Г. И. Кустова,
С. Ю. Толдова

НКРЯ: семантические фильтры для разрешения многозначности глаголов[1]

1. Введение

Нам уже приходилось писать в ряде публикаций (см. Кустова и др. 2005, 2006; Шеманаева и др. 2007; Кустова, Толдова 2008a,b) о том, как устроена семантическая разметка в Национальном корпусе русского языка (НКРЯ) и как работает программа снятия неоднозначности[2]. Однако, чтобы не затруднять читателя поиском этих публикаций, мы сочли целесообразным вкратце повторить некоторые основные тезисы, касающиеся проблемы многозначности в НКРЯ и методов ее автоматического разрешения.

Слова в текстах основного корпуса (*http://www.ruscorpora.ru*) имеют грамматическую и семантическую разметку, которая значительно расширяет возможности пользователя при создании поиско-

---

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

2 См. также статью Е. В. Рахилиной и др. в настоящем сборнике (сс. 215–239), которая затрагивает проблемы снятия многозначности на материале адъективной лексики.

вых 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).

Благодаря наличию семантической разметки 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 (impact)», *пилить* (*мужа*) — «речь (speech)».

Если в словаре пометы распределены по значениям, то в текстах Корпуса каждому вхождению слова приписываются все пометы, которые были у него в словаре, т.к. пометы расставляются автоматически, и программа «не знает», в каком значении употреблено слово в каждом отдельном случае. Для снятия «лишних» помет нужна другая программа—программа разрешения многозначности, которая 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, который, в свою очередь, тоже имеет семантическую помету. Если удается сформулировать простое семантическое правило вида «в контексте существительного семантического класса X у глагола реализуется значение семантического класса Y», оно и становится основой для семантического фильтра.

Например, глагол *красоваться* имеет в словаре Корпуса два значения: «поведение человека (behav)» (*Мальчик красовался перед нерусскими ребятишками* (В. Месяц)) и «местонахождение (loc)» (*В кабинете над камином красовался герб князей Черкасских* (газ.); *Среди горелых построек красовался барак* (В. Астафьев)); соответственно, каждое его вхождение в текстах Корпуса имеет эти две пометы. Первое значение (behav) реализуется в контексте существительных класса ‘лицо’, и семантический фильтр для него включает соответствующий признак. Получая на вход такой контекст, программа оставляет у глагола нужную помету и автоматически удаляет ненужную:

красоваться (behav; loc) + сущ.: лицо → *красоваться* (behav)

В остальных контекстах программа оставляет помету «loc».

Разумеется, разработчики заинтересованы в том, чтобы составлять фильтры не для отдельных глаголов, а для целых классов глаголов. Но для этого нужно сначала найти такие классы глаголов, у которых в определенном контексте одинаковым образом меняется значение. Регулярные семантические сдвиги чаще развиваются, как известно, на базе метонимических отношений. Например, многие глаголы звучания (*звонить, трезвонить, тарахтеть, шипеть* и др.) в контексте личных существительных приобретают значение «речь»; многие глаголы деформации (*резать, ломать, колоть*) имеют значение ущерба (*порезать палец*) и значение физиологического (обычно болезненного) ощущения (*режет в животе; колет в боку; меня всего ломает*). Обнаружение таких классов не только позволяет оптимизировать работу программы автоматического снятия многозначности, но и помогает формулировать семантические закономерности в области сдвигов значений.

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.

Неоднозначность может иметь разное происхождение:

а) омонимия, ср. *найти 1* и *найти 2:*
Я ***нашел*** *этот дом легко* vs. ***Нашла*** *коса на камень*;

б) полисемия, ср. *найти 1:*
Я ***нашел*** *этот дом легко* vs. ***Нашла*** *возможным помочь нам;*

в) «искусственная» неоднозначность (ср. *болеть*: *болеет* vs. *болит*): *люди меньше* ***болели*** vs. *уши привыкли к давлению и не так* ***болели***.

Для фильтров это безразлично.

В фильтрах могут использоваться не только семантические, но и грамматические признаки, прежде всего—модель управления глагола или ее элементы. Например, для глагола *болеть* предложная группа ***за*** **+ сущ. Вин**. задает только одно значение: *Он болеет за «Динамо»*,—поэтому для идентификации данного значения удобно использовать именно грамматический контекст.

