Г.И. Кустова, С.Ю. Толдова

НАЦИОНАЛЬНЫЙ КОРПУС РУССКОГО ЯЗЫКА: СЕМАНТИЧЕСКИЕ ФИЛЬТРЫ ДЛЯ РАЗРЕШЕНИЯ МНОГОЗНАЧНОСТИ ГЛАГОЛОВ[1]

1. Введение

В нашем предыдущем сообщении[2] речь шла о создании семантических фильтров по снятию многозначности прилагательных в Национальном корпусе русского языка (НКРЯ). Все тексты основного корпуса (**http://www.ruscorpora.ru**), помимо общей метатекстовой разметки (автор, жанр и т.п.), имеют также грамматическую и семантическую разметку, которая значительно расширяет возможности пользователя при создании поисковых запросов и улучшает качество результатов поиска. Лингвистическая разметка может использоваться и для нужд самого Корпуса, а именно – для снятия лексической неоднозначности (что, в свою очередь, отвечает интересам пользователей).

---

[1] Работа выполнена при поддержке РГНФ, проект № 08-04-00181а. Примеры взяты из Национального корпуса русского языка.

[2] *Кустова Г.И.*, *Ляшевская О.Н.*, *Рахилина Е.В*. Семантическая разметка и семантические фильтры для Национального корпуса русского языка // Труды международной конференции «Корпусная лингвистика – 2006», СПб., 2006. С. 209–218; см. также *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; *Шеманаева О.Ю.*, *Кустова Г.И.*, *Ляшевская О.Н.*, *Рахилина Е.В*. Семантические фильтры для разрешения многозначности в Национальном корпусе русского языка: прилагательные // Иомдин Л.Л., Лауфер Н.И., Нариньяни А.С., Селегей В.П. (ред.). Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2007». М., 2007. С. 582–587.

Значения многозначных слов в Корпусе различаются не номерами, как в обычных толковых словарях, а семантическими пометами: значения, относящиеся к разным семантическим классам, имеют разные пометы, например: *пилить* (*бревно*) – «физическое воздействие», *пилить* (*мужа*) – «речь». Если в словаре каждая помета сопоставлена соответствующему значению, то в текстах Корпуса каждому вхождению слова приписываются все пометы, которые были у него в словаре, т.к. при автоматической расстановке помет значения слова различить невозможно. Программа разрешения многозначности использует семантические фильтры, основанные на принципе контекстной однозначности. В предложении многозначное слово употреблено в одном определенном значении (не считая случаев языковой игры). Это значение согласовано с контекстом. Например, глагол *разбушеваться* имеет в словаре Корпуса два значения: «природное явление» и «поведение человека»; соответственно, каждое его вхождение в текстах Корпуса имеет эти две пометы. Первое значение реализуется в контексте существительных класса ‘природное явление’ (*Вьюга разбушевалась*), второе – в контексте существительных класса ‘лицо’ (*Сосед разбушевался*). Семантический фильтр включает признаки контекста, соответствующие данному значению. Обнаруживая соответствующий контекст, программа снимает ненужную помету и оставляет нужную. Неоднозначность, таким образом, снимается с точностью до семантического класса, т.е. с точностью до семантической пометы (разумеется, не все значения глаголов имеют отдельные пометы. Мы берем глаголы, достаточно хорошо обеспеченные пометами. Именно для таких глаголов пишутся семантические фильтры).

Теоретически есть два ключевых параметра глагола, важных для составления семантических фильтров: модель управления (МУ) и семантические классы актантов (при широком понимании МУ семантические характеристики актантов включаются в нее

наряду с грамматическими; мы придерживаемся узкого понимания МУ как «падежной рамки» глагола).

