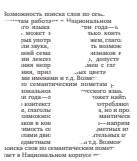
Е. В. Рахилина,
Г. И. Кустова,
О. Н. Ляшевская,
Т. И. Резникова,
О. Ю. Шеманаева

Задачи и принципы семантической разметки лексики в НКРЯ[1]

1. Введение


Возможность поиска слов по семантическим пометам работает в Национальном корпусе русского языка вот уже три года—пользователь может найти не только контексты, в которых употребляются, скажем, глаголы запаха или звука, но и проверить возможность сочетаний семантических признаков в комбинации лексем—например, допустимость сочетания непредметных имен с глаголами движения, прилагательных цвета с непредметными именами и т.д.

Исходно лексико-семантическая классификация в НКРЯ базировалась на принципах системы «Лексикограф» (*http://www.lexicograph.ru*, [Красильщик, Рахилина 1992; Кустова, Падучева 2004]); при этом для целей разметки Корпуса эта система была

[1] Исследование выполнено при частичной финансовой поддержке Российского фонда фундаментальных исследований, грант №08-06-00197-а.

в свое время существенно изменена и дополнена, см. [Кустова и др. 2005]. Однако уже будучи интегрирована в корпус, семантическая разметка продолжает меняться и совершенствоваться. Технология этих изменений такова: имеется базовая нотация, по которой есть поиск в открытом доступе, и экспериментальная, которая проходит апробацию—ею пользуются только разработчики Корпуса. После тестирования новые пометы внедряются в систему общедоступного поиска. В частности, в самые ближайшие планы входит расширение системы семантических помет за счет включения новых топологических типов имен и новых словообразовательных классов.

Понятие топологического типа имени восходит к [Talmy 1983], где обращается внимание на лингвистическую релевантность геометрических характеристик объектов внешнего мира. Мы применяли его к широкому русскому материалу в работах, касающихся именной сочетаемости, см. [Рахилина 2000, Десятова и др. 2008] и показали, что имена физических объектов, относящихся к классам «поверхности», «контейнеры», «веревки» и т.д. по-разному сочетаются с пространственными операторами—такими как прилагательные размера и формы, пространственные предлоги, глаголы локализации и движения и др. Сегодня поиск в Корпусе идет только по топологическим признакам «поверхность» и «контейнер», планируется добавить в поисковую форму признаки «выступ», ср. *нос, бородавка, грудь, балкон* и др., «вертикальная поверхность», ср. *забор, стена, стенд* и др., «отверстие», ср. *дыра, горлышко, окно,* а также ряд других топологических признаков.

Что касается словообразовательных помет, то в Корпусе уже сейчас доступен поиск разнообразных дериватов: приставочных глаголов, вторичных имперфективов (глаголов на *-ыва-* типа *выпивать*), семельфактивов (на *-ну* типа *мигнуть*), а также—в зоне предметных существительных—диминутивов (ср. *домик*), аугментативов (ср. *домище*), в зоне прилагательных—каритивов (ср. *безглазый, бездыханный*) и некоторых других. В ближайшее время станет возможен поиск словообразовательного класса сложных слов (ср. *авианосец, густонаселенный, боготворить* и др.).

С другой стороны, помимо «плановых» изменений имеющаяся на сегодняшний день разметка редактируется, так сказать, «внепланово»—благодаря замечаниям пользователей корпуса. Одно-

временно, помимо частных помет, интерес у пользователей—конечно, прежде всего у активных пользователей—вызывают и сами принципы, заложенные в основу корпусной разметки. Например, Алексей Кретов обратился к нам с целой статьей по этому поводу—ее мы публикуем ниже—она стала хорошим стимулом для нас, чтобы еще раз продумать возможные альтернативы «семантических шагов», предпринятых в свое время в Корпусе. Таким образом, следующий раздел нашей статьи будет посвящен обсуждению общей идеологии корпусной разметки в семантической зоне (раздел 2), а затем—на примере конкретных спорных решений—мы обсудим «приложение» этих принципов—сначала к разметке как таковой (раздел 3), а потом—к снятию семантической омонимии (раздел 4).

