теоретически возможные валентности слов. Ситуации, изображенные в литературных примерах, бывают подчас далеки от прототипических для описываемого слова (например, из-за того, что они "выдернуты" из контекста, или, скажем, в силу естественного стремления художников слова к оригинальности произведения). Так, можно ли считать прототипическим для имени *библиотекарь* следующий почти наугад взятый из повести контекст:

"*Библиотекарь, с которым [Бейль] часто просиживал у Флориана или*

Гвадри, встречает его возгласом и поздравлениями, засыпает его *вопросами о Франции."* (А.Виноградов)? Ведь в описании ситуации нет ни библиотеки, ни книг, ни других признаков привычной для этой профессии обстановки. Найденные примеры к тому же зачастую приходится подвергать редактированию: отсекать части слишком длинных предложений, вставлять референты местоимений, убирать лишнюю фактографию (цифры, фамилии и пр.), осовременивать орфографию. (В качестве иллюстрации такого редактирования авторского текста приведем "исправленную" пушкинскую фразу, которую было решено использовать в статье слова *цыган* в РУСЛАНе:

Цыгане [а не *цыганы*, как в оригинале] *шумной* [а не *шумною*] *толпой по Бессарабии кочуют*).

Искусственные модельные примеры можно придумать и без поисковых средств; обычно за такими примерами не приходится "ходить далеко" или "лезть в карман" (хотя перед занесением в словарь придуманного примера не мешало бы заглянуть, скажем, в конкордансы или в Интернет, чтобы пример был корректным, усредненным по каким-то реальным контекстам). На модельных примерах нетрудно бывает пояснить свойства, отражаемые в описании, например, заполнить все валентности (вспомним хотя бы известный пример с "пятивалентным" глаголом *командировать* в [1]), а также показать прототипическую таксономию участников и т.п. (Кстати, в обоих словарях предусмотрены места для записи минимального контекста лексемы, каковым, в частности, является контекст отдельно взятого i-ого участника ситуации. В глагольной базе "Лексикографа" такой контекст считается фрагментом заголовка словарной статьи; в РУСЛАНе для этого отведены поля ИЛЛ1, ИЛЛ2, ... Если слово со своим участником образует более или менее лексикализованное сочетание, то участник в РУСЛАНе может быть внесен и в поле так называемых актантных лексических функций: АЛФ1("*экологически*") = *чистый* [4].)

Недостатком сконструированных примеров является некоторая искусственность, "подогнанность" иллюстративного материала под словарное описание, а иногда даже отсутствие уверенности в нормативности примера, ведь лексикограф иногда рискует, сознательно или бессознательно, сочиняя контексты, как бы идет по грани языковой нормы. Плохо также то, что при преобладании таких примеров семантический словарь приобретает малоакадемичный, "кустарный" облик.

Упорядочения, определенной унификации требуют оба способа получения иллюстраций, тем более, что большинство современных словарных систем, и два данных словаря в частности, являются плодами коллективного творчества, и разница в авторских подходах к заполнению данной словарной зоны бывает весьма заметной при переходе от статьи к статье.

Рассмотрим некоторые конкретные моменты иллюстраторской работы.

В качестве одной из основных целей такой работы над обоими словарями выступает демонстрация способов насыщения валентностей. С реализации этой процедуры и надлежит, по-видимому, начинать иллюстрирование, так как заполненные должным образом валентности сразу "погружают" лексему в характерную ситуацию, что удобно и потенциальному читателю словаря (а мы надеемся, что даже АОТ-ориентированный РУСЛАН может быть читаем человеком), и разработчику.

Наиболее просто дело обстоит с иллюстрированием описаний одноактантных слов, да еще и если участник грамматически однороден. Таково, в частности, большинство прилагательных, имеющих пассивную валентность на определяемое имя. В РУСЛАНе в словарных статьях прилагательных обычно бывает достаточно раскрыть типовые таксономические классы участника, перечисленные в описании, а именно, в поле семантических характеристик 1-го актанта [8]:

СХ1 ("*контрольный*") = ДЕЙСТВИЕ U ВЕЛИЧИНА U ОРГАНИЗАЦИЯ U УСТРОЙСТВО U НОС_ИНФ;

ИЛЛ ("контрольный") = *контрольное считывание информации;*

контрольная сумма; контрольный разряд;

контрольный протокол

/все четыре примера - реальные, из области информатики/;

контрольная комиссия при партийном аппарате;

контрольно-счетные органы РФ;

контрольные весы; акустические контрольные системы;

контрольно-измерительное оборудование.

