081-Toldova:_Layout 1 15.05.2008 22:15 Page 522

СЕМАНТИЧЕСКИЕ ФИЛЬТРЫ ДЛЯ РАЗРЕШЕНИЯ МНОГОЗНАЧНОСТИ В НАЦИОНАЛЬНОМ КОРПУСЕ РУССКОГО ЯЗЫКА: ГЛАГОЛЫ*

SEMANTIC FILTERS FOR THE WORD SENSE DISAMBIGUATION IN RNC: VERBS

Толдова С.Ю. *(toldova@yandex.ru)*
Московский государственный университет им. М.В. Ломоносова
Кустова Г.И. *(galina03@mtu-net.ru)*
Московский государственный педагогический университет
Ляшевская О.Н. *(olesar@mail.ru)*
Всероссийский институт научной и технической информации РАН

В статье обсуждаются результаты эксперимента по разработке системы семантических фильтров глаголов, используемых для разрешения неоднозначности лексико-семантической разметки в Национальном корпусе русского языка. Основные задачи эксперимента: проверить, в какой степени можно использовать специализированные лексикографические источники для создания таких фильтров (в качестве основного источника использовался словарь глагольного управления [Апресян-Палл 1982]); какие ограничения на актанты (семантические, лексические, грамматические) наиболее значимы для фильтров.

1. Введение

Настоящая статья продолжает серию публикаций в «Диалоге», освещающих работу над созданием лексико-семантической разметки Национального корпуса русского языка (http://www.ruscorpora.ru), см. [18], [19], [21], [22]. Все тексты Основного корпуса содержат три вида лингвистической разметки: метатекстовую (автор, жанр текста и т.д.), грамматическую (лемма и грамматические признаки) и лексико-семантическую (разметка по лексико-семантическим группам и словообразовательным типам). Сейчас на первый план выходит задача повышения точности разметки и снижения уровня «шума» в результатах поиска. Ее решение связано с учетом разных значений многозначных и омонимичных слов и с правильным распознаванием этих значений в тексте. В статье обсуждаются результаты экспериментов по разрешению неоднозначности семантической разметки глаголов.

В словаре Корпуса каждое значение слова снабжено семантическим ярлыком, показывающим его принадлежность к тому или иному таксономическому (семантическому) классу, например:

валяться

1) ‘движение: движение субъекта’ (*поросята валяются в грязи*);

2) ‘местонахождение’ (*бумаги валяются на полу*).

Таким образом, в словаре у многозначного слова обычно имеется несколько семантических помет, причем эти пометы распределены по разным значениям. Однако когда программа автоматически расставляет пометы в тексте, то она каждому вхождению слова приписывает все пометы, которые есть у слова в словаре (поскольку программа не может определить, в каком значении употреблено слово в каждом конкретном случае). В результате многозначное слово в тексте имеет все множество возможных помет. Это часто мешает более точному поиску в корпусе, создает «шум», а также иногда является источником ошибок морфологической разметки и лемматизации. В этой связи возникает задача найти способ снизить количество семантических помет для конкретного контекста без использования ручной разметки.

Для разрешения неоднозначности семантической разметки в Корпусе была разработана технология специальных фильтров. Семантический фильтр основан на принципе контекстной однозначности, т.е. на том, что в каждом конкретном контексте слово имеет одно значение (за исключением случаев языковой игры). Семантический фильтр – это правило, задающее некоторый минимальный контекст, в котором реализуется определенное значение слова. Если программа, содержащая фильтры, обнаруживает в предложении такой контекст, то она оставляет соответствующую ему семантическую помету, а остальные пометы удаляет. Таким образом,

* Работа выполнена при поддержке РГНФ, проект № 08-04-00181а.

Примеры взяты из Национального корпуса русского языка.

