Г. И. Кустова, О. Н. Ляшевская, Е. В. Падучева, Е. В. Рахилина

СЕМАНТИЧЕСКАЯ РАЗМЕТКА ЛЕКСИКИ В НАЦИОНАЛЬНОМ КОРПУСЕ РУССКОГО ЯЗЫКА: ПРИНЦИПЫ, ПРОБЛЕМЫ, ПЕРСПЕКТИВЫ

1. Введение

Естественным расширением и продолжением разметки Корпуса по морфологическим признакам является таксономическая классификация словника Корпуса (семантическая разметка). В принципе, даже наличие одной только морфологической разметки значительно расширяет возможности лингвистов при использовании Корпуса, поскольку позволяет искать примеры употреблений не вообще слов, а слов в определенных формах (например, глаголов в императиве, существительных в том или другом падеже и т. д.). Уже морфологическая разметка содержит какую-то, пусть и довольно скудную, семантическую информацию (например, признак «одушевленность» у существительных). Внедрение же семантической разметки открывает целый спектр новых возможностей.

Семантическая разметка создается в Отделе лингвистических исследований ВИНИТИ и предназначена для того, чтобы усилить возможности поиска по лингвистическим параметрам — в частности, в тех случаях, когда поиск в текстах ведется не по изолированным словам, а по их сочетаниям, т. е. по конструкциям. В принципе, предполагается, что весь Корпус современного русского языка (т. е. все запланированные 100 млн. словоупотреблений) будет размечен таким образом, однако в настоящее время семантическая разметка применена только к той части корпуса, где снята морфологическая омонимия.

Параллельно группой Ю. Д. Апресяна ведется работа над значительно более сложной системой семантической разметки, ориентированной на экспериментальный синтаксически размеченный фрагмент корпуса (объемом около 0,5 млн. словоупотреблений) со снятой омонимией (подробнее см., например, [Чардин 2003]). По-

ка этого фрагмента корпуса нет в открытом доступе — полное описание принципов его семантической разметки дается в статье [Апресян и др. 2004].

2. Семантическая разметка и система «Лексикограф»

Семантическая разметка морфологически размеченного Корпуса была осуществлена на основе лексической базы данных «Лексикограф-эксперт», которая разрабатывалась в Отделе лингвистических исследований (ОЛИ) ВИНИТИ РАН с 1992 г. Наиболее «продвинутыми» частями «Лексикографа» является база глаголов, создаваемая группой исследователей под руководством Е. В. Падучевой[1], и база предметных имен Е. В. Рахилиной, см. [Красильщик, Рахилина 1992]. Система «Лексикограф» создавалась как высокотехнологичный инструмент лингвистического исследования; она была предназначена для того, чтобы получать списки слов по заранее заданным семантическим параметрам (прежде всего, таксономическим) и исследовать языковое поведение полученных классов лексики — их сочетаемость, типы моделей управления, диатетические сдвиги как проявление системных отношений в лексике. С этой целью разрабатывались системы семантической классификации глаголов, предметных имен, прилагательных и наречий. На основе таких классификаций и в процессе самой работы над системой «Лексикограф» были получены серьезные лингвистические результаты, которые отражены в нескольких монографиях, см. [Падучева 1996], [Рахилина 2000], [Падучева 2004], [Кустова 2004], [Ляшевская 2004].

Словник базы «Лексикограф» составлен на основе «Русско-французского словаря» [Зализняк 1972] и содержит около 10 тыс. слов — в том числе, База предметной лексики содержит 4 тыс. слов. Этот словник был признан достаточным для данной задачи. Для сравнения скажем, что объем словаря DIALING[2], который был размечен для Корпуса, составляет порядка 120 тыс. слов, а

[1] В настоящее время основными участниками проекта являются Г. И. Кустова и Р. И. Розина; в разные периоды времени в работе над этой базой также принимали участие О. Н. Ляшевская, С. Ю. Семенова, Е. В. Рахилина, М. В. Филипенко, Н. М. Якубова и Т. Е. Янко; подробнее см. [Кустова, Падучева 1994].

[2] Подробнее см. статью О. Н. Ляшевской, В. А. Плунгяна, Д. В. Сичинавы «О морфологическом стандарте Корпуса современного русского языка» в настоящем сборнике.

значит, превышает объем базы «Лексикограф» более чем в 10 раз. Однако, несмотря на небольшой объем, в теоретическом отношении экспериментальная исследовательская база данных «Лексикограф» сыграла огромную роль и фактически стала основой для создания лексико-семантического словаря Корпуса. Именно десятилетний опыт работы над системой «Лексикограф» позволил разработчикам решить эту трудоемкую и крайне нетривиальную задачу в рекордные сроки: семантический поиск на сайте начал работать уже в начале октября 2004 года.