2. Наши цели

2.1 Лексико-семантическая классификация и корпусная разметка

Сегодня создано множество лексико-семантических классификаций, в том числе на русском материале—см., например, [Кузнецова 1989, Бабенко 1999, Шведова 2000]; есть и примеры компьютерных систем, опирающихся на такого рода классифицирование лексики, ср. например, систему WordNet для разных языков мира (*http://wordnet.princeton.edu*), онлайн-словарь английских глаголов VerbNet (*http://verbs.colorado.edu/~mpalmer/projects/verbnet.html*), также посвященный глаголам ресурс VerbOcean (*http://demo.patrickpantel.com/Content/verbocean*) или систему USAS (Lancaster, *http://ucrel.lancs.ac.uk/usas*), не говоря уже о базе данных «Лексикограф». Они отражают чисто семантический подход к лексической классификации, подразумевающий максимально дробную признаковую базу. Действительно, чем больше используется семантических признаков, тем надежнее (за счет дробности классификации) можно предсказать сочетаемостные особенности конкретных слов. Лучше всего эти задачи решает лексическая база данных с максимально жесткой структурой и максимально повторяющимися признаками—транскатегориальными, т.е. действующими в зоне любой части речи, так что, например, признак ‘движение’ характеризует и глагол *идти*, и прилагательное *пеший*, и существительное *ноги*. Пользователь такой базы данных оперирует списками лексем разной степени общ-

ности, которые могут быть релевантны для решения самых разных задач—от составления списков квазисинонимов или, скажем, онтологий для информационного поиска до сопоставления лексических систем разных языков.

Скажем сразу, что данный (чисто семантический) подход в полном объеме на нашем Корпусе реализован быть не может—во-первых, по техническим причинам. Многоступенчатая семантическая разметка, порождая все новую и новую омонимию, значительно бы «утяжелила» и без того объемный Корпус (по своему «весу» приближающийся к 200 млн словоупотреблений, к тому же снабженных морфологической и метатекстовой информацией), существенно замедляя его работу вплоть до угрозы сбоев при поиске. Во-вторых, даже если бы техника выдерживала такую нагрузку без снижения быстродействия, транскатегориальный подход к корпусной разметке устроил бы далеко не всех. Действительно, какому пользователю понравилось бы, если бы на запрос ‘движение’ выдавался огромный массив предложений, содержащий не только глаголы и отглагольные имена, но и прилагательные типа *быстрый / медленный*, а также предметные имена типа *ноги, колеса, лыжи* и т. д. и т. п., и даже существительное *часы* (они ведь тоже *ходят*!). А ведь именно с таким эффектом мы столкнемся, если, по предложению А. А. Кретова, «отменим» частеречные противопоставления, действующие сегодня в системе семантических классов Корпуса.

Нам скажут: такое легко исключить, запросив только грамматический класс глаголов с пометой «движение». А если пользователю нужны как раз отпредикатные имена (типа *хождение, вращение, полет* и др.)? Здесь грамматические ограничения не помогут, и в ответ на запрос о ‘движении’ будут выданы все те же *часы* вкупе с *лыжами*. Если же исследователю действительно интересны предметные имена, связанные с ‘движением’, он и в нынешней версии семантической разметки может запросить все субъекты (т.е. предшествующие глаголу существительные в именительном падеже) при глаголах движения, ср. запрос:

сущ. & им. п. + глагол : движение & личная форма

С лингвистической точки зрения ответ многомиллионного корпуса на этот запрос будет гораздо точнее, чем априорное классифика-

ционное решение лингвиста-разметчика, которое базируется исключительно на его интуиции. В отличие от лингвиста, корпус не будет «раздумывать» над тем, приписывать ли помету ‘движение’, прямо скажем, нестандартным с этой точки зрения именам типа *часы, дорога, дым, пар, газ* и им подобным, а просто выдаст исследователю весь объем сочетаний—чтобы тот дальше мог выбрать нужные ему лексемы по своему усмотрению, а не был вынужден следовать чьей-то интуиции. Собственно, именно поэтому разработчики корпусной разметки, опираясь на базу данных «Лексикограф», которая в части, касающейся предметной лексики, содержит для каждого имени информацию о функциональном предикате, см. [Красильщик, Рахилина 1992], сознательно «отрезали» эту семантическую зону, исключив ее из дерева разметки.