081-Toldova:_Layout 1 15.05.2008 22:15 Page 523

многозначность снимается с точностью до семантического класса (т.е. с точностью до семантической пометы). В фильтрах для разрешения неоднозначности прилагательных (см. [22]) используются только семантические характеристики определяемого существительного (поскольку грамматические характеристики существительного – род, падеж и число – не влияют на семантический класс (помету) прилагательного). В глагольных фильтрах тоже используются признаки связанных с глаголом существительных, но глагольные фильтры, в отличие от фильтров прилагательных, устроены более сложно. В них могут использоваться два параметра – семантические классы существительных, связанных с глаголом, и модель управления (МУ) глагола (т.е. морфолого-синтаксические характеристики существительных; учитываются также другие виды зависимых – придаточные, наречия). Мы исходили из того, что:

▪ глагол является синтаксическим и семантическим ядром предложения, а его базовые свойства определяются его МУ;

▪ значение глагола непосредственно связано с его МУ: с морфологическими (синтаксическими) и семантическими характеристиками его актантов (см. [6]).

В простейшем случае достаточно значения какого-то одного параметра – (1) модели управления глагола или (2) семантического класса существительного (далее МУ понимается в узком смысле – как «падежная рамка» глагола; возможно и широкое понимание МУ, когда в нее включаются не только грамматические, но и семантические характеристики актантов).

(1) Для идентификации значения может быть достаточно модели управления, если она является уникальной для данного значения. Например, у глагола *достать* в Корпусе (на уровне помет) различается три значения: ‘движение’ (*достать книгу с полки*), ‘обладание’ (*достать лекарство*, *достать билет в театр*) и ‘контакт’ (*достать рукой до потолка*). Если у первых двух значений модель управления может совпадать при неполной реализации (ср. *достать книгу* и *достать лекарство*), то последнее значение связано с уникальной моделью управления, и даже при ее неполной реализации (сущ.: Им. + *достать* + ***до*** сущ.: Род.) отличимо от первых двух. (2) Иногда для различения двух значений решающую роль играет семантическая характеристика актанта. Рассмотрим пример простейшего фильтра, различающего два значения глагола только за счет семантического класса управляемого существительного при совпадении модели управления. 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, в котором реализуется одно из двух значений (контекст включает существительное с нужными грамматическими и семантическими характеристиками). Получая на вход такой контекст, программа оставляет у глагола нужную помету и автоматически удаляет ненужную:

(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).

Конечно, в большинстве случаев ситуация намного сложнее.

Первая сложность связана с недостаточной различительной «мощностью» моделей управления. Часто у разных значений совпадают простейшие («минимальные») модели управления, включающие подлежащее и дополнение, ср.: *Он бросил снежок* vs. *Он бросил школу* vs. *Он бросил упрек.* Могут совпадать МУ с предложными группами: (а) *Он вернулся к столу* – (б) *Он вернулся к жене* – (в) *Сознание вернулось к нему* – (г) *Докладчик вернулся к первому вопросу*. Есть случаи, и их немало, когда совпадают не только минимальные, но и «расширенные» МУ: *X бросил сумку на диван / в шкаф / за ширму* vs. *X бросил взгляд на дверь / в окно / за ширму*; *Следователь вызвал свидетеля на допрос* vs. *Следователь вызвал свидетеля на откровенность*. В таких случаях нельзя обойтись только указанием МУ, необходимо включать в фильтр и семантическую информацию об актантах.

Другая сложность состоит в том, что количество именных групп в предложении, как правило, 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]]*. С другой стороны, в реальном корпусе достаточно высок процент неполных предложений (около 10%), состоящих, например, только из одного глагола (ср. *Нашел.*). 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.*

Если говорить о семантических характеристиках, то здесь также возникает немало проблем. Во-первых, существуют классы неодушевленных существительных, для которых характерны стандартные метонимические

081-Toldova:_Layout 1 15.05.2008 22:15 Page 524

переносы, меняющие семантическую характеристику, например: организация → множество работающих в ней людей, ср. *Партия создана в 2001 г.* vs. *Партия решила…* Во-вторых, сложность в том, что 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.). Кроме того, множества абстрактных и конкретных существительных неоднородны, поэтому иногда для различения значений необходимо выделять частные подклассы внутри класса абстрактных или конкретных существительных, ср., например: *Свет горит* vs. *План горит.*

2. Исходные данные

Каковы источники двух типов информации (МУ и семантические ограничения на актанты), используемой в фильтрах?