Разумеется, семантическая классификация лексики, принятая в «Лексикографе», не могла быть применена к Корпусу в неизменном виде. Дело в том, что «Лексикограф» по своему замыслу ориентировался практически только на лингвистов-лексикологов, тогда как круг пользователей Национального корпуса русского языка значительно шире: с Национальным корпусом работают не только лингвисты, но и преподаватели и учащиеся (в процессе обучения школьников русскому языку и подготовки студентов филологических специальностей), иностранцы (для изучения русского языка и просто получения образцов качественных текстов на русском языке), программисты (в процессе тестирования программ и совершенствования поисковых систем), редакционно-издательские работники и другие группы пользователей. В силу этого принятая в «Лексикографе» семантическая классификация претерпела некоторые изменения. Стратегия этих изменений была следующей: с одной стороны, перенести из «Лексикографа» в Корпус то, что может представлять интерес и ценность для лингвистов, с другой — выдержать определенный уровень доступности для пользователя, не имеющего специальной лингвистической подготовки. Семантическая разметка, принятая в Корпусе, в идеале должна быть не только лингвистически содержательной, но и общепонятной — поэтому, наряду с лексической базой «Лексикограф» (электронным словарем), разработчики использовали обычные толковые словари [МАС], [БАС], [Даль 1907], [Ожегов 1972], [Ожегов, Шведова 1999], [Ушаков 1935-1940], а также лингвистические описания, отражающие традиционные принципы классификации лексики [Грамматика 1980], [Кузнецова 1989], [Бабенко 1999], [Шведова 2000], [Саяхова и др. 2000] (ср. еще [Levin 1993]).

Разметка внедряется в Корпус по этапам. В настоящее время разработаны наборы признаков для всех знаменательных частей

речи и размечены основные значения слов этих частей речи. Поиск сейчас возможен не по всем признакам из тех, по которым осуществлена разметка словаря, — для каждого класса слов выбран набор от 3 до 50 признаков. Отдельную проблему представляет лексическая многозначность в Корпусе (подробнее см. ниже). В ближайшее время у пользователей появится возможность выбора: искать слова только по первому значению — или учитывать весь спектр словарных значений.

4. Семантическая разметка: признаки и классы

Как уже было сказано, семантическая разметка является естественным расширением и продолжением уже прочно интегрированной в Корпус морфологической разметки. Она включает 3 группы признаков[3]:

(а) признаки, выражаемые словообразовательными показателями: словообразовательные корреляты соответствующих исходных слов получают пометы «диминутив», «аугментатив», «аттенуатив», «nomen agentis», «nomen femininum», «отыменное прилагательное», «отглагольное прилагательное», «отглагольное существительное», «семельфактив», «префиксальный глагол» и т. п.;

(б) признаки, соответствующие так называемым лексико-грамматическим разрядам: качественные, относительные и притяжательные прилагательные, предметные и непредметные существительные, имена собственные и т. п.;

(в) собственно семантические признаки: тематический (таксономический) класс; «оценка»; «каузативность» (у глаголов и отглагольных имен) и др.[4]

Как и в системе «Лексикограф», в Корпусе выделяется своя система признаков для каждой части речи: свой набор имеют глаголы, прилагательные, числительные, наречия, местоимения, и отдельно предметные и непредметные имена.

В качестве примеров выделяемых семантических классов можно привести следующие:

---

[3] Полную информацию о признаках, по которым доступен поиск в настоящее время, см. на сайте www.ruscorpora.ru, раздел «Семантика».

[4] Каждому значению слова соответствует свой набор признаков; кроме того, в семантическую разметку входит информация о статусе значения (первое/непервое значение).

- для глаголов: движение, физическое воздействие, создание, уничтожение, обладание, эмоция, речь, поведение человека;
- для прилагательных: размер, форма, цвет, вкус, запах, температура, место, время, свойство человека;
- у непредметных существительных, поскольку значительная часть их образована от глаголов и прилагательных, классы пересекаются с глагольными и адъективными, ср.: движение, физическое воздействие, создание, уничтожение, обладание, эмоция, речь и т. п. (для отглагольных) и цвет, вкус, температура, место, время, свойство человека (для отадъективных); кроме того, у них есть и «собственные» классы, такие, как мероприятие, болезнь, спорт, игра, единица измерения;
- для предметных имен: лица, животные, растения, вещества и материалы, здания и сооружения, инструменты, транспортные средства, и т. п.

Как уже было отмечено ранее, разработчики старались соблюдать принцип традиционности, т. е. избегать явных расхождений с традиционно принятой грамматической и таксономической номенклатурой. С другой стороны, были добавлены и некоторые из тех классификационных рубрик, разработанных в рамках системы «Лексикограф», которые отсутствуют в других классификациях. Например, для предметных имен это информация о мереологии (включающая, прежде всего, указания на отношения «часть-целое» и «элемент-множество», в которых участвует данный объект) и о топологии объекта (включающая такие классы, как «вместилища», «поверхности» и т. п.). Мереология и топология — это независимые параметры классификации, существующие параллельно таксономии, поэтому одно и то же существительное может характеризоваться по всем трем параметрам. Например, *ковш* (*экскаватора*) будет относиться к приспособлениям по таксономической классификации, к вместилищам по топологической характеристике и будет частью с точки зрения мереологии.

Заметим, однако, что и внутри таксономической классификации для Корпуса наиболее удобным был признан не древесный, а фасетный принцип классификации — и, следовательно, одно и то же слово может попадать сразу в несколько классов, если это необходимо. Действительно, в языке, как известно, есть множество случаев, когда одна лексема совмещает свойства нескольких классов. В такой ситуации разработчик, по нашему мнению, должен не

навязывать свое однозначное решение (как того требовала бы древесная классификация), а ориентироваться на весь спектр классификационных возможностей. Скажем, глагол *убедить* относится и к глаголам речи, и к глаголам воздействия на ментальное состояние; *вытребовать* — к глаголам речи и к поссесивным глаголам; *наполнить* — к глаголам помещения и изменения признака; *забить* <гвоздь> — к глаголам помещения и воздействия, и т. д.