2.2. Древесная VS. фасетная классификация

Кстати, о деревьях. На первом этапе работы над разметкой мы считали, что наша классификация должна быть не древесной, а фасетной. Древесный принцип в чистом виде, реализованный, например, в Семантическом словаре Шведовой [2000], где предметное имя попадает *или* в класс контейнеров, *или* в класс приспособлений, а глагол—*или* в класс речевых, *или* в класс посессивных, для Корпуса не годится, и нужно иметь возможность приписывать слову несколько семантических помет сразу, что как раз и отражает идею фасетности, см. [Кустова и др. 2005: 160]. Однако в ходе работы выяснилось, что и фасетный принцип в чистом виде как основа корпусной разметки тоже оказывается опасным. Действительно, он хорошо работает и широко применяется в Корпусе для поиска по полностью независимым признакам—скажем, с одной стороны, таксономическим, как ‘движение’, ‘лицо’, ‘физическое качество’, т.е. отражающим собственно онтологию, а с другой—так сказать, «квазиграмматическим» пометам—отражающим мереологию (‘части’–‘целое’ & ‘элементы’–‘множество’), топологию (‘поверхности’, ‘контейнеры’ и др.), словообразование (уменьшительные суффиксы, приставки и др.), оценку (положительная / отрицательная) и под. Именно за счет такой комбинации (и даже практически всегда только за счет нее) возникает сложная многопризнаковая разметка в се-

мантической зоне предметных имен[2].

Другое дело—возможность фасетной организации разметки внутри чисто таксономических признаков, которые часто не-независимы друг от друга. Ведь при таком способе структурирования семантической информации в один и тот же класс попадают, скажем, и глаголы, у которых данный признак является вершинным, и те, у которых он совершенно второстепенный. В качестве примера удобно вернуться к признаку ‘движение’. Всякий человек (даже и не лингвист) знает, что такое «глаголы движения»—это *бегать, лететь, плыть, вертеться, катиться* и т.д. и т.п.—довольно большой класс (общий его объем по нашей базе данных составляет для русского языка более 1000 единиц). Все это те глаголы, в толковании которых признак ‘движение’ является базовым или, говоря в синтаксических терминах, занимает вершинную позицию. Но если иметь в виду глубокую детальную семантическую разметку, ориентированную на систематизацию лексики, о которой говорит А. А. Кретов, то по признаку ‘движение’ придется разметить гораздо большее количество глаголов, у которых этот признак входит в толкование, но не как вершинный. Тогда на запрос о глаголах движения в Корпусе найдутся не только предложения с «классическими» предикатами типа *бегать* или *лететь*, но и, например, предложения с глаголом *закрыть* <*дверь*> (≈ ‘каузировать дверь, д в и г а я с ь , начать находиться в контакте со стеной’), и отделить их друг от друга будет невозможно. Понятно, что пользователь в этом случае будет разочарован, а значит, практическая задача, которую Корпус призван решать, не будет выполнена. Однако такой «провал» прикладных функций не случаен, он имеет и теоретическое объяснение.

Фактически идеология «универсальной» семантической разметки (о которой, в частности, идет речь в работе А. А. Кретова и которая при поиске дает эффект фасетности в полном объеме) восходит к семантическим примитивам Г. В. Лейбница и компонентному анализу Й. Трира и Дж. Катца. Для них такое разложение на минимальные смыслы было самоценно и представляло собой

[2] Примером, иллюстрирующим принцип возникновения редких исключений здесь может служить комбинация ‘вещества и материалы’ и ‘еда и напитки’, ср. *сахар, творог, спирт* и т.п.