МУ можно извлекать как из текстов (из корпусов), так и из специальных и обычных словарей. Задача выделения моделей управления, актуальная во многих системах автоматической обработки языка как для синтаксического анализа, так и для разрешения семантической неоднозначности, может решаться либо чисто статистическими способами[1], что приводит к потере точности, либо «вручную», т.е. силами экспертов. Во втором случае речь идет о создании специальных лексикографических ресурсов, таких как WordNet, FrameNet[2]. Активно разрабатывается такой ресурс – RussNet – и для русского языка в группе под руководством И.В. Азаровой[3]. В своей работе мы опирались на опыт группы разработчиков RussNet, однако наш эксперимент был призван оценить, каким образом можно использовать готовые

---

[1] *Brown P.F.*; *Della Pietra St.A.*; *Della Pietra*, *V.J.*; *Mercer R.* Word-sense disambiguation using statistical methods // ACL. 1991. V. 29. P. 264–270.

[2] *Dagan I.*, *Itai A.*, *Schwall U.* Two languages are more informative than one // Proceedings of the ACL, 1991 (29). P. 130–137; *Fellbaum Chr.* (ed.) WordNet: An Electronic Lexical Database. MIT Press, 1998; *Gale W.A.*; *Church*, *Kenneth W.* and *Yarowski*, *David*. A method for disambiguating word senses in a large corpus. // Computers and the Humanities. 1992. Vol. 26. P. 415–439; *Gildea D.*; *Jurafsky D.* Automatic Labeling of Semantic Roles // Computational Linguistics. 2002. Vol. 28. No 3. P. 245–288.

[3] *Азарова И.В.*, *Синопальникова А.А.*, *Яворская М.В.* Принципы построения wordnet-тезауруса RussNet // Кобозева И.М., Нариньяни А.С., Селегей В.П. (ред.), Компьютерная лингвистика и интеллектуальные технологии: труды международной конференции «Диалог 2004». М., 2004. С. 542–547; *Митрофанова О.А.*, *Кадина В.В.*, *Савицкий В.С.* Экспериментальное исследование синтагматических свойств лексем на основе лексикографических описаний и корпусов текстов // Труды международной конференции MegaLing'2006: Горизонты прикладной лингвистики и лингвистических технологий. Украина, Крым, Партенит. 20–27 сентября 2006 г.

лексикографические источники и каким образом дополнять извлеченную из этих источников информацию с использованием обучающего корпуса.

В качестве основного источника МУ глаголов использовался словарь глагольного управления[1]. Из словаря извлекалась информация о различных возможных наборах актантов и сирконстант для разных значений глагола, о грамматических ограничениях на них (часть речи, падеж, иногда – число). Для простоты все глагольные зависимые, в том числе наречия и предложно-падежные адвербиалы, мы будем далее называть актантами.

Информация по второму параметру – семантическим ограничениям на актанты – была взята из Корпуса: использовалась таксономическая разметка существительных в НКРЯ. Первоначально учитывалась только минимальная семантическая и лексико-грамматическая информация об актантах: одушевленность / неодушевленность и абстрактность / конкретность. Если минимального набора признаков оказывалось все-таки недостаточно, привлекалась более детальная информация о таксономическом классе соответствующих существительных. Имеющаяся в Корпусе семантическая разметка была дополнена новыми пометами, а именно: (а) была расширена система таксономических классов; (б) учитывались метафорические переносы (помета «metaph»); (в) для служебных значений (лексических функций[2], ср., например, *найти* в *найти возможность*) была введена помета «LF».

Для уменьшения ошибок, связанных с отсутствием синтаксического анализа, мы использовали преобразования исходного контекста, моделирующие неполный синтаксический анализ. Материалом эксперимента послужил корпус со снятой морфологи-

---

[1] *Апресян Ю.Д.*, *Палл Э.* Русский глагол – венгерский глагол. Управление и сочетаемость. Будапешт, 1982.

[2] *Апресян Ю.Д.* Лексическая семантика: Синонимические средства языка. М., 1974.