Представляется, что фасетная классификация отражает интересы пользователя наиболее полно: именно поэтому фасетная система принята в такой области, как библиотечное дело, см. [Эйдельман 1977]: ведь, подобно Корпусу, библиотечный поиск ориентирован на самый широкий круг людей.

5. Лингвистическая релевантность семантических классов

В современной лингвистике стал практически общепринятым тезис, который на протяжении многих лет последовательно отстаивала Анна Вежбицкая, см. [Wierzbicka 1985; 1988]. Этот тезис состоит в том, что семантика языковых единиц отражается в их поверхностных свойствах — соответственно, если слова принадлежат к одному семантическому классу, то у них должны обнаруживаться определенные особенности языкового поведения, обусловленные их семантикой и общие для разных представителей данного класса. С точки зрения задачи семантической разметки Корпуса это означает, что классы разметки — в идеале − должны иметь лингвистическую релевантность.

Например, в Корпусе выделяется класс параметрических существительных типа *длина*, т. е. слов, которые имеют валентность на значение параметра (*длина — 2 м*), причем параметры могут быть не только числовыми, ср.: *место встречи — вокзал*; *исполнитель главной роли — Сидоров*. Ясно, что сам класс параметрических существительных можно выделить просто из соображений «здравого смысла». Но его состав будет определяться и уточняться исходя из того, какие характерные лингвистически релевантные проявления имеет параметрическая семантика. Вот несколько таких проявлений (подробнее см. [Падучева 2004]):

1. Параметрические существительные способны заполнять обязательную семантическую валентность интеррогативов (равно как и предикатов выбора, зависимости и под.): *определил величину*, *изменил условия приема, узнал его отношение* (но: **узнал его при-*

езд). При этом их параметрический актант (т. е. значение параметра) в этом контексте всегда отсутствует (**узнал длину 2 м*).

Ср. те же эффекты для нечисловых параметров: *Место встречи специально не* оговаривается; *Исполнитель главной роли не упоминается*.

2. Параметрический актант параметрического существительного плохо выражается синтаксически: как правило, ему не соответствует предложно-падежное управление, так что обычно он выражается с помощью предикативной, атрибутивной или аппозитивной связи: *длина забора — 2 метра*, *двухметровая высота*, *на глубине 2 м*[5]. Это ограничение сочетаемости имеет естественное объяснение: в контексте интеррогативов (а он является первичным для параметрических слов) интеррогативному актанту соответствует переменная, связанная «оператором вопроса» [Падучева 1985].

3. В *неинтеррогативном* контексте (параметрическая) валентность параметрического имени всегда заполнена, ср.: *Валиком хорошо красить плоские поверхности с большой площадью*, но **Валиком хорошо красить плоские поверхности с площадью*). Правда, поверхностно эта валентность может не выражаться — в этом случае способ заполнения восстанавливается слушающим из контекста, причем не всегда однозначно, ср. *Меня огорчили размеры его квартиры.*

В принципе, разные языковые явления могут потребовать разной глубины выделения семантических классов. Так, инструменты, устройства, транспортные средства, предметы посуды, одежды и мебели могут объединяться в более крупный класс приспособлений-артефактов. Этот класс противопоставляется именам людей, животных и природных объектов, в частности, по способу интерпретации сочетаний с относительными прилагательными, производными от названий пространственных объектов: *морской, речной, степной* и др. Сочетаясь с именами одушевленных и природных объектов, прилагательные такого рода описывают местонахождение объекта; а в сочетании с названиями артефактов прилагательные описывают место использования данного изделия: ср. *морская рыба, волна / речной песок, ил, риф / степная трава*, но *мор-*

[5] Из этого правила имеются редкие исключения. Так, Ю. Д. Апресян отметил, что в творительном падеже параметрическое существительное может иметь предложное управление, ср.: *длиной в 2 м* [Апресян 1974: 146].

ской бинокль, речной катер, кухонный нож, *полевая форма*, см. [Рахилина 2000].

В связи с этим в разметке используется понятие вложенных классов. Слово, относящееся к классу инструментов, будет найдено также по запросу на класс приспособлений, имена частей тела (*рука, коготь* и т. п.) — по запросу на класс частей, а имена с отрицательной оценкой — по запросу на класс оценочных слов.

В то же время необходимо подчеркнуть, что семантическая разметка, представленная в Корпусе, не нацелена и не может в полной мере служить инструментом контент-анализа текстов. При контент-анализе в поисковый образ некоторой категории (например, «переговоры») включаются любые единицы текста, способные кодировать данное понятие, в том числе коллокации: *сесть за стол, достичь договоренности, прошли успешно, противоположная сторона* и т. п., а в семантической разметке Корпуса в один класс входят только лексемы, принадлежащие одной части речи[6]. Кроме того, в практике контент-анализа часто используются очень узкие понятия, такие как «фамилии бизнесменов» или «внешний долг России». Очевидно, что подобные классы не будут релевантны в лингвистическом отношении.

Еще одним следствием принципа лингвистической релевантности является то, что слова, попадающие в один класс, имеют сходную структуру толкования. Например, из всех глаголов, в толкование которых входит компонент ‘контакт’, к классу глаголов контакта относятся только те, у которых этот компонент находится в вершине толкования, ср. *касаться, трогать, ухватиться* (глаголы контакта) и *передать, есть, копать, идти* (глаголы других классов).

Итак, важное направление работы над классификацией состоит в проверке и обосновании ее лингвистической релевантности. Эта задача решается не только в процессе разработки классификации, но и процессе работы над ней уже в онлайновом режиме.