Отметим попутно, что часть иллюстративного материала, наиболее лексикализованная, может отражаться в других полях данного словаря, например, в поле устойчивых словосочетаний:

СЛСЧ ("*контрольный*") = *контрольная работа* /по физике/;

контрольный пакет акций;

контрольный выстрел в затылок и т.п.

На примерах видно также, что иллюстрируется не только слово-заголовок, но и некоторые прагматически важные его дериваты, если они не описываются в данном словаре в самостоятельных статьях: *контрольно-счетный, контрольно-измерительный.* Если валентностей более одной, или даже если она единственная, но по-разному выражается грамматически, то иллюстрирование в РУСЛАНе становится комбинаторным; в примерах должны быть отражены различные допустимые варианты сочетания таксономии участника и поверхностной модели его присоединения:

ИЛЛ ("*очень*") = *Чеченская война длится очень долго;*

Переход через ущелье оказался очень долгим;

Читать в полутьме очень вредно;

Пенсионеры, вынужденные работать, очень устают;

Если очень постараться, то можно добиться известности и т.п.

(варьируется часть речи и таксономия участника, в данных примерах - глагольного).

Вот пример прилагательного, обладающего собственными активными валентностями:

ИЛЛ ("*известный*") = *политик, известный [каждому телезрителю] своим популизмом;*

С.Сорокина известна телеаудитории по многим публицистическим передачам;

Каждому известны анекдоты про жителя Севера и т.п. (здесь варьируются таксономия определяемой группы, наличие-отсутствие участников "источник" /откуда известен/ и "аспект"/чем именно известен/).

Наиболее выразительно комбинаторность иллюстраций проявляется в РУСЛАНе для глагольной [многоактантной] лексики, когда примеры отражают разнообразные [и не структурированные пока в этом словаре] метонимические и залоговые сдвиги (ср. с обрисованной ниже несколько другой ситуацией в системе

"Лексикограф"). Конечно, показать все сочетания типовой таксономии с моделями управления при быстром пополнении АОТ-словаря малореально, это ведет к бессмысленному "комбинаторному взрыву" в зоне иллюстраций, и потому актуально стоит (еще не проработанный пока) вопрос об отборе прагматически значимых вариантов.

Практика показывает, что при конструировании примеров подчас самым трудным для "красивой" текстовой интерпретации оказывается синтаксически стандартный и, казалось бы, простейший первый участник, будь то прототипический агенс глагола или, скажем, идентификатор - имя собственное - для нарицательной одушевленной лексики. Его словесное воплощение, быть может в наибольшей степени по сравнению с другими, требует от лексикографа эрудиции и вкуса.

В самом деле, какое имя собственное уместно использовать, например, в контексте имени деятеля? Если описываются имена из таких "публичных" сфер, как политика, литература, культура, спорт, история и нек. др., то вполне естественно будет привести имена реальных лиц: иракский лидер Саддам Хусейн, великий русский поэт А.С.Пушкин, чемпион мира по фигурному катанию Е.Плющенко, советский контрразведчик Павел Анатольевич Судоплатов и т.п. А как быть с гораздо менее публичными названиями массовых профессий и должностей, работников различных отраслей хозяйства, с именами аксиологических оценок, исполнителей тех или иных социальных функций и именами ряда других классов (слесарь, железнодорожник, лентяй, заявитель, верующий, язвенник и др.)? Не привлекать же на помощь "самого лингвистичного человека" - "Васю" (или "дядю Васю"; вспомним известное своей банальностью сочетание "слесарь дядя Вася")! И не прибегать же к помощи местоимений, свидетельствующих о скудости фантазии лексикографа и делающих ситуацию обезличенной, а пример - практически бесполезным: он лентяй, мой сосед - железнодорожник, этот человек выступил в роли заявителя и т.п.!