ческой омонимией объемом 4,5 млн. словоупотреблений. Исследовались глаголы из высокочастотной части списка. Эксперимент должен был ответить на вопрос: в какой степени данные о МУ глагола с использованием минимальной информации о семантическом классе актантов (одушевленность vs. неодушевленность, абстрактность vs. конкретность) позволяют понизить степень многозначности.

В простейшем случае достаточно значения какого-то одного параметра – (1) модели управления глагола или (2) семантического класса актанта / актантов.

(1) Для идентификации значения может быть достаточно модели управления, если она является уникальной для данного значения. Например, у глагола *следовать* в словаре Корпуса (на уровне помет) различаются значения: ‘движение’ (*следовать из Москвы в Казань*; *следовать за проводником*), ‘существование’ (*событие следовало за событием*), ‘локативное’ (*далее следовала подпись и печать*; *за отелями следовали рестораны и бары*), ‘поведение’ (*Он во всем следует примеру отца*), модальное (*Этого следовало ожидать*), лексическая функция (*Из этого положения следует вывод*). У некоторых значений модели управления могут совпадать (ср. *X следует из Y-а*, *X следует за Y-ом*), но есть значение, связанное с уникальной моделью управления (*X следует Y-у*) и тем самым однозначно определимое по синтаксическому контексту.

(2) Иногда для различения двух значений решающую роль играет семантическая характеристика актанта. Так, среди значений глагола *бродить* в Корпусе различаются физическое движение (move): *Дачники долго бродили по его огромному саду* – и метафорическое движение (metaph_move): *Грустная улыбка бродила по его лицу*. Поскольку их МУ совпадают, фильтр, снимающий одну из помет, использует сведения о семантическом классе первого актанта (подлежащего):

сущ.: Им.: конкр.: лицо, животное → *бродить*: move;
сущ.: Им.: абстр. → *бродить*: metaph_move.

Однако в реальных текстах ситуация намного сложнее. Во-первых, часто приходится задействовать оба признака. Во-вторых, и для МУ, и для семантических ограничений на актанты есть факторы, препятствующие или, наоборот, способствующие снятию неоднозначности.

2. Роль информации о грамматических и семантических ограничениях на актанты при создании семантических фильтров для разрешения глагольной многозначности

2.1. Модель управления

Реализация в предложении того или иного варианта МУ может как препятствовать (I), так и способствовать (II) автоматическому различению значений многозначного слова.

I. Факторы, препятствующие различению значений.

(1) Первая сложность связана с недостаточной различительной «мощностью» моделей управления.

(1а) Реализована базовая МУ.

Базовая, «стандартная» МУ, характерная для данного глагола или класса глаголов, во-первых, обычно обладает наибольшей степенью многозначности, а во-вторых, имеет, как правило, наибольшее покрытие. Так, базовая МУ глагола *отдать* / *отдавать* (и других глаголов этого класса) <именительный, винительный, дательный> представлена в целом ряде значений: исходное значение – ‘каузация обладания’ (*Он всегда отдает долги друзьям*), метафорическое от ‘каузации обладания’ (*Он отдает все силы борьбе*); лексические функции (*Командир отдает приказы бойцам*; *Бойцы отдают честь командиру*), ‘движение’ (*Нападающий отдал мяч защитнику*). В таких случаях нельзя обойтись только указанием МУ, необходимо включать в фильтр и семантическую информацию об актантах.

(1б) Модель управления реализована не полностью.

Два значения глагола *кричать* – «звук» (*Раненый кричал от боли*) и «речь» (*Командир кричал, чтобы бойцы отходили к лесу*) – различаются на уровне полных МУ. Однако при неполной реализации МУ совпадают (ср.: *Перевязка закончилась, а раненый все кричал* vs. *Командир все кричал, а бойцы не двигались*).