[6] Сказанное не исключает симметричности классификаций для разных частей речи и для разных полей признаков: например, классы расстояния, скорости, количества присутствуют и у прилагательных, и у наречий, а структура таксономии предметных имен сохраняется в поле мереологических коррелятов (части тела человека, животного, части инструментов, одежды и т. д.).

6. Размеченный корпус как инструмент лингвистических исследований

6.1. Поиск по морфологическим и семантическим признакам: новые возможности лингвистического исследования

Семантическая разметка ориентирована на общую идею Корпуса, который был задуман не только и не просто как собрание текстов, но и как эффективный инструмент для решения разного рода научных и прикладных задач. Наличие значительного по объему и общедоступного морфологически и семантически размеченного Корпуса русского языка создает качественно новые возможности для лингвистических исследований.

Прежде всего, Корпус является источником примеров употреблений слов в текстах, т. е. языкового материала, необходимого для различных лингвистических исследований[7]. Если раньше на поиски примеров у лингвистов уходили месяцы, а иногда и годы, то теперь поиск примеров осуществляется автоматически в течение нескольких секунд, а количество получаемых примеров измеряется сотнями и тысячами. Правда, у этого расширения возможностей, как это всегда бывает, есть и оборотная сторона: эти тысячи примеров необходимо прочитать, осмыслить и расклассифицировать. Однако эта работа поддается оптимизации с помощью семантической разметки, т. к. семантическая разметка позволяет не только искать, но и исключать из поиска какие-то примеры.

Например, если предметом изучения является дательный субъекта в безличном предложении или дательный адресата, то нужны только существительные со значением лица. При поиске по морфологическому признаку «дательный падеж» будет получено много «лишних» примеров. При наличии же семантической разметки, даже таких простейших признаков, как ‘лицо’, ‘предмет’, ‘вещество’ и под., можно задать комбинированный запрос **сущ.: дат. & лицо**, благодаря чему отсеется большое количество «шума».

Вообще, очевидно, что наличие двух рядов признаков — морфологических и семантических — позволяет более точно и лингвистически содержательно формулировать любые запросы.

---

[7] Еще раз отметим, что корпус ориентирован на широкий круг пользователей, и примеры употреблений слов могут понадобиться не только лингвистам и не только для лингвистических исследований.

Для формулирования запросов и получения соответствующих примеров употреблений из текстов используется программа GRAMFIND, разработанная А. Е. Поляковым. Она позволяет осуществлять поиск не только по отдельным словам, но и по целым конструкциям, состоящим из 2-х или 3-х элементов с заданными морфологическими и семантическими признаками, например:

(а) частица *хоть* + глагол в форме императива (*Хоть плачь*);

(б) частица *как* + глагол СВ в буд. вр. (*Как крикнет*);

(в) предлог *у* + сущ. (с признаком 'лицо') или личн. мест. в Род. п. + сущ. в Им. п. (*У нас гости*; *У Пети грипп*); и т. п.

Существенным для разработчиков и ценным для пользователей является то, что наличие семантической разметки не только оптимизирует поиск примеров, но и позволяет решать разнообразные лингвистические задачи: изучение семантической сочетаемости, описание конструкций русского языка, уточнение свойств классов слов и отдельных слов, проверка гипотез и т. д.

Так, например, сделав запрос (в), пользователь, прежде всего, получает сам языковой материал, который ему нужен. Но, кроме того, по полученной выборке примеров можно выяснить, какие семантические классы глаголов, существительных и т. д. встречаются (а какие — не встречаются) в заданных конструкциях (например, для конструкции «*как* + глагол СВ в буд. вр.» не характерны ментальные глаголы, поскольку это конструкция со значением интенсивного действия).

С другой стороны, даже если у глагола семантический класс соответствует некоторой конструкции, это еще не значит, что он реально в ней употребляется. С помощью корпуса можно узнать, какие глаголы на самом деле бывают в этой конструкции в современном русском языке. Совершенно очевидно, что вручную собирать примеры по такому огромному количеству текстов чрезвычайно трудно. Использование же размеченного Корпуса делает подобную задачу вполне обозримой.

Вот примеры запросов (использующих комбинацию морфологических и семантических признаков), которые позволяют проверять гипотезы и получать лингвистически релевантную информацию:

(г) Запрос: **сущ. + *не* + глагол: движение**

позволяет выяснить, каков падеж субъекта в данном типе отрицательных предложений для существительных разных семантических классов. Ср. реально встретившиеся примеры из Корпуса:

*А **мальчик** во двор **не выходил***! (А. Алексин);

*И **бабушка не пошла** под венец* (А. Алексин);

Выражений благодарности не последовало (С. Довлатов);

Ответ не последовал (А. Азольский);

***Ответа не последовало**, и братья, ловившие каждое слово и движение через замочную скважину, возмущенно бабахнули кулаками по двери* (А. Азольский);

Танки** в прорыв еще **не вошли (В. Гроссман);

*Зачислили нас матросами на сейнеры, но оказалось, что наши **сейнеры** еще **не пришли** в колхоз* (В. Аксенов).

Данные Корпуса показывают, что имена лиц допускают только именительный падеж, ср.: *Мальчик не выходил — *Мальчика не выходило*; обозначения речевых действий допускают как именительный, так и родительный падеж (выбор падежа, по-видимому, зависит от определенности), ср.: *Ответ не последовал* (ожидавшийся) и *Ответа не последовало* (никакого); наконец, существительные класса «транспортное средство» допускают только именительный падеж, ср.: *Сейнеры еще не пришли — *Сейнеров еще не* пришло[8].