В массе случаев приходится "стыдливо" опускать первый актант и приводить в словаре контексты типа работать слесарем в автомеханическом цехе, эндокринолог районной поликлиники, мастер по станкам с числовым программным управлением и т.п., и это, по-видимому, приемлемо при иллюстрировании слов, для которых валентность на персону не является единственной. Таковы большинство имен должностей и профессий (регулярно имеющих две валентности - на персону и на "уточнитель" - название организации или наименование "обрабатываемой" сущности; методика их словарного описания излагается в [6]), таковы и широкие круги глагольной лексики, в частности, обозначающей агентивные действия. Так, для иллюстрации одного из употреблений глагола обсудить можно использовать обезличенный инфинитивный вариант обсудить диссертацию на семинаре отдела. Подобный контекст, "прилагаемый" к слову, содержит иллюстративный материал, по крайней мере, по другим валентностям, для которых конкретика выражается легче, чем для валентности субъекта. (Для иллюстрирования глагольной лексики можно использовать и различные нарицательные наименования человека: Армянская семья бежала в Москву из Баку, когда там начались погромы, или Бастующие рабочие перекрыли железнодорожное полотно, или Пожилой человек теперь не может один перейти Садовое кольцо и т.п.).

Другой выход из затруднительного положения при иллюстрировании одушевленной лексики, особенно одновалентной - это прибегание к контекстам, не заполняющим валентности имени, но показывающих прототипическую ситуацию для обозначаемого деятеля: В состав экспертной комиссии по анализу причин аварии на подводной лодке были включены кораблестроители; Лентяя трудно застать за работой; Либералы в Думе проголосовали за жилищно-коммунальную реформу и т.п., когда другие участники или сам предикат приводимой ситуации связаны с именем деятеля устойчивыми ассоциативными связями (кораблестроитель – подводная лодка, лентяй - работа /хоть и противопоставление, но все-таки связь!/, либерал - реформа и т.п.); ср. непрототипичный пример со словом библиотекарь в начале статьи, когда подобные связи отсутствуют.

Подобный сорт "невалентных" иллюстраций применим и для слов, обладающих более чем одной валентностью, ср.: Токарь за смену выточил X деталей (вспомним типовые школьные задачи по арифметике, составители которых, как и лексикографы, сталкиваются с необходимостью подбора правдоподобных контекстов).

Нам трудно пока указать конкретные типы "хороших" иллюстрирующих ситуаций для конкретных лексических классов и подклассов, исчислить критерии релевантности, но можно надеяться, что со временем определенные рекомендации в этом плане могут быть выработаны.

В работе над РУСЛАНом, нацеленным прежде всего на АОТ, внимание к полю иллюстраций обусловлено не только тем, что эстетичные и убедительные примеры призваны сделать словарь более наглядным для самих разработчиков и других потенциальных "человеческих" читателей, но и тем, что позволят в будущем организовать обработку текста по аналогиям (так называемая example-based technology), что полезно, например, при автоматическом различении полисемии и омонимии.

В примерах, в частности, приводятся типовые представители таксономических классов участников (желательно непересекающихся у разных выделяемых лексем [5]):

СХ1 ("короткий 1") = ПРЕДМЕТ U ПРОТЯЖ[енность] U ПРОСТР[анственность] U (ДЕЙСТВИЕ & СВЯЗАНО & ПРОСТР);

ИЛЛ ("короткий 1") = короткий брусок; короткая длина; бег на короткие дистанции; короткий прыжок;

СХ1 ("короткий 2") = ВРЕМЯ U ЯВЛЕНИЕ U ДЕЙСТВИЕ U ПРОЦЕСС U ИНФОРМАЦИЯ U НОС_ИНФ U КОММУНИК[ативное];

ИЛЛ ("короткий 2") = короткий день; короткий сигнал;

короткий сон; короткий разговор; короткое письмо.

При описании многозначности особенно важно наличие таких примеров, в которых каждая выделяемая лексема погружена в прототипический, "родной" для нее и максимально информативный контекст, где, в частности, должно быть минимизировано количество стоп-слов и следует по возможности избегать окказиональностей, выбивающихся за рамки прототипа и делающих текст маркированным (опять см. выше "плохой" пример с именем библиотекарь).

Отметим, что местоименные слова не всегда малоинформативны; они могут, например, нести полезную для распознавания многозначности грамматическую нагрузку; ср.: Кто является аудиторией Жириновского? и На каком этаже находится главная лекционная аудитория? (противопоставляются одушевленная аудитория - публика и неодушевленная аудитория - комната).