самостоятельную научную проблему, ориентированную на поиск универсального метаязыка. Конечно, с тех пор прошло много лет и сменилось много лингвистических теорий, но и сегодня жива точка зрения, согласно которой решение этой задачи могло бы способствовать построению лексической типологии и диахроническим исследованиям лексики. Это не так. И теория [Atkins, Fillmore 2000; Lakoff 1987], и практика (ср. [Viberg 2001, Goddard 2003, Majid, Bowerman 2007]), в том числе и собственные исследования по лексической типологии авторов этой статьи [Копчевская-Тамм, Рахилина 1999; Майсак, Рахилина 2007, Резникова и др. 2008] говорят о том, что восприятие лексики носителями и ее классификация в естественном языке опирается не на дискретные признаки, а на целостные гештальты. Именно поэтому для семантического моделирования в лексической типологии гораздо удобнее использовать теоретический аппарат фреймов и конструкций, который как раз апеллирует к «не-независимости» отдельных семантических признаков друг от друга. Так, признак ‘движение’ в семантике глагола *закрыть* настолько необходим для перехода объекта в результирующее состояние, что является неотъемлемой частью этой ситуации. В этом смысле идея движения для глагола *закрыть* ни с точки зрения типологии, ни с точки зрения диахронии, скорее всего, релевантна не будет, потому что она присутствует в соответствующей внеязыковой ситуации обязательно.

В то же время, в семантике многих глаголов (а соответственно, и отпредикатных имен со значением ситуации) есть не одна (как у предметных имен), а две в равной степени базовые таксономические зоны—причем достаточно независимые друг от друга: это способ действия и результат. Соответствующие им признаки организуются фасетно и ищутся независимо друг от друга. Именно так устроен глагол *вытребовать*, о котором шла речь в [Кустова и др. 2005: 160]: с одной стороны, *вытребовать*—это посессивный глагол, квазисинонимичный таким как *взять, получить, приобрести* и под., а с другой—для него, как и для глагола *требовать*, важна речевая составляющая, описывающая способ действия. По тому же принципу размечены в Корпусе глаголы *ткнуться* (‘движение’ + ‘контакт’), *барабанить* (‘движение’ + ‘звук’), *мелькать* (‘движение’ + ‘восприятие’), *продрогнуть* (‘изменение состояния’ + ‘физиоло-

гическая сфера') и др. под[3]. Понятно, что этих двух признаков недостаточно ни для полного описания соответствующих глаголов, ни для их типологического сравнения с другими языками. Но поскольку Корпус в принципе не может ставить перед собой задачу «описания лексико-семантической системы русского языка» (см. статью А. А. Кретова в настоящем сборнике), это и не так важно. Его задача—обеспечение максимально удобного поиска примеров для максимально широкого круга пользователей. Что же можно сделать для решения этой задачи?

По нашему опыту, пользователю Корпуса легче формулировать запросы, апеллируя к базовым категориям—и именно они лучше всего приспособлены для такой пользовательско-ориентированной системы, как Корпус. Если говорить о глаголах, то это ментальные, речевые, позиционные, бытийные, движения, контакта и др., если о прилагательных—цвета, размера, формы и др., в сфере предметной лексики—лица, вещества, инструменты и проч. С одной стороны, такие классы интуитивно понятны неподготовленному пользователю (хотя в Корпусе все равно каждая такая помета прямо в таблице снабжена всплывающей подсказкой и в будущем планируется разместить на сайте списки классов), а с другой—именно на эти базовые классы, как выясняется, опирается большинство правил выбора значения при разрешении многозначности (см. раздел 4). Ясно, что оба эти обстоятельства вовсе не случайны: как раз такого рода свойства и лежат в основе определения базовой лексики.