Таким образом, теоретически есть два ключевых параметра глагола, важных для составления семантических фильтров:

1. модель управления (МУ);
2. 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 извлечения моделей управления из текстов решается в рамках создания специальных лексикографических ресурсов, таких как WordNet, FrameNet[3], а также – для русского языка—RusNet (разрабатывается группой исследователей под руководством И. В. Азаровой[4]), однако она требует значительного времени и усилий квалифицированных экспертов. Решение же такой задачи чисто статистическими способами[5] приводит к потере точности.

Мы в своей работе в качестве основного источника МУ глаголов использовали словарь глагольного управления: *Апресян Ю. Д., Палл Э.* Русский глагол—венгерский глагол. Управление и сочетаемость. Будапешт, 1982. Вот как выглядит, например, словарная статья глагола *бродить* в этом словаре:

Номер значения	**Модель управления**	**Пример**
1	N1/n_ V PR1 N2/x_	*Они бродили в лесу*.
1	N1/n_ V PR1 N2/x_	*Дачники бродили по дорожкам сада*.
2	N1/n_ V	*Вино бродит*.
3	N1/n_ V в N2/x_	*Странные мысли бродили в его голове*.
4	N1/n_ V по N2/d_	*Грустная улыбка бродила у девушки по лицу*.
5	N1/n_ V	*Ветер бродит*.

[3] См. [Dagan et al. 1991; Fellbaum (ed.) 1998; Gale et al. 1992. Gildea, Jurafsky 2002; Lopatková et al. 2005].

[4] См. [Азарова и др. 2004; О. А. Митрофанова и др. 2006].

[5] См. [Lesk 1986; Brown et al. 1991; Gale et al. 1992; Manning, Schütze 1999].

Из словаря можно извлечь информацию о различных возможных наборах актантов и сирконстантов 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. Это связано с одной из задач эксперимента по составлению глагольных фильтров—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 снизить степень многозначности. Если минимального набора признаков оказывалось все-таки недостаточно, привлекалась более детальная информация о таксономическом классе соответствующих существительных.

При составлении фильтров имеющаяся в Корпусе семантическая разметка была дополнена новыми пометами, а именно: (а) была расширена система таксономических классов; (б) учитывались метафорические переносы: к помете исходного значения, от которого образовалось метафорическое, прибавлялась помета «metaph», например: *дышать* «physiol» (*Трудно было дышать сырым воздухом*)—*дышать* «metaph physiol» (*Чем дышит сейчас столица?*); (в) для служебных значений (лексических функций[6], ср., например, *найти* в *найти возможность*) была введена помета «LF».

Так, значения упомянутого выше глагола *бродить* получили следующие семантические пометы:

---

[6] О понятии «лексической функции» см. Апресян 1974, Мельчук 1974.

Животные бродили с одного пастбища на другое [*из конца в конец деревни*].	move
Они бродили в лесу [*в незнакомом городе…*].	move
Дачники бродили по рощам [*по дорожкам сада*].	move
Солдаты долго бродили, искали свою часть.	move
Вино бродит.	changest
Грустная улыбка бродила у девушки по лицу.	metaph move
Странные мысли бродили в его голове.	metaph move
Ветер бродит.	metaph move

Для уменьшения ошибок, связанных с отсутствием синтаксического анализа, мы использовали преобразования исходного контекста, моделирующие неполный синтаксический анализ. Материалом послужил корпус со снятой морфологической омонимией объемом 4,5 млн. словоупотреблений. Исследовались глаголы из высокочастотной части списка.

Как показала практика составления фильтров, в простейшем случае для смыслоразличения достаточно задать какой-то один из обсуждавшихся выше параметров—(1) модель управления глагола или (2) семантический класс актанта / актантов.

1. Моделью управления можно ограничиться в тех случаях, когда 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 (так, каждому из контекстов *X следует из Y-а*, *X следует за Y-ом* могут соответствовать разные интерпретации), но есть значение, связанное с уникальной моделью управления (*X следует Y-у—следует примеру отца*),—оно однозначно определяется по синтаксическому контексту.