МУ можно извлекать как из текстов (из корпусов), так и из специальных и «обычных» словарей.

Задача выделения моделей управления (МУ) во многих системах автоматической обработки языка является актуальной как для синтаксического анализа, так и для разрешения семантической неоднозначности. Данная задача решается либо чисто статистическими способами (см. например, [1]), что приводит к потере точности, либо является трудоемким ручным процессом. Одним из способов преодоления указанных трудностей является создание специальных лексикографических ресурсов общего доступа таких, как WordNet, FrameNet и др. (см. [2], [3], [4], [5], [9]). Активно разрабатывается такой ресурс и для русского языка – RusNet – в группе под руководством И.В.Азаровой [15], [20]. Создание такого ресурса также достаточно трудоемкий процесс.

Что касается статистических методов, то для разрешения многозначности используются как контролируемые методы обучения, так и неконтролируемые ([1], [4], [8] и др., см. также обзор в [10]). Большинство таких систем базируются на баесовской модели или на модели канала с шумом. В работе [4] сообщается о достижении 90% точности для шести существительных с достаточно четко различимыми смыслами. Данный метод активно разрабатывался на материале английского языка. Он требует большого корпуса, размеченного вручную. Потенциальными признаками контекста являются все лексемы из достаточно большого окна. При неконтролируемом обучении (см., например [10]) невозможна семантическая разметка с приписыванием тому или иному слову семантического тэга, задача сводится к кластеризации множества контекстов на группы и их различение (discrimination).

Для задач нашего проекта было важно учесть опыт использования специализированных лексикографических ресурсов. Такие методы предполагают либо первичную полуавтоматическую разметку тренировочного корпуса (ср. проект Senseval [11]), либо использование тезаурусов и словарных систем, таких как Wordnet, FrameNet, VerbNet. Технологии применения данных систем активно разрабатываются в проектах по семантическому аннотированию корпусов на многих языках, в том числе при разрешении многозначности глаголов с использованием моделей управления: [6], [7], [12], [13], [14].

Для русского языка в рамках проекта RusNet проводились пилотные эксперименты по применению лексикографических источников для извлечения моделей управления (см. [20]). Однако данный проект предполагает задачу лексикографического описания глагола, а не снятие омонимии в корпусе.

В своей работе мы опирались на опыт группы разработчиков RusNet, однако наш эксперимент был призван оценить, каким образом можно использовать готовые 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 Апресян-Палл 1982 [17]. Из словаря извлекалась информация о различных возможных наборах актантов и сирконстант для разных значений глагола, о грамматических ограничениях на них.

Что касается второго параметра, то в качестве источника информации о семантических ограничениях использовалась таксономическая разметка существительных НКРЯ. Первоначально учитывалась только минимальная семантическая и лексико-грамматическая информация об актантах: одушевленность / неодушевленность и абстрактность / конкретность. Несмотря на перечисленные в разделе 1 сложности, минимальная грамматическая и семантическая информация способна существенно снизить степень многозначности. Если минимального набора признаков оказывалось все-таки недостаточно, привлекалась более детальная информация о таксономическом классе соответствующих существительных. Имеющаяся в Корпусе семантическая разметка для целей эксперимента была дополнена новыми пометами, а именно: (а) была расширена система таксономических классов; (б) учитывались метафорические переносы (помета «metaph»); (в) для служебных значений (лексических функций в смысле [16], 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.

081-Toldova:_Layout 1 15.05.2008 22:15 Page 525

Материалом эксперимента послужил корпус со снятой морфологической омонимией объемом 4,5 млн. словоупотреблений. Исследовались глаголы из высокочастотной части списка.

Эксперимент должен был ответить на следующие вопросы:

- ✓ в какой степени можно использовать информацию автоматически или полуавтоматически извлеченную из лексикографических источников;
- ✓ 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;
- ✓ каким образом извлекать информацию об актантах глагола в конкретном контексте, не используя полный синтаксический анализ;
- ✓ какова технология пополнения исходного списка МУ с использованием обучающего корпуса;
- ✓ каков должен быть формат глагольного фильтра для разрешения семантической неоднозначности;
- ✓ как взаимодействуют различные таксономические, грамматические и лексические ограничения.