Конечно, базовые классы могут дальше специфицироваться—уже по древесной схеме, так что, например, вещества будут делить-

[3] Понятно, что сам таксономический признак далеко не всегда просто сформулировать. Например, для разбиравшегося выше глагола *закрыть*, который относится к классу 'физическое воздействие' наряду с *резать, целовать, нажимать, касаться* и др. под., определить результат не так уж просто. С сугубо теоретической точки зрения, это, наверное, мог бы быть 'контакт', но всегда контакт предмета с предметом (двери с притолокой, например). Между тем класс глаголов контакта интуитивно определяется (видимо, ввиду общей антропоцентричности картины мира) как состоящий из глаголов, способных описывать контакт предмета с человеческим телом – ср. те же *целовать, нажимать, касаться*. В таких трудных случаях лучше, конечно, оставить лексему недоопределенной – именно такая стратегия и принята в Корпусе.

ся на жидкие, твердые и газообразные, а физические свойства—на форму, цвет, температуру и проч. Одновременно на таксономическое дерево в Корпусе, как мы уже говорили, накладывается еще несколько «квазиграмматических» классификаций, и комбинация этих признаков уже происходит по фасетной схеме. При этом «прозрачность» классификации, конечно, сохраняется: если *здание* относится к топологическому типу контейнеров, то и его разновидность—*дом*—тоже.

Итак, дело не в том, что разработчики Корпуса случайно или по недосмотру допускают непоследовательности в использовании древесного или фасетного принципов классификации, а в том, что, учитывая специфику своего продукта и его отличия от лексических баз данных и словарей, они вполне сознательно *отказались* от этих принципов как однозначной догмы и применили более эффективную в условиях он-лайнового поиска стратегию их совмещения. Конечно, такой подход не дает возможности (и даже не ставит задачи) построить общезначимую надъязыковую онтологию на базе универсальных лексико-семантических констант, а проще говоря, компонентного анализа или (внечастеречных) семантических множителей—зато позволяет довольно эффективно искать если не отдельные слова по заданному семантическому признаку, то по крайней мере эти же слова в составе последовательностей словоформ.

2.3 Семантика и синтаксис

И здесь мы переходим к ответу на еще один распространенный упрек: почему же в Национальном корпусе русского языка нет синтаксической разметки?

Во-первых, строго говоря, она есть: в рамках семейства подкорпусов имеется небольшой экспериментальный синтаксически размеченный подкорпус (см. *http://www.ruscorpora.ru/search-syntax.html*). Работа над ним показала, насколько это трудоемкая задача. У осуществляющей этот проект лаборатории ИППИ РАН под руководством Л. Л. Иомдина к началу работы имелся огромный опыт такого рода деятельности в рамках работ по машинному переводу; имелся и задел—в виде серии систем ЭТАП на базе русского поверхностного синтаксиса, принятого в модели «Смысл⇔Текст», а также пилотного корпуса новостных текстов, уже размеченных

к тому времени тем же анализатором. Тем не менее потребовалось 6 лет для того, чтобы разметить корпус в пределах полумиллиона словоупотреблений. Если даже представить себе, что дальше работа будет продвигаться в разы быстрее, то для такой разметки всего массива НКРЯ потребуются десятки лет. Одновременно детальная синтаксическая разметка в том виде, в котором она принята в синтаксическом подкорпусе, требует не только профессиональной подготовки разметчика, но и дополнительной подготовки пользователя—«новичку» она недоступна.

Таким образом, подробный синтаксический анализатор не может быть пока применен к Корпусу в целом—во-первых, ввиду его объема, а во-вторых, ввиду отсутствия единой—одновременно общезначимой и общедоступной—модели русского синтаксиса. Можно было бы пофантазировать и попытаться себе представить, как мог бы выглядеть специальный модуль корпусного синтаксиса—чтобы он был и общезначимым, и общеполезным, и автоматическим. Один из вариантов решения этой проблемы нам видится в том, чтобы указывать сам факт синтаксической связи, не специфицируя ее природу. Можно ли будет добиться на этом пути интересного результата—пока до конца не ясно.