(2) Еще одна сложность состоит в том, что количество именных групп в предложении часто не совпадает с количеством именных групп, указанных в словарном источнике. В предложении могут содержаться именные группы, которые входят в состав других именных групп и не являются непосредственно актантами глагола: *Он нашел [для меня] [квартиру]* vs. *Он нашел [нож [для чистки картофеля]]*. Мешают однозначно выделять актанты в реальном предложении и такие специальные конструкции, как комитативные и дистрибутивные группы, ср., например: *Он дал Пете по голове* vs. *Он дал каждому по прянику*. Наконец, в Корпусе достаточно высок процент неполных предложений, где глагол употреблен без актантов (около 10%), ср. *Нашел*; *…потому что думал* и т.п.

II. С другой стороны, есть факторы, способствующие понижению неоднозначности (сокращению числа помет).

(1) Модель управления, включающая «специфичные» актанты, существенно сужает число возможных значений вплоть до одного. Например, для глагола *болеть* предложная группа *за*+S&acc в МУ задает только одно значение: *Он болеет за «Динамо»*; глагол *отдавать* в контексте Твор.п. реализует значение ‘запах’ (*Чай отдает рыбой*; посессивное значение тоже допускает Твор.п., но предполагает еще и Вин.п., ср.: *Отдает долги борзыми щенками*). Реализация валентности инструмента у «физического» значения глагола *пилить* (*пилить бревно пилой* (Твор.)) позволяет однозначно отличить его от речевого значения (*пилить мужа*). У речевого значения, в свою очередь, есть валентность мотивировки (*пилить за что*), которой тоже доста-

точно для его идентификации. Разное падежное оформление второго актанта при глаголах движения также позволяет существенным образом сузить класс значений. Так, глагол *идти* имеет по разметке НКРЯ 8 тэгов. Для значения ‘движение’ возможно более 20 МУ. Однако каждая из этих МУ либо связана только с данным значением, либо максимальная величина кластера не превышает 3-х значений.

Таким образом, МУ может быть надежным критерием для идентификации значения: если в предложении помимо собственно синтаксических валентностей (соответствующих подлежащему и прямому дополнению) реализуются специфичные валентности, обусловленные особенностями семантики конкретного глагола, а также факультативные валентности или некоторые сирконстанты, учет этих распространителей нередко позволяет отличить одно значение от другого, не прибегая к семантическим признакам существительных.

(2) Отсутствие в реальном предложении каких-либо именных групп не обязательно ведет к повышению неоднозначности (к реализации всех или большинства возможных значений); для некоторых глаголов такой контекст, наоборот, снижает число возможных семантических тэгов. Например, для глагола *получить* МУ с отсутствием прямого дополнения в винительном падеже может сигнализировать о том, что реализовано значение ‘физическое воздействие’: *Ты у меня получишь!*; *Получишь по шее!*; *Получил в рожу*; отсутствие актанта в дательном падеже (входящего в базовую МУ для исходных значений глаголов передачи) характерно для некоторых лексических функций (*дать течь; дать эффект*). Для многих глаголов надежным показателем типа значения является неопределенно-личная конструкция: часто (хотя и не всегда) она возможна только для первого значения (*Сзади толкают*; *Улицу не освещают*).

2.2. Семантические ограничения на актанты

Вторым важнейшим диагностическим признаком (наряду с МУ) является семантический класс актанта. Однако данная характеристика, как и МУ, может выступать в роли диагностического признака далеко не всегда.

(1) Есть сложности, связанные с использованием минимального исходного набора различительных признаков (абстрактность / конкретность, одушевленность / неодушевленность). Во-первых, существуют классы неодушевленных существительных, для которых характерны стандартные метонимические переносы, меняющие семантическую характеристику, например: организация → множество работающих в ней людей, ср. *Партия создана в 2001 г.* vs. *Партия решила*… Во-вторых, иногда важно не противопоставление актантов по абстрактности / конкретности, а их объединение по некоторому семантическому компоненту, ср. *Горит свет* (абстр. сущ.) и *Горит лампа* (конкр. сущ., осветительный прибор).