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

3.1. Вклад информации о составе и грамматических характеристиках элементов МУ в разрешение глагольной многозначности

Сначала был проведен эксперимент, имеющий целью установить, каков вклад собственно грамматической информации об актантах в разрешение многозначности глагола.

Для каждого исследованного глагола составлялся тестовый корпус предложений с данным глаголом (в них встречались и полные МУ, соответствующие словарному источнику [17], и не полностью реализованные МУ, и вхождения глагола без распространителей). В качестве примера ниже приводится диаграмма 1, которая показывает распределение моделей управления глагола *давать* в Корпусе:

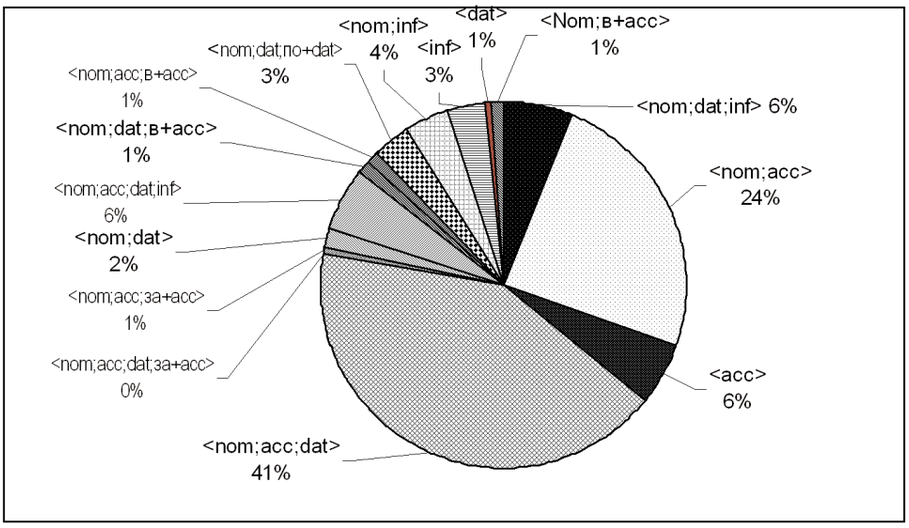

Диаграмма 1.

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

Анализ тестового корпуса позволил выявить (1) случаи, препятствующие разрешению омонимии, и (2) случаи, способствующие ее разрешению.

(1) Факторы, повышающие неоднозначность.

(а) Реализована базовая МУ.

Показательным является тот факт, что базовая, «стандартная» МУ, характерная для данного глагола или класса глаголов, обычно обладает наибольшей степенью многозначности. Так, базовая МУ глагола *дать* / *давать* (и других глаголов этого класса) <именительный, винительный, дательный> представлена почти во всех воз-

081-Toldova:_Layout 1 15.05.2008 22:15 Page 526

можных для данного глагола значениях: прямое значение – класс ‘каузация обладания’ (*Мать дала ему пирожок*), лексические функции (*Мы дадим вам такую возможность*; *Войска дали отпор врагу*), класс ‘физическое воздействие’ (*Она дала ему пощечину* – это лексическая функция, однако она может быть «семантизирована»). При этом такая модель, как правило, имеет наибольшее покрытие (37% для глагола *дать*). 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*).

(б) Модель управления реализована не полностью.

В корпусе встречается достаточно много случаев, когда некоторая МУ реализована не полностью или в доступном анализу отрезке текста представлен только глагол без актантов. Вот несколько примеров для глагола *думать*: *в министерстве действительно могут так думать*; *чтобы она не мешала мне думать*; *надо думать*; *я думаю*; *потому что думал*; *и думать не хочу*; *продолжал мучительно думать*; *а по-настоящему думать* и т.п.

(2) Факторы, понижающие неоднозначность.