Еще пример. У глагола *достать* в Корпусе различается три значения: ‘движение’ (*достать чашку с полки*), ‘обладание’ (*достать*

дефицитное лекарство, *достать билет на Таганку*) и ‘контакт’ (*достать рукой до потолка*). Если у первых двух значений модель управления при неполной реализации может совпадать (ср. *достать чашку* и *достать дефицитное лекарство*), то последнее значение отличимо от первых двух по модели управления даже при неполной ее реализации (сущ.: Им. + *достать* + ***до*** сущ.: Род.).

2. Иногда для противопоставления двух значений решающую роль играет, напротив, 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):

(а) *бродить* (move, metaph move) + сущ.: Им.: конкр.: лицо, животное → *бродить* (move);
(б) *бродить* (move, metaph move) + сущ.: Им.: абстр. → *бродить* (metaph move).

Глагол *разбушеваться* имеет в словаре Корпуса два значения: «природное явление» и «поведение человека». Первое значение реализуется в контексте существительных класса ‘природное явление’ (*Вьюга разбушевалась*), второе— в контексте существительных класса ‘лицо’ (*Сосед разбушевался*).

Многие глаголы физического воздействия имеют производное значение, относящееся к классу ‘речь’ (*пилить бревно* vs. *пилить мужа*, *резать хлеб* vs. *резать правду*, *молоть муку* vs. *молоть чушь*). Любое вхождение такого глагола в текстах Корпуса имеет две пометы—«физическое воздействие» (impact) и «речь» (speech). Фильтр содержит контекст (существительное с нужными грамматическими и семантическими характеристиками), в котором реализуется одно из двух значений:

(а) *пилить* (impact, speech) + сущ.: Вин.: конкр.: физич. предмет (*пилить бревно*) → *пилить* (impact);
(б) *пилить* (impact, speech) + сущ.: Вин.: конкр.: лицо (*пилить мужа*) → *пилить* (speech);

(а) *молоть* (impact, speech) + сущ.: Вин.: конкр.: вещество (*молоть муку*) → *молоть* (impact);
(б) *молоть* (impact, speech) + сущ.: Вин.: абстр.: речь (*молоть чушь*) → *молоть* (speech).

В отличие от словаря, куда попадают специально подобранные, а иногда и специально составленные предложения, в Корпусе мы имеем дело с реальными предложениями, «вырванными» (извлеченными) из их реального контекста. Иногда в таких предложениях отсутствует необходимая для анализа информация, а иногда присутствует ненужный «шум». Чтобы учесть все эти случаи, материал Корпуса подвергался предварительной обработке.

Для каждого исследованного глагола составлялся тестовый корпус предложений с данным глаголом (в них встречались и полные МУ, соответствующие словарному источнику [Апресян, Палл 1982], и не полностью реализованные МУ, и вхождения глагола без распространителей). Приводимая ниже Диаграмма 1 дает представление о количественном соотношении разных моделей управления на примере глагола *давать*.

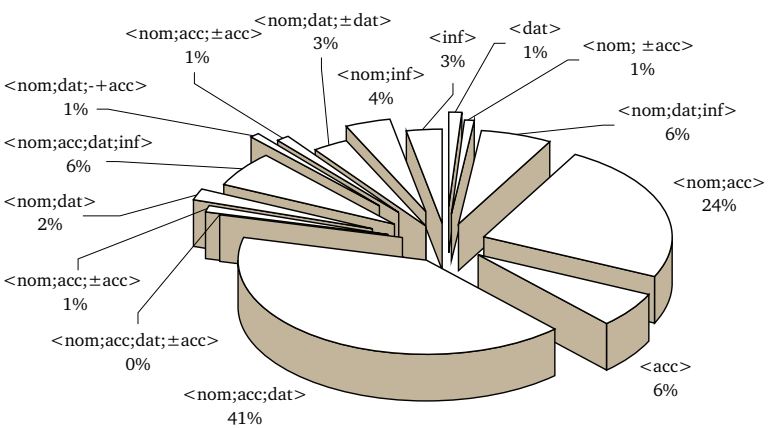

Диаграмма 1. Распределение моделей управления глагола *давать* в Корпусе

Как видно из диаграммы, МУ, включающие базовые актанты (<nom, acc, dat> и <nom, acc>), составляют бóльшую часть примеров Корпуса.

Анализ тестового корпуса позволил выявить случаи, препятствующие разрешению омонимии, и случаи, способствующие ее разрешению. К их рассмотрению мы и переходим.