Вместе с тем неправы те, кто говорит, что сейчас в НКРЯ нет *никакой* синтаксической разметки, см., например, [Копотев, Мустайоки 2008]. Во-первых, в Корпусе имеется частеречная разметка—а это не только морфология, но и синтаксис; плюс к этому—(морфологическая по природе) информация о падежном маркировании: она тоже дает представление о синтаксических связях. Во-вторых, не так давно была введена опция поиска по знакам препинания, так что теперь на всем массиве текстов можно находить вопросы и восклицания, а также вводные слова или сложноподчиненные предложения разных видов. Все это, конечно, не полноценный синтаксис, но, что называется, *элементы синтаксиса в Корпусе* [там же]. Не забудем и о возможности задавать строгий порядок следования единиц поиска. Таким образом, в совокупности для запросов оказываются доступны *конструкции*—т.е. (как правило) сложные синтаксические единицы со своим значением, часто фиксированным набором и порядком лексических переменных, заданным грамматическим оформлением и лексическим наполнением разной степени свобо-

ды: от почти застывших фразеологизмов до свободных сочетаний с минимальными ограничениями на составляющие.

Термин «конструкция» удобен тем, что, как говорится, «проверен временем» и до сих пор используется самыми разными школами, причем примерно в одном и том же значении. Главную особенность конструкций лучше всего эксплицировал Ч. Филмор в теории Грамматика конструкций [Fillmore et al. 1988], см. также [Goldberg 1995]: конструкция—это минимальная языковая единица, в которой ограничения разного уровня (морфологические, лексические, семантические, синтаксические, а иногда и фонетические) взаимозависимы, так как мотивированы семантикой конструкции в целом. Филмор же предложил компьютерную модель для своей теоретической идеи—систему Framenet (см. *http://framenet.icsi.berkeley.edu*), в которой воплощается комплексная, многоступенчатая разметка контекстов употребления лексических единиц.

Понятно, что НКРЯ, в сущности, воплощает ту же идею: лексическая семантика в языке существует не сама по себе, а в теснейшей связи с так называемым «малым синтаксисом» (см. также последние работы Л. Л. Иомдина на эту тему, например, [Иомдин 2003]), следовательно, семантическая разметка в Корпусе должна встраиваться в морфосинтаксическую и взаимодействовать с ней. И действительно, наиболее эффективен Корпус тогда, когда задан сложный запрос, комбинирующий лингвистическую информацию разной природы. В этом случае он, во-первых, незаменим, потому что никакая обычная интернет-поисковая система в принципе не может осилить такой запрос (а ведь как часто критики говорят, что корпуса не нужны—достаточно интернета!). Во-вторых, именно в сложных запросах (а не в запросах по одному независимому признаку), в том числе с учетом семантических параметров, пользователь получает наиболее аккуратную выдачу, с минимальным шумом, который как раз и снимается дополнительными условиями поиска.

Более того, именно возможность построить запрос на конструкцию, характеризующуюся, в частности, определенными семантическими признаками, позволяет оперировать существенно более простой системой помет, не перегружая ее лишними параметрами. Например, теоретически можно было бы (как предлагает А. А. Кретов) приписать значению слова *утихнуть* помету weather: fin, напри-

мер, *(метель) утихла*. И действительно, с этим глаголом сочетаются и *дождь,* и *буря,* и *вьюга*, и *шторм,* и *гроза* и т. д. Однако природные явления, как показывает соответствующий запрос, составляют лишь малую часть субъектов глагола *утихнуть*—среди них есть и *крик,* и *голос,* и – метонимически—имена, обозначающие людей (*женщина, ребенок* и др.), а также *ненависть, аплодисменты, боль* (и даже—метонимически—*висок*) и др. В то же время *утихнуть,* как и все глаголы с подобным значением, легко находится в современной версии разметки при поиске конструкции: непредметное имя класса «природное явление» + глагол «прекращения существования».

Теперь суммируем все сказанное о принципах выделения семантических классов для корпусной разметки. По нашему мнению, классифицирующие таксономические признаки должны быть:

- независимыми,
- базовыми,
- выделять крупные классы,
- порождать минимальный шум,
- оптимальный результат при их использовании можно ожидать в случае сложного поискового запроса (т.е. конструкции).