(а) Модель управления, включающая «специфичные» актанты, существенно сужает число возможных значений вплоть до одного. Так, например, для глагола *болеть* предложная группа *за*+S&acc в МУ задает только одно значение: *Он болеет за «Динамо»*. Значение глагола *найти* в контексте прилагательного в творительном падеже относится к классу ‘ментальные действия’ или ‘восприятие’ (*Книгу я нашёл весьма грамотной*). Глагол *дать* при наличии предложных групп *в*+Вин. или *по*+Дат. реализует значение ‘физическое воздействие’ (*Здорово ему давеча Кирилл Анатольевич дал по башке*). Для глагола *толкать* актант *на*+S&acc в МУ задает только одно значение (*толкать на преступление*). Реализация валентности инструмента у «физического» значения глагола *пилить* (*пилить бревно* (Вин.) *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 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 сужения 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.

(б) 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 вдвое. Например, для глагола *дать* МУ с отсутствием прямого дополнения в винительном падеже может сигнализировать о том, что реализовано значение ‘физическое воздействие’ (*А он ему как дал*); отсутствие актанта в дательном падеже характерно для некоторых лексических функций (*дать течь; дать свисток; дать эффект*).

(в) Неполная реализация МУ, ее редукция, вплоть до отсутствия синтаксически выраженных актантов в определенных конструкциях, также не всегда является негативным фактором, в некоторых случаях она может, наоборот, служить для разрешения неоднозначности, сужая число возможных значений. Так, употребление глагола *толкать* без падежных распространителей в неопределенно-личной конструкции (*Сзади толкают*) возможно только для первого (физического) значения.

3.2. Семантические ограничения

Следующим диагностическим признаком является семантический класс актанта. Однако данная характеристика играет роль диагностического признака далеко не всегда. 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 одушевленных, так и для неодушевленных объектов, при этом и тот, и другой класс может участвовать в метафорических переносах и сочетаться с лексическими функциями (ср. *Поезд идет из Москвы ~ Человек идет из сада ~ Чай идет из Индии* и т.п.). Для глаголов же восприятия или ментальных глаголов наличие неодушевленного подлежащего очень маловероятно. А это значит, что, например, глагол *найти* не может реализовывать одно из своих переносных значений, относящихся к классу ментальных, в контексте, когда позицию подлежащего занимает неодушевленный актант (ср. *Метод нашел применение*…). Семантические ограничения в сочетании с синтаксической ролью образуют иерархию с точки зрения надежности отсечения лишних значений. Абстрактность актанта чаще играет решающую роль в

081-Toldova:_Layout 1 15.05.2008 22:15 Page 527

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 надежный признак, чем абстрактность локативного актанта.

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. Так, если в случаях *бросать гневные упреки* / *горькую правду* / *чудовищные обвинения* не ограничиваться пометой «абстр.» для сущ. в Вин., а использовать помету «речь», то данное значение глагола *бросать* тоже можно будет идентифицировать как «речь» (а не как бессодержательную «лексическую функцию»). В случаях *оторвать голову от подушки*, *не отрывать глаз от книги* общую характеристику существительного в Вин. «конкр.» имеет смысл дополнить более частной характеристикой «часть тела» (впрочем, эта помета может использоваться для идентификации данного значения только совместно с грамматической характеристикой другого актанта «*от* + сущ: Род.», т.к. семантическая характеристика «часть тела» может быть у актанта и в другом значении, ср.: *взрывом оторвало ногу*).

В некоторых случаях приходится даже использовать лексические фильтры, т.е. правила, в которых фигурируют конкретные лексемы. Например, для глагола *болеть* важно, что отдельно необходимо рассматривать словосочетание *болеть* с существительным *душа*: *болеть душой* – наличие в предложении слова *душа* однозначно указывает на непрямое, метафорическое значение глагола. В контексте данного существительного значение глагола следует отнести к классу ‘выражение эмоций’. Т.е. почти со 100% точностью можно во всех подобных примерах оставить ровно одно значение.

3.3. Некоторые результаты эксперимента

Эксперимент показал, что грамматические характеристики актантов и сирконстант позволяют существенным образом понизить многозначность глаголов. Особенно информативны 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. Что касается лексико-грамматических и семантических характеристик, то самых общих признаков «одушевленность» / «неодушевленность» и «конкретность» / «абстрактность» иногда оказывается достаточно для существенного понижения степени многозначности.