Как видим, поиск по данному запросу подтверждает гипотезу о том, что класс обозначений транспортных средств с точки зрения своих лингвистических свойств часто ведет себя как класс людей, а не физических предметов, ср. [Ляшевская 2004].

(д) В конструкциях вида: *До X <осталось> 3 часа / 3 дня / неделя...* предлог *до* управляет существительными, обозначающими временной ориентир. Сформулировав запрос:

***до* + сущ.: род. п. + сущ.: время** (*до концерта ... неделя*)

или: ***до* + сущ.: род. п. + (числ. + сущ.: время)** (*до концерта ... 3 часа*)

мы можем выяснить, какие семантические классы существительных в русском языке способны обозначать временные ориентиры. Наряду с ожидаемыми классами:

[8] Мы не можем здесь входить в детали семантического основания такого поведения лексики в генитивной конструкции субъекта — на эту тему имеется чрезвычайно обширная литература, ср., например [Babby 1980], а также [Падучева 1997], [Борщев, Парти 2002] и др. Речь идет только о том, что корпус представляет возможность изучить эту конструкцию с той же степенью подробности, с какой это сделано в [Mustajoki, Heino 1991] для отрицательной конструкции с генитивом объекта.

1) время (включая подклассы «время суток»; «день недели»; «месяц» и т. п.): *до **пяти** еще 3 часа*; *до **понедельника** еще три дня*;
2) события с фиксированной датой, в том числе праздники: *до **зарплаты** / до **Нового года** — неделя*;
3) мероприятия (у которых время назначается): *до **концерта** / до **конференции** — неделя*

сюда попадают еще и транспортные средства. То есть названия транспортных средств могут обозначать не только пространственный ориентир (*До самолета было не так далеко, а он всё шел*, И. Грекова), но и — метонимически — временной (*До самолета в Рим — неделя*, А. Азольский)[9]. В первую очередь это относится к таким транспортным средствам, которые ходят по расписанию: так, возможно *до самолета еще неделя / до автобуса полчаса* — но невозможно (во временном значении): **до кареты / BMW еще 15 минут*. Можно сказать *До машины еще полчаса*, но в таком случае машина ведет себя как «маршрутное» транспортное средство (прибытие которого заранее планируется и ожидается).

Таким образом, данный запрос к Корпусу выявляет еще одну группу лингвистически интересных особенностей класса «транспортные средства».

Запросы подобного рода, касающиеся одновременно морфологической и лексико-семантической информации, служат мощным инструментом лингвистического анализа. Если морфологическая разметка русских текстов, в особенности со снятием омонимии, расширяет возможности исследователя прежде всего в области морфологии и лексикографии (поскольку лингвист получает доступ к примерам на употребления форм слов или слова во всех формах), то включение семантических признаков расширяет возможности изучения *конструкций русского языка* (о понятии конструкций см. [Шведова 1966], [Апресян 1967]; ср. также [Золотова 1988], а кроме того теоретические работы Ч. Филлмора [Fillmore 1989]; [Fillmore 1996]; [Fillmore, Kay 1992]; А. Голдберг [Goldberg

[9] Речь идет о метонимическом обозначении времени отхода: *до поезда еще час* = ‘до отхода поезда еще час’; время прибытия транспортного средства с точки зрения встречающих так обозначить нельзя (ср. **до поезда еще час* → *до прибытия поезда еще час*), что имеет свое объяснение.

1995], У. Крофта [Croft 2001] и др. в области Грамматики конструкций.

6.2. Проблема снятия семантической неоднозначности как инструмент исследования языка

Одна из глобальных задач семантической разметки — снятие семантической неоднозначности.

Разные значения слова часто (хотя и не всегда) относятся к разным семантическим классам, и, следовательно, получают разные семантические пометы, ср. глагол *ныть*: 'звук' (*Ноет саксофон*) / 'речь' (*Не ной*, *никто тебя не пожалеет*) / 'физиологическое ощущение' (*Ноет рука*). В машинном словаре, обслуживающем Корпус, семантические признаки приписываются отдельно каждому значению слова, однако при автоматической семантической разметке текста каждое вхождение слова получает всю совокупность помет, которые есть у разных значений этого слова в словаре (например, у любого вхождения *ныть* будут признаки 'звук', 'речь' и 'ощущение'). Это создает шум при семантическом поиске. Поэтому нужно убрать «лишние» признаки и у каждого вхождения слова оставить тот единственный правильный признак, который соответствует его значению в данном контексте. Для Корпуса объемом в 100 млн. словоупотреблений это непростая задача. И дело не только в оптимизации семантической разметки — вопрос имеет более общий и принципиальный характер. С нашей точки зрения, для Национального корпуса русского языка проблема фиксации (определения, выделения) отдельных значений многозначных слов[10], в принципе, имеет два способа решения.

Первый способ — снимать неоднозначность «вручную», т. е. усилиями эксперта-разметчика (таким «ручным» способом была снята морфологическая омонимия для подкорпуса объемом около 5 млн. словоупотреблений). Второй способ — снимать неоднозначность автоматически, разработав какую-то новую программу семантической дизамбигуации в тексте.

---

[10] В практике автоматической обработки текстов эта проблема называется word sense disambiguation (т.е. проблема разрешения лексической многозначности) и имеет свою историю; подробное обсуждение подходов к ее решению, применяемых в настоящее время в мировой практике, их достоинств и слабых мест приводится в обзоре [Кобрицов 2004а].