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

2.1. Модель управления (грамматические ограничения)

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 (ср. выше диаграмму для глагола *давать*). 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*).

Базовая модель <именительный, винительный> глагола *покинуть* также представлена в разных значениях: прямое значение—класс ‘движение’ (*Новобранцы покинули родное село*), лексическая функция (*Смелость покинула его*—‘исчезновение’), фазовое значение (*Певица покинула сцену*).

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.: …*потому что думал*; *надо думать*; *и думать не хочу*; *продолжал мучительно думать*; *а по-настоящему думать* 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:

- значение глагола *найти* в контексте **прилагательного / причастия в Твор**. относится к классу ментальных или перцептивных (*Книгу я нашёл весьма грамотной*; *Иван нашел сестру плачущей*);
- глагол *дать* при наличии предложных групп ***в* + сущ. Вин**. или ***по* + сущ. Дат**. реализует значение ‘физическое воздействие’ (*Здорово ему давеча Кирилл Анатольевич дал по башке*);
- для глагола *толкать* актант ***на* + сущ. Вин**. в МУ задает только одно значение (*толкать на преступление*);
- глагол *отдавать* в контексте **сущ. Твор**. реализует значение ‘запах’ (*Чай отдает рыбой*; посессивное значение тоже допускает Твор., но предполагает еще и Вин., 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 ***за*** **+ сущ. Вин**. (*пилить за что*), которой тоже достаточно 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; для некоторых глаголов такой контекст, наоборот, снижает число возможных семантических тэгов—иногда даже вдвое.

Например, для глагола *получить* МУ без прямого дополнения в винительном падеже может сигнализировать о том, что реализовано значение ‘физическое воздействие’: *Ты у меня получишь!*; *Получишь по шее!*; *Получил в рожу*; аналогично у глагола *дать* (*А он ему как дал!*); отсутствие у *дать* актанта в дательном падеже характерно для некоторых лексических функций (*дать течь; дать свисток; 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., о с в е т и т е л ь н ы й 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 мы сталкиваемся даже с необходимостью 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. Подобные лексические фильтры почти со 100%-ной точностью предсказывают значение анализируемого глагола.

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 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 более «благоприятной» для разрешения многозначности, чем наличие полной стандартной модели, 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 на основе 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 в коей мере не снижать многозначности. 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 с неодушевленным субъектом, 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 инструментом снятия омонимии и в идеале для различения значений нужно приписывать актанту его «точный» (терминальный) семантический класс. Однако в ходе эксперимента обнаружилось, что даже этих общих признаков во многих случаях оказывается достаточно для существенного снижения степени многозначности глаголов в Корпусе.

В целом работа над фильтрами показала, 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 однозначно указывает на то, что данный глагол употреблен здесь как лексическая функция. Более того, абстрактность как смыслоразличительный признак имеет разную эффективность для существительных с разной синтаксической ролью: абстрактность актанта, занимающего позицию подлежащего, более значима, чем, например, абстрактность локативного актанта.

В заключение приведем Диаграмму 2, в которой отражена различительная сила грамматических и обобщенных семантических признаков актантов для некоторых глаголов:

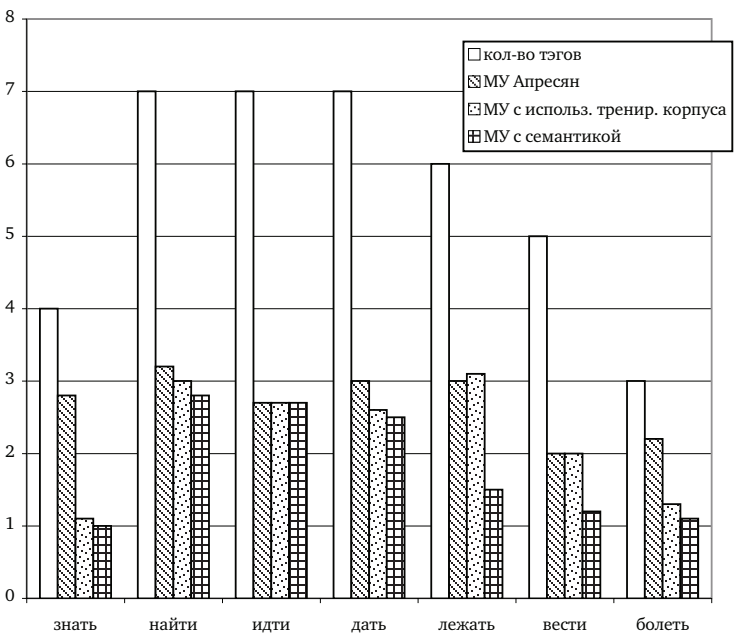

Диаграмма 2. Результаты эксперимента

Для глаголов *найти*, *идти*, *дать, лежать* информация о грамматических свойствах актантов (на диаграмме—«МУ Апресян») позволяет снизить число возможных значений более чем в два раза. При этом использование корпусных данных (на диаграмме—«МУ с использованием тренировочного корпуса») в ряде случаев сущест-

венно улучшает результаты применения грамматических фильтров (ср., например, данные для глаголов *знать, болеть*). Семантические ограничения (на диаграмме —«МУ с семантическими характеристиками актантов») также имеют разное значение для разных классов глаголов. Так, включение в число ограничений обобщенных семантических характеристик актантов глагола *идти* никак не влияет на уровень его многозначности. Для глаголов же *лежать, вести* такие характеристики позволяют снизить многозначность почти до одного тэга на глагол, т.е. полностью снимают полисемию в большинстве контекстов их употребления.

Литература

Азарова и др. 2004—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

Апресян 1974—Апресян Ю. Д. Лексическая семантика. М., 1974.

Апресян, Палл 1982—Апресян Ю. Д., Палл Э. Русский глагол—венгерский глагол. Управление и сочетаемость. Будапешт, 1982.

Кустова и др. 2005—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. С. 155–174.

Кустова и др. 2006—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.

Кустова, Толдова 2008*a*—Кустова Г. И., Толдова С. Ю. Национальный корпус русского языка: семантические фильтры для разрешения многозначности глаголов // Труды международной конференции «Корпусная лингвистика—2008». СПб., 2008. С. 240–252.

Кустова, Толдова 2008*b*—Кустова Г. И., Толдова С. Ю. Семантические фильтры для разрешения многозначности в Национальном корпусе русского языка: глаголы // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог-2008». М, 2008. С. 522–529.

Мельчук 1974—Мельчук И. А. Опыт теории лингвистических моделей «Смысл ⇔ Текст». М., 1974.

Митрофанова и др. 2006—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. 20–27 сентября 2006 г., Украина, Крым, Партенит.

Шеманаева и др. 2007—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.

Brown et al. 1991—Brown P.F., Della Pietra S.A., Della Pietra V.J., Mercer R. Word-sense disambiguation using statistical methods // ACL. 1991. V.29. P. 264–270.

Dagan et al. 1991—Dagan I., Itai A., Schwall U. Two languages are more informative than one // Proceedings of the ACL, 1991 (29). P. 130–137.

Fellbaum (ed.) 1998—Fellbaum Ch. (ed.) WordNet: An Electronic Lexical Database. MIT Press. 1998.

Gale et al. 1992—Gale W.A., Church K.W., Yarowski D. A method for disambiguating word senses in a large corpus. // Computers and the Humanities. 1992. Vol. 26. P. 415–439.

Gildea, Jurafsky 2002—Gildea D., Jurafsky D. Automatic Labeling of Semantic Roles // Computational Linguistics. 2002. Vol. 28. No 3. P. 245–288.

Lesk 1986—Lesk M. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from a ice cream cone

// Proceedings of SIGDOC '86. New York. Association for Computing Machinery. 1986. P. 24–26.

Lopatková et al. 2005—Lopatková M., Bojar O., Semecký J., Benešová V., Zabokrtský Z. Valency Lexicon of Czech Verbs VALLEX: Recent Experiments with Frame Disambiguation // V. Matoušek, P Mautner, and T. Pavelka (eds.) Text, Speech and Dialogue: 8th International Conference, TSD 2005.—Karlovy Vary, Czech Republic, September 12–15, 2005. Proceedings, volume LNAI 3658. Springer Verlag. 2005. P. 99–106.

Manning, Schütze 1999—Manning C.D., Schütze H. Foundations of Statistical Natural Language Processing. Chapter 7. Cambridge, Massachusetts: The MIT Press. 1999. P.230–262.