Рассмотрим диаграмму 2, в которой отражены свойства грамматических и обобщенных семантических ограничений для некоторых глаголов.

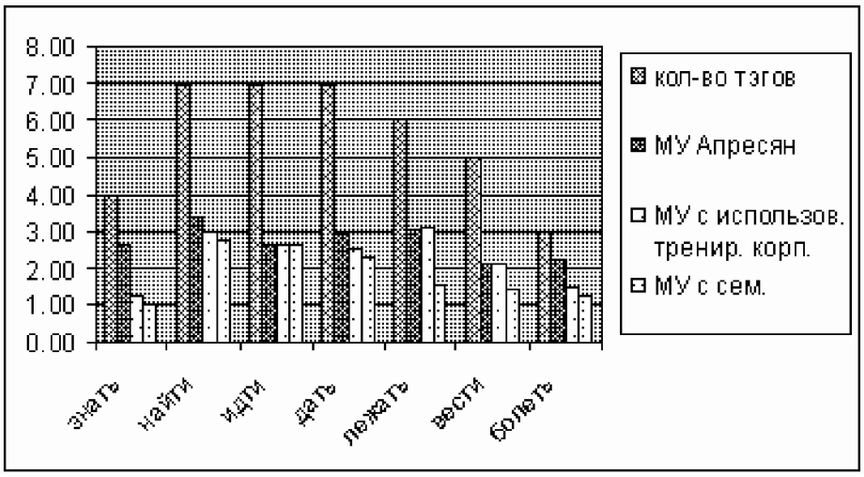

Диаграмма 2.

Для глаголов *найти*, *идти*, *дать, лежать* информация о грамматических свойствах актантов позволяет снизить число возможных значений более чем в два раза. При этом использование корпусных данных в ряде случаев существенно улучшает результаты применения грамматических фильтров (ср., например, данные для глаголов *знать, болеть*). Как видно из диаграммы, семантические ограничения также имеют разное значение для разных классов глаголов. Так, включение в число ограничений обобщенных семантических характеристик актан-

081-Toldova:_Layout 1 15.05.2008 22:15 Page 528

тов глагола *идти* совсем никак не влияет на снижение степени многозначности. Для глаголов же *лежать, вести, болеть* такие характеристики позволяют снизить многозначность почти до одного тэга на глагол, т.е. полностью разрешают многозначность в большинстве контекстов.

4. Выводы

Как синтаксические характеристики актантов, так и семантические ограничения на них могут иметь разную различительную силу.

С точки зрения различительной силы актанты образуют иерархию. Их можно разбить на два класса:

(а) базовые актанты, такие как S&nom, S&acc, а также актанты, соответствующие семантическому классу глагола в его прямом значении (например, актант, указывающий на место, для глаголов движения; датив для глаголов класса *давать*);

(б) уточняющие актанты.

МУ, содержащая базовые актанты, приводится первой в словарных источниках. Базовые актанты наиболее частотны при данном глаголе в корпусе (более 60%). Они, как правило, содержатся в нескольких МУ данного глагола и реализуются с несколькими значениями.

Второй класс актантов включает более специфические, необязательные актанты, например, предложные группы (*за*+S&acc, *по*+S&dat и др.) или инфинитив. Они обладают большей «различительной» силой. Наличие такого актанта в МУ существенным образом сужает множество соответствующих значений вплоть до одного, таким образом, он может служить диагностическим признаком для некоторого значения даже при отсутствии в контекстной МУ других актантов.

Описанное выше противопоставление «различительных» и «неразличительных» актантов вполне предсказуемо. Неожиданным результатом явился тот факт, что для многих глаголов ситуация, когда по некоторым причинам в предложении не хватает актантов, оказывается также более «благоприятной» для разрешения многозначности, чем полная стандартная модель. Так, например, в предложении *Он дал* глагол *дал* с достаточно высокой степенью вероятности не может иметь ни значения лексической функции, ни значения физического воздействия, ни значения ‘позволить’, а относится к исходному классу ‘каузировать иметь’. Множество значений глагола *найти* в предложении с опущенными актантами также уменьшается с пяти максимально возможных до двух (обладание (*найти кошелек*) и метафорический перенос по этому значению (*найти выход*), ср.: *Он долго искал кошелек* / *выход. И наконец нашел*). Подобную информацию можно извлечь только из размеченного обучающего корпуса, либо опираясь на интуицию эксперта, поскольку никакие лексикографические источники такой информации не дают по определению. Это не входит в их задачу.