«Ручной» способ решения этой проблемы предлагается, в частности, в недавней публикации С. Св. Волкова и В. П. Захарова [Захаров, Волков 2004]: квалифицированные лингвисты, опираясь на данные авторитетного словаря, идентифицируют значения слов в тех или иных текстовых употреблениях и составляют электронную картотеку примеров на употребления слов в отдельных значениях, подобную традиционным «бумажным» картотекам. Однако если учесть, что в 100-миллионном корпусе на каждое слово могут быть получены тысячи примеров, то становится ясно, что для одного только распределения этих примеров по значениям, зафиксированным словарями (не говоря уже об учете новых значений, возникающих в современном языке), понадобится такое количество высококвалифицированных лексикографов и такие значительные временные и финансовые ресурсы, которые в требуемых объемах едва ли удастся мобилизовать.

При этом нужно понимать, что стратегия в области морфологической разметки и снятия морфологической омонимии оказывается принципиально другой, чем при разметке лексико-семантической: частотность *лексем* в тексте на много порядков ниже частотности *грамматических значений*. Поэтому, если для морфологических исследований корпус со снятой омонимией объемом 5-10 млн словоупотреблений будет заведомо достаточен, для изучения лексики он наверняка окажется слишком мал: ступив на такой путь, можно получать пусть и качественные, но относительно небольшие и не слишком представительные выборки примеров на отдельные значения, несопоставимые по объему с бумажными картотеками академических институтов, над созданием которых работали целые коллективы специалистов на протяжении многих десятилетий.

Другой путь — автоматическое снятие семантической неоднозначности с использованием семантической разметки.

Здесь тоже есть возможность выбора решения. Одно из них — статистическое. Известно, что имеются статистические методики, позволяющие прогнозировать решение (в нашем случае, выбор значения) на основе тестового корпуса, в котором омонимия снята. Эффективность таких методик может быть очень высока — до 90%, в зависимости от решаемой задачи. В частности, именно эти методики применяют разработчики Чешского национального корпуса для снятия морфологической омонимии. Но опять-таки, вви-

ду того, что разница в частотности морфологических и лексических значений, как мы уже говорили, огромна, минимальный объем тестового корпуса в этих случаях для эффективной работы статистики тоже оказывается несопоставим. Так, если в Чехии морфологическая омонимия была снята на корпусе примерно в 1 млн. словоупотреблений (тот же объем, видимо, был бы достаточен и для русского языка), то, по оценкам экспертов, хорошим тестовым корпусом для разработки программы по снятию лексической многозначности будет корпус объемом 100 млн. словоупотреблений. Как видим, в таком случае задача сводится к предыдущей: ручной обработке огромных массивов.

Мы предлагаем не статистическое, а, так сказать, лингвистическое решение данной проблемы для Корпуса: создание системы лингвистических правил (фильтров), которые бы позволяли снимать неоднозначность для целых классов употреблений лексемы в тексте. Эти фильтры будут приводиться в действие программно, а значит, лексическая неоднозначность в Корпусе будет сниматься все-таки не вручную. Разумеется, эвристика фильтров, их разработка и проверка потребуют существенных усилий от лингвистов-экспертов; однако фильтры такого рода, как нам представляется, имеют самостоятельное значение, причем и для грамматики данного языка, и для проектов его автоматической обработки.

Поясним суть предлагаемого решения.

Известно, что каждое значение многозначного слова реализуется в определенном семантическом контексте и, как правило, связано с одной или несколькими конструкциями. Задача, таким образом, состоит в том, чтобы фильтры задавали нужные контексты и конструкции — тогда применение фильтров позволит автоматически убирать признаки, не соответствующие значению слова в данном контексте, то есть выбирать нужное значение. Фильтры могут опираться или только на морфологию, или на морфологию и семантику, или только на семантику или даже только на лексику. Заметим, что во всех случаях, кроме последнего, фильтры позволяют работать не просто с отдельными словами, но и с целыми классами слов — и это тоже способствует ускорению и автоматизации работы по снятию неоднозначности в Корпусе.

Рассмотрим примеры фильтров, использующих как морфологические, так и семантические признаки.

(А) У слов *куча, груда, туча, пропасть, капля, горстка* и под. имеется неоднозначность «физический объект» vs «квантификатор» (ср. *капля дождя* vs. *капля терпения*). В семантически размеченном тексте при каждом таком слове будет две пометы, например: *Тучи* {физич. объект|квантификатор} *плыли совсем низко над землей.*

Фильтры, позволяющие снять одну из помет, будут иметь следующий вид:

a) *туча* без следующего за ним сущ. в род. п. → {физич. объект}, ср.: *грозовая туча*;
b) *туча* + сущ.: род. п. → {квантификатор}, ср.: *туча народа*;
c) *куча* + сущ.: вещество: сыпучее & ед., род. п. → {физич. объект|квантификатор}, ср.: *куча песка* (*куча песка* — это физический объект, но одновременно слово *куча* указывает на большое количество);
d) *куча* + предмет & мн., род. п. → {физич. объект|квантификатор}, здесь указание класса (‘предмет’) не приводит к снятию неоднозначности, т. к. возможны обе интерпретации, ср.: *Он перебил уже кучу тарелок* (‘количество’) и *На полу он увидел кучу старых ботинок* (т. е. ботинки были свалены в кучу; ‘физич. объект’);
e) *куча* + сущ.: вещество & мн., род. п. → {квантификатор}, ср.: *куча вин*;
f) *куча* + сущ.: человек/животное & род. п. → {квантификатор}, ср.: *куча детей*;
g) *куча* + сущ.: абстр. & мн., род. п. → {квантификатор}, ср.: *куча неприятностей*;