Разнообразие текстового материала, который можно привлечь для иллюстрирования словарных статей, в принципе, требует ранжирования примеров (что пока только в проекте). В дальнейшем текстовые образцы могут быть разделены на ряд типов (по крайней мере, на два типа). В одно, обязательное, поле могут быть занесены "хорошие"

примеры, отвечающие определенным методическим критериям (иллюстрирующие нужные валентности и лексическую сочетаемость, лаконичные, четко обозначающие различия в употреблении лексем многозначных слов), и эти примеры должны получить высокий ранг. В другое поле (или в поля, если их несколько) - факультативные примеры, по большей части, периферийные, дающие лексикографу "информацию к размышлению" и к дальнейшей детализации словарных описаний. Ранг таких примеров должен быть снижен, и в соответствующие поля алгоритм распознавания значений "заглядывать" не должен.

В системе "Лексикограф", в ее глагольной базе данных, где у глагола обычно выделяется значительное число типов употребления и, как правило, изменения таких параметров лексического значения, как 'Тематический класс', 'Таксономическая категория обозначаемой ситуации' ('агентивное действие', 'состояние', 'происшествие' и нек. др., со своими внутренними градациями), 'Диатеза' (т.е. конкретный вариант экспликации участников ситуации), 'Таксономия участников' [7 и др.] фиксируются в отдельных записях базы, трудности иллюстрирования связаны, в основном, с необходимостью отнести конкретный литературный (а подчас и придуманный) пример к тому или иному [узкому] типу употребления. (На несоответствие между литературным примером и толкованием лексемы, нередкое в реальной лексикографической практике, обращено внимание, в частности, в [2]).

Неоднозначные интерпретации возможны по каждому из указанных параметров. Например, по таксономической категории глагола известное высказывание В.И.Ленина Декабристы разбудили Герцена скорее всего, следует отнести к абстрактно трактуемым 'происшествиям' (когда Герцен узнал о деятельности декабристов, в его душе произошел перелом), но не исключена и интерпретация 'агентивное действие; конатив', поскольку декабристы, в числе прочих целей, ставили цель "разбудить" последующие поколения граждан и прилагали усилия к тому, чтобы в будущем "из искры возгорелось пламя".

Нечеткость примеров, их переходный характер наблюдается и при иллюстрировании таксономии участников; так, при мелком дроблении употреблений иллюстратору приходится решать, одинаково ли "бегут" ручьи по асфальту (континуальные сущности, относящиеся к концептам "масса" и "бесконечная полоса") и облака по небу (сущности, чаще мыслимые как дискретные, тяготеющие к форме неправильного овала).

При иллюстрировании диатетических сдвигов не всегда ясно, например, являются ли конкретные вариации в выражении валентностей, связанные со спецификой обозначаемой ситуации и со способом ее видения и не содержащие грамматических запретов на тех или иных участников, свидетельством принадлежности примеров к одной или к разным диатезам. Так, наличие участника "источник" и отсутствие участника "адресат" в первом из контекстов речевого глагола перечислить и противоположная ситуация во втором, ср.:

В своем выступлении докладчик перечислил наличествующие признаки социально-экономического кризиса и Жена перечислила мужу все свои покупки за день, можно трактовать и в пользу единой диатезы, ведь недостающие участники в обоих случаях, в принципе, могут быть добавлены ("докладчик перечислил - нам - признаки ...", "в своем письме с курорта - жена перечислила мужу ..."), и в пользу разных диатез, поскольку в первом примере, где акцентируется прежде всего нетривиальное содержание речевого акта, поверхностное выражение адресата все же воспринимается напряженно, а во втором – прототипическом контексте устного бытового разговора - участник "источник", как правило, "уходит в тень".

Дрейф тематического класса можно наблюдать, например, при рефлексивном употреблении некоторых речевых актов: глагол сформулировать в контексте Ученый сформулировал новую гипотезу сначала для себя, а потом уже изложил ее в докладе - занимает промежуточное положение между собственно речевым глаголом (что характерно для его исходных употреблений, нацеленных на нормального адресата - на другое лицо или на аудиторию) и глаголом ментальной деятельности (осознал идею и придумал формулировку) [9].