Наибольшую сложность для снятия многозначности представляют случаи, когда для разных значений набор основных актантов совпадает. В такой ситуации признаки образуют некоторую иерархию с точки зрения их различительной силы. Наибольшей степенью различительности обладают лексические ограничения (случаи, когда глагол реализует данное значение только в устойчивом словосочетании), далее следуют периферийные (факультативные актанты), а также такие семантические характеристики, как абстрактность.

Список литературы

1. Brown, P.F., Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert Mercer. Word-sence 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.

3. Fellbaum, Christian (ed.) WordNet: An Electronic Lexical Database. MIT Press. 1998.

Gale, William 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.

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

5. Johnson,C., Fillmore,C., Petruck,M., Baker,C.,Ellsworth,M., Ruppenhofer, J., and Wood, E. FrameNet: Theory and Practice. 2002. [Electronic resource]. 2002. Mode of access: http://www.icsi.berkeley.edu/framenet.

6. Kingsbury, P., Palmer, M., and Marcus, M. Adding semantic annotation to the Penn TreeBank. // Proceedings of the Human Language Technology Conference HLT-2002. San Diego, California, 2002.

7. Lesk M. Automatic sense disambiguation using machinereadable 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.

081-Toldova:_Layout 1 15.05.2008 22:15 Page 529

8. Lopatková, Markéta, Ondřej Bojar, Jifí Semecký, Václava Benešová, and Zdenék _Zabokrtský. Valency Lexicon of Czech Verbs VALLEX: Recent Experiments with Frame Disambiguation. // Václav Matoušek, Pavel Mautner, and Tomáš Pavelka, editors. 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.

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

10. Mihalcea R., Chklovsky T., Kilgarriff A. Framework and results for English SENSEVAL // Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, July 2004, Barcelona. Barselona, Spain, 2004. P. 25–28. ftp://ftp.itri.bton.ac.uk/reports/ITRI-04-09.pdf.

11. Ng H.T., Lee H.B. Integrating multiple knowledge sources to disambiguate word sense: An examplar-based approach // Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL-96). Santa Cruz, 1996.

12. Scott Songlin Piao, Rayson P., Archer D., McEnery T. Comparing and combining a semantic tagger and a statistical tool for MWE extraction // Computer Speech & Language. Vol. 19. No 4. 2005. P. 378–397.

13. Shi, L., and Mihalcea, R. Semantic parsing using FrameNet and WordNet. // Proceedings of the Human Language Technology Conference (HLT/NAACL 2004). Boston, 2004.

14. 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.

15. Апресян Ю.Д. Лексическая семантика. – М.: «Наука», 1974. – 368 с.

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

17. Кобрицов Б.П., Ляшевская О.Н., Шеманаева О.Ю. Снятие лексико-семантической омонимии в новостных и газетно-журнальных текстах: поверхностные фильтры и статистическая оценка // Интернет-математика – 2005. М.: 2005. С. 38–57.

18. Кустова Г.И., Ляшевская О.Н., Падучева Е.В., Рахилина Е.В. Опыт семантического расширения морфологической разметки: таксономическая классификация лексики в национальном корпусе русского языка // НТИ, сер. 2. Информационные процессы и системы. № 6. 2005.

19. О.А. Митрофанова, В.В. Кадина, В.С. Савицкий. 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.

20. Рахилина Е.В., Ляшевская О.Н., Кобрицов Б.П., Кустова Г.И., Шеманаева О.Ю. Многозначность как прикладная проблема: Лексико-семантическая разметка в Национальном корпусе русского языка // Лауфер Н.И., Нариньяни А.С., Селегей В.П. (ред.). Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2006». 2006. С. 445–450.

21. 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». 2007. С. 582–587.