(Б) У глагола движения *довести* имеется неоднозначность, поэтому в Корпусе вхождения этого глагола получают двойную помету {движение|воздействие на человека}, ср.: *довести до подъезда* vs. *довести до бешенства.*

Однако если мы зададим ограничение для существительного (фильтр):

довести + *до* + сущ.: род. п. & **эмоциональное состояние или проявление** (ср.: *довел до истерики/до бешенства/до исступления/до слез/до инфаркта*),

то сможем удалить помету «движение».

Попутно выясняется, что, с одной стороны, не все эмоции допустимы в этой конструкции, т. е. не все слова с эмоциональной семантикой в русском языке могут обозначать высшую, предельную степень проявления внутреннего состояния (ср. **довел до страха* / **до возмущения*), а с другой стороны, что некоторые болезненные состояния и проявления типа *инфаркта* и *слез* приравниваются в русской языковой картине мира к предельным эмоциональным состояниям.

(В) Субъектом глагола поведения (*кривляться*, *важничать*, *буянить*) является лицо. Это дает возможность построить фильтр, разрешающий неоднозначность для таких глаголов, для которых «поведение» не является единственным значением (ср. *забыться, распуститься, рисоваться, ломаться, красоваться*). В контексте неодушевленного субъекта помета «поведение» у них снимается, в контексте названия лица — сохраняется, ср.:

Приборы ломались; *Шутки забылись*; *Ветка распустилась*; *Звезда красовалась на груди*; *Стены рисовались темным массивом* vs. *Петя ломается / рисуется / красуется* и т. д.

Принцип семантического фильтра можно использовать и при поиске: по заданным параметрам контекста (в частности, по заданной конструкции) получать примеры употребления слова в определенном значении.

Например, разные значения слова *серп* имеют семантические характеристики ‘инструмент’ или ‘форма’. Чтобы получить примеры употребления слова *серп* в значении ‘инструмент’ (*серп крестьянина*), нужно задать ограничение:

серп + сущ.: человек & род. п.;

чтобы получить примеры употребления слова *серп* в значении ‘форма’ (*серп месяца*), задаем ограничение:

серп + сущ.: предмет & род. п.

Разумеется, снятие неоднозначности с помощью фильтров, как любая серьезная и масштабная задача, имеет свои сложности. Во-первых, какую-то часть лексики, возможно, всё-таки придется обрабатывать вручную (чтобы на этой основе создать исходный набор фильтров); во-вторых, как уже было сказано, потребуются значительные интеллектуальные усилия как лингвистов-экспертов, так и программистов по созданию семантических фильтров; в-третьих, задача снятия неоднозначности на всем 100-миллионном

корпусе, конечно, не может быть решена в полном объеме; наконец, сама работа фильтров неизбежно будет давать какое-то количество «шума».

С другой стороны, у этого пути есть очевидные и очень существенные преимущества. Решая важную практическую задачу по снятию неоднозначности в Корпусе, мы попутно получаем массу полезной лингвистической информации о свойствах слов и конструкций, которая может использоваться для решения самых разных исследовательских и прикладных задач, в том числе и для совершенствования самого Корпуса — уточнения номенклатуры семантических признаков и состава соответствующих им классов слов и значений (подробнее об этом см. [Кобрицов 2004б]).

Сказанное относится и к работе с Корпусом в целом. Решение внутренних, технических задач (уточнение семантической рубрикации, разработка фильтров и др.), помимо практической важности получаемых результатов, имеет значительные научные перспективы: оно не только расширяет информационно-поисковые возможности пользователей и повышает качество результатов поиска, но и дает огромный материал для теоретических выводов и обобщений.

Библиография

Апресян 1967 — Апресян Ю. Д. Экспериментальное исследование семантики русского глагола. — М.: Наука, 1967.

Апресян 1974 — Апресян Ю. Д. Лексическая семантика: Синонимические средства языка. — М., 1974.

Апресян и др. 2004 — Апресян Ю. Д., Иомдин Л. Л., Санников А. В., Сизов В. Г. Семантическая разметка в глубоко аннотированном корпусе русского языка // Труды Международной конференции «Корпусная лингвистика — 2004». — СПб.: Изд-во Санкт-Петербургского университета, 2004. — С. 41-54.

Бабенко 1999 — Бабенко Л. Г. Толковый словарь русских глаголов: Идеографическое описание. Английские эквиваленты. Синонимы. Антонимы. М.: АСТ-Пресс, 1999.

БАС — Словарь современного русского литературного языка. Т. 1-17. — М.-Л., 1948-1964.

Борщев, Парти 2002 — Борщев, В. Б., Парти, Б. Х. О семантике бытийных предложений // Семиотика и информатика. Вып. 37. М.: ВИНИТИ, 2002. С. 59-77.

Грамматика 1980 — Русская грамматика. Т. 1-2. — М., 1980.

Даль 1907 — Даль В. И. Толковый словарь живого великорусского языка. Тт. I–IV. — СПб.—М., 1907.

Зализняк 1972 — Зализняк А. А. Русско-французский учебный словарь. — М., 1972.