Для тематических классов один, чисто внешний, затрудняющий момент связан с тем, что разные тематические разделы в "Лексикографе" описываются разными исполнителями, и при изменении тематики глагола (вследствие метафоризации или десемантизации) производные употребления, вышедшие за пределы определенной локальной базы данных, вообще могут остаться непротолкованными.

Конечно, можно было бы отбрасывать примеры, допускающие неоднозначные интерпретации либо отклонения от указанного в толковании прототипа, но тогда будет ослаблена исследовательская "жилка" системы, которая как раз и составляет одну из ее сильных сторон. По-видимому, более рационально будет сохранять "второсортные" примеры: сомнительные с точки зрения языковой нормы, размытые, периферийные для данного тематического класса и т.п., но записывать такие примеры следует в специальные поля словарной статьи, отдельно от примеров "вполне хороших". Накапливаемая таким образом "свалка" примеров может послужить материалом для дальнейших исследований. То есть вновь, при работе со словарной информацией другого типа, чем в РУСЛАНе, мы приходим к выводу о целесообразности ранжирования иллюстраций.

В АОТ-ориентированном словаре при выборе и ранжировании контекстов может быть учтена и их потенциальную значимость для конкретных задач текстовой обработки. Так, при описании параметрической лексики с целью

использования словарных данных при фактографическом поиске, ранг примеров общеродового употребления параметрических имен (типа трудовая теория с т о и м о с т и, при нагревании о б ъ е м тела увеличивается, тонометр - это прибор для измерения кровяного д а в л е н и я и т.п.) может быть снижен из прагматических соображений: подобные употребления с неполной диатезой не выражают конкретных значений параметров. Такие употребления, очевидно, следует приводить в словаре, так как они раскрывают нюансы поведения слова, но лучше их выносить в периферийную часть зоны иллюстраций. (Аналогично, и у других лексических классов в употреблениях с неполной диатезой информационная значимость имеет тенденцию к уменьшению, ср. разную информативность прямой и параметрической диатез глагола в [7]: Выбрали председателем Петра и Выбрали председателя. Правда, в "Лексикографе" для подобных примеров вопрос о ранжировании отдельно не стоит, поскольку, как мы уже отметили, там они окажутся в разных статьях).

В заключение несколько слов еще об одной полезной функции примеров, которую мы упомянули в самом начале. Примеры дают материал для пополнения словаря, и потому иллюстрирование может рассматриваться как один из вспомогательных механизмов формирования словника. В словаре РУСЛАН, например, благодаря тому, что его нынедействующая программная оболочка (которую разработал А.В.Сокирко) поддерживает высветку в поле иллюстраций незнакомых словарю слов, ввод примеров показывает, какие интересные и практически значимые слова следует в ближайшее время описать в словаре.

Литература

1. Апресян Ю.Д. Лексическая семантика. Синонимические средства языка. М.: Наука, 1974. - 367с.
2. Виноградов В.В. О некоторых вопросах теории русской лексикографии // Лексикология и лексикография. Избранные труды. - М.: "Наука", 1977. - С.243-264.
3. Кустова Г.И., Падучева Е.В. Словарь как лексическая база данных. - Вопросы языкознания, 1994, N 3. - С.96-105;
4. Леонтьева Н.Н. К теории автоматического понимания естественных текстов. Часть 2. Семантические словари: состав, структура, методика создания. - М., МГУ, 2001. - 41 с.
5. Леонтьева Н.Н., Семенова С.Ю. Об отражении полисемии в прикладном семантическом словаре // Компьютерная лингвистика и интеллектуальные технологии. Труды международного семинара Диалог '2002'. Протвино, 6-11 июня 2000 года. - М., "Наука", 2002. - Т.2. - С. 489-496.
6. Леонтьева Н.Н., Семенова С.Ю. Инструменты построения файла ПЕРСОНА // НТИ.- Сер.2. - 2001. - N 8. - С. 9-20.
7. Падучева Е.В. О параметрах лексического значения глагола: таксономический класс участника // Русский язык в научном освещении. - N 1(3). - 2002. - С.87-111.
8. Семенова С.Ю. Прилагательные в семантическом словаре одной прикладной системы // Труды Международного семинара Диалог'98 по компьютерной лингвистике и ее приложениям. - М., 1998. - С.553-564.
9. Семенова С.Ю. Порождение Текста как выход из Хаоса (в печати).