(2) Нередки случаи, когда исходного набора признаков недостаточно. Анализ данных показывает, что чем специфичней ограничения, тем точнее может быть разрешена многозначность. Иногда приходится прибегать к более частным семантическим признакам в рамках широких классов абстрактности / конкретности. Например, для глагола *оторвать* – (1) *оторвать листок от календаря* (‘воздействие: ликвидация контакта’) vs. (2) *оторвать голову от подушки* (‘движение’) vs. (3) *оторвать детей от матери* (‘метаф.: ликвидация контакта’) vs. (4) *оторвать студентов от учебы* (‘фаза’) – три значения из четырех не только имеют одинаковые модели управления, но и одинаковую характеристику актантов – ‘конкр.’. Для различения этих значений актантам должны быть приписаны дополнительные признаки: «сущ. Вин. = часть тела» в (2) и «сущ. Вин. = лицо» в (3) (при этом характеристика «часть тела» может использоваться для идентификации значения (2) только совместно с грамматической

характеристикой другого актанта «*от* + сущ.: Род.», т.к. актант «часть тела» есть и в другом значении, ср.: *взрывом оторвало ногу*). В классе абстрактных существительных для различения значений иногда также приходится указывать более частные подклассы, ср., например: *Свет горит* vs. *План горит*.

В некоторых случаях приходится даже использовать лексические фильтры, т.е. правила, в которых фигурируют конкретные лексемы. Например, для глагола *болеть* словосочетание *болеть душой* однозначно указывает на метафорическое значение (класс эмоций), глагол *сбить* в сочетании *сбить с ног* реализует значение ущерба. Т.е. почти со 100% точностью можно во всех подобных примерах оставить ровно одно значение.

3. Некоторые результаты эксперимента

Эксперимент показал, что несмотря на перечисленные выше сложности (неполная реализация МУ в тексте, совпадение МУ у разных значений и под.), грамматическая и минимальная семантическая информация об актантах способна существенно снизить степень многозначности (т.е. уменьшить количество семантических помет) глаголов в текстах Корпуса.

Как синтаксические характеристики актантов, так и семантические ограничения на них могут иметь разную различительную силу. Эксперимент подтвердил ряд исходных гипотез, но в то же время дал и некоторые неожиданные результаты.

(а) В сфере морфолого-синтаксических характеристик, как и ожидалось, более информативными оказываются более периферийные актанты. При этом можно разбить глаголы на классы в зависимости от того, в какой степени именно грамматическая информация позволяет уменьшать число возможных значений.

К неожиданным результатам относится, например, тот факт, что для многих глаголов ситуация, когда в предложении не хватает каких-то актантов, оказывается также более «благоприятной» для разрешения многозначности, чем полная стандартная

модель, т.е. отсутствие одного или нескольких актантов иногда может служить не менее надежным критерием для идентификации значения в тексте, чем наличие специфичных актантов. Неполные реализации МУ и специальные конструкции с отсутствующими (с другой точки зрения – нулевыми) актантами (неопределенно-личная, безличная) в каких-то случаях не препятствуют, а способствуют разрешению неоднозначности. Этот практический результат эксперимента может послужить базой для важного теоретического и лексикографического вывода: значения глаголов и других предикатных слов должны описываться не только с точки зрения того, какая модель управления их характеризует (и различает), но и с точки зрения того, какие специальные синтаксические конструкции и какие неполные реализации МУ они допускают.

(б) Что касается семантических характеристик актантов, то они тоже не обладают каким-то постоянным «коэффициентом» различительности для всех глаголов. Один и тот же семантический признак актанта для одних глаголов может быть решающим, а для других – ни о чем не говорить. Так, для глаголов движения прямое значение физического перемещения характерно как для одушевленного, так и для неодушевленного субъекта, при этом и тот, и другой класс может участвовать в метафорических переносах и сочетаться с лексическими функциями (ср. *Дети прыгают ~ Мяч прыгает ~ Сердце прыгает ~ Что ты прыгаешь с одной работы на другую?*; *Человек идет ~ Поезд идет ~ Товар идет хорошо ~ Почему ты идешь на это?*). Для глаголов же восприятия или ментальных глаголов наличие неодушевленного подлежащего в исходном значении очень маловероятно, так что контекст неодушевленного подлежащего, как правило, указывает на полуслужебное значение (лексическую функцию: ср. *Окна смотрят на юг*; *Метод нашел применение*…; *Этот дом знал лучшие времена*).