Захаров, Волков 2004 — Захаров В. П., Волков С.Св. Корпус текстов и исторический словарь // Русский язык XIX века: проблемы изучения и лексикографического описания. СПб.: Наука, 2004.

Золотова 1988 — Золотова Г. А. Синтаксический словарь: Репертуар элементарных единиц русского синтаксиса. — М., 1988.

Кобрицов 2004а — Кобрицов Б. П. Модели многозначности русской предметной лексики: глобальные и локальные правила разрешения омонимии. Автореф. ... канд. филол. наук. — М, 2004.

Кобрицов 2004б — Кобрицов Б. П. Методы снятия семантической неоднозначности // НТИ, сер.2. — 2004. — № 3. — С. 15-27.

Красильщик, Рахилина 1992 — Красильщик И. С., Рахилина Е. В. Предметные имена в системе «Лексикограф» // НТИ, сер. 2. — 1992. — № 9. — С. 24-31.

Кузнецова 1989 — Кузнецова Э. В. Лексико-семантические группы русских глаголов. — Иркутск, 1989.

Кустова 2004 — Кустова Г. И. Типы производных значений и механизмы языкового расширения. — М.: Языки слав. культуры, 2004.

Кустова, Падучева 1994 — Кустова Г. И., Падучева Е. В. Словарь как лексическая база данных // ВЯ. — 1994. — № 4.

Ляшевская 2004 — Ляшевская О. Н. Семантика русского числа. М.: Языки слав. культуры, 2004.

МАС — Словарь русского языка в четырех томах. Т. I-IV. — М., 1957-1961.

Ожегов 1972 — Ожегов С. И. Словарь русского языка. — Изд. 9-е, испр. и доп. Под ред. Н. Ю. Шведовой. — М., 1972.

Ожегов, Шведова, 1999 — Ожегов С. И., Шведова Н. Ю. Толковый словарь русского языка. — 4-е изд. — М.: Азбуковник, 1999.

Падучева 1985 — Падучева Е. В. Высказывание и его соотнесенность с действительностью. — М., Наука, 1985.

Падучева 1996 — Падучева Е. В. Семантические исследования: Семантика времени и вида в русском языке. Семантика нарратива. — М.: Языки рус. культуры, 1996.

Падучева 1997 — Падучева Е. В. Родительный субъекта в отрицательном предложении: синтаксис или семантика? // ВЯ. — 1998. — № 5. С. 3-23.

Падучева 2004а — Падучева Е. В. Динамические модели в семантике лексики. — М.: Языки слав. культуры, 2004.

Падучева 2004б — Падучева Е. В. О параметрах лексического значения слова: онтологическая категория и тематический класс // Русский язык сегодня. — Вып. 3. Проблемы русской лексикографии. — М., 2004.

Рахилина 2000 — Рахилина Е. В. Когнитивный анализ предметных имен: семантика и сочетаемость. — М.: Рус. словари, 2000.

Саяхова и др. 2000 — Саяхова Л. Г., Хасанова Д. М., Морковкин В. В. Тематический словарь русского языка / Под ред. В. В. Морковкина. — М., 2000.

Ушаков 1935-1940 — Ушаков Д. Н. (ред.). Толковый словарь русского языка. Тт. I–IV. — М., 1935-40.

Чардин 2003 — Чардин И. С. Лингвистические корпусы с разметкой на основе грамматики зависимостей и их применение при автоматическом синтаксическом анализе, Дисс... канд. филол. наук. — М., 2003.

Шведова 1966 — Шведова Н. Ю. Активные процессы в современном русском синтаксисе. — М., 1966

Шведова 2000 — Русский семантический словарь. Толковый словарь, систематизированный по классам слов и значений. Под общ. ред. Н. Ю. Шведовой. Т. 1-4. — М.: Азбуковник, 2000.

Эйдельман 1977 — Эйдельман Б. Ю. Библиотечная классификация и систематический каталог. Учебное пособие. — М.: Книга, 1977.

Babby 1980 — Babby L. Existentional Sentences and Negation in Russian. — Ann Arbor: Caroma Publishes, 1980.

Croft 2001 — Croft W. Radical Construction Grammar. — Oxford: Oxford University Press, 2001.

Fillmore 1989 — Fillmore Ch. J. Grammatical construction theory and the familiar dichotomies. // In R. Dietrich and C. F. Graumann, eds., Language processing in social context. — Amsterdam: Elsevier. 1989. — P. 17-38.

Fillmore 1996 — Fillmore Ch. J. The pragmatics of constructions. // In Dan I. Slobin, ed., Social interaction, social context, and language. — Mahwah: New Jersey: Lawrence Erlbaum Associates, Inc, 1996.

Fillmore, Kay 1992 — Fillmore Ch. J., Kay P. Construction Grammar Course Book. — Berkeley, 1992.

Goldberg 1995 — Goldberg A. E. Constructions: A Construction Grammar Approach to Argument Structure. — Chicago: Chicago University Press, 1995.

Levin 1993 — Levin B. English verb classses and alternations: A preliminary investigation. — Chicago: Chicago UP, 1993.

Mustajoki, Heino 1991 — Mustajoki A., Heino H. Case Selection for the Direct Object in Russian Negative Cases. Part II. — Helsinki 1991.

Wierzbicka 1985 — Wierzbicka A. Lexicography and conceptual analysis. — Ann Arbor: Karoma, 1985.

Wierzbicka 1988 — Wierzbicka A. The semantics of grammar. — Amsterdam: Benjamins, 1988.