В сфере лексико-грамматических и семантических характеристик эксперимент также дал некоторые неожиданные результаты. Априори можно было предположить, что столь общие характеристики актантов, как «одушевленность» / «неодушевленность» и «конкретность» / «абстрактность», не являются эффективным различающим инструментом и в идеале для различения значений нужно приписывать актанту его «точный» (терминальный) семантический класс. Однако результаты эксперимента показали, что даже этих общих признаков во многих случаях оказывается достаточно для существенного понижения степени многозначности.

В целом эксперимент показал, что семантические ограничения в сочетании с синтаксической ролью образуют иерархию с точки зрения надежности отсечения лишних значений. Абстрактность актанта чаще играет решающую роль в определении значения глагола, чем одушевленность. Так, для глагола *дать* абстрактность существительного в позиции прямого дополнения является решающим ограничением для выделения употреблений данного глагола как лексической функции. При этом абстрактность актанта, занимающего позицию подлежащего, более сильный различительный признак, чем, например, абстрактность локативного актанта.

В заключение приведем диаграмму, в которой отражена «различительная сила» грамматических и обобщенных семантических признаков актантов для некоторых глаголов: см. рис. 1.

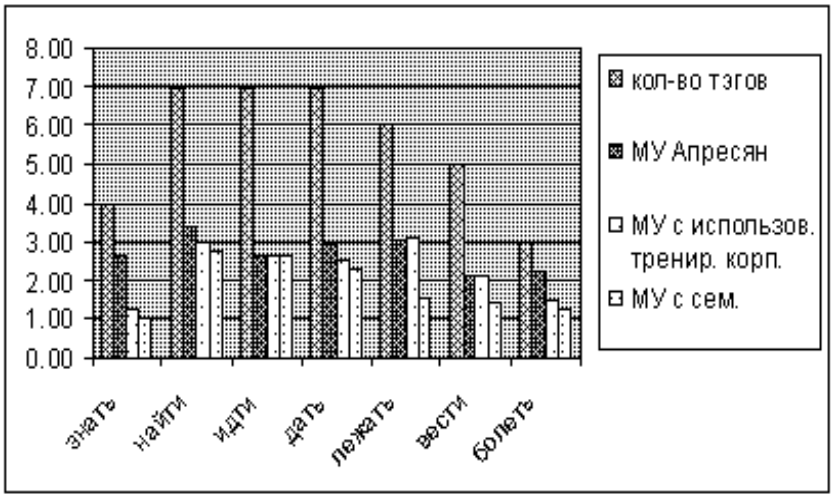

Рис. 1.

Для глаголов *найти*, *идти*, *дать, лежать* информация о грамматических свойствах актантов («МУ Апресян») позволяет снизить число возможных значений более чем в два раза. При этом использование корпусных данных («МУ с использованием тренировочного корпуса») в ряде случаев существенно улучшает результаты применения грамматических фильтров (ср., например, данные для глаголов *знать, болеть*). Семантические ограничения («МУ с семантическими характеристиками актантов»), как видно из диаграммы, также имеют разное значение для разных классов глаголов. Так, включение в число ограничений обобщенных семантических характеристик актантов глагола *идти* совсем никак не влияет на снижение степени многозначности. Для глаголов же *лежать, вести* такие характеристики позволяют снизить многозначность почти до одного тэга на глагол, т.е. полностью разрешают многозначность в большинстве контекстов.