3. Разметка: вопросы и ответы

Итак, принципы обозначены. Но реальный словарь, который лежит в основе семантической базы данных, очень большой, а его разметка предполагает преимущественно ручную работу. И конечно, здесь могут быть ошибки и непоследовательности, так что процесс «чистки» семантического словаря идет непрерывно. Мы благодарны всем нашим «семантическим» оппонентам, и прежде всего А. А. Кретову, за то, что они своими вопросами и замечаниями помогают нам в этой работе. Однако здесь мы хотели бы обсудить не случайные ошибки, а принципиальные решения и сложные случаи—в качестве иллюстрации наших теоретических установок.

3.1. Независимость признаков

О необходимости этого принципа мы говорили выше. Теперь о трудностях. Трудности его применения хорошо иллюстрируются материалом имен собственных.

В Корпусе собственные имена представляют собой отдельный класс—наравне с предметными и непредметными, так что им свойствен свой тип разметки. Это очень естественно, потому что в число собственных имен не входят, с одной стороны, ни инструменты, ни вещества, ни иные классы конкретной лексики, а с другой—ни периоды времени, ни звуки, ни иные классы абстрактной лексики. Одновременно, собственные имена не являются ясным подклассом ни для предметных, ни для непредметных имен—они бывают и теми, и другими (ср. *МГУ* как здание—предметное имя—и «*Кинотавр*» как мероприятие—абстрактное имя). Именно поэтому система их разметки представляется в Корпусе как независимая от других имен. Пока она включает только имена, отчества, фамилии, топонимы, а также словообразовательные корреляты—стяженные формы (типа *Николаич* и др.) и аббревиатуры (типа *МММ, ГРУ* и под.). В дальнейшем могут быть добавлены клички животных, марки машин и другие дополнительные разряды.

Эта работа, однако, не так проста, как кажется, потому что здесь мы столкнемся с практически обязательной полисемией типа: *Волга*–топоним / «*Волга*»–марка машины, *Васька*–кличка кота и *Васька*–имя человека, «*Стрела*»–название поезда и *стрела*–предметное имя, *Форд*–фамилия человека и «*Форд*»–марка машины и т. д. и т. п. Именно поэтому разработчики не торопятся с простым расширением числа помет в этой зоне: до того, как разрешится проблема снятия омонимии, оно не будет способствовать оптимизации поиска, а наоборот, только «утяжелит» Корпус за счет дополнительной многозначности.

Раз собственные и предметные имена представляют разные классы, пометы из этих классов оказываются в разметке независимо друг от друга. Нужно только проследить, чтобы эта независимая разметка была проведена. Проведена она для имен лиц, так что в Корпусе можно найти и *Александров*, и *Сергеевичей*, и *Пушкиных* по запросу на имена лиц, но, конечно, только если убрать из поисковой строки ограничение на конкретность имени. Тогда собственные имена найдутся наравне с нарицательными. Не проведена такая разметка для местоположений, и пока названия городов и стран не ищутся как представители таксономического класса

space (а только на запрос «топоним» или его объемлющий класс «собственное имя»).

Независимая разметка предметных и собственных имен имеет тот недостаток, что пользователю требуется дополнительная подсказка о том, как ему найти одновременно все существительные со значением «лица», включая имена, фамилии и отчества. По умолчанию он получит только нарицательные существительные и должен будет дополнительно искать контексты с собственными именами. Но альтернатива, которая предложена А. А. Кретовым, тоже не кажется нам оптимальной. Его решение ввести разметку типа *t:hum:persn* для имен (*Александр*), *t:hum:patrn* для отчеств (*Сергеевич*) и *t:hum:famn* для фамилий (*Пушкин*) осуществимо только в том случае, если мы аннулируем имена собственные как отдельную категорию, т.е. фактически сделаем их подклассами конкретных. Тогда потребуется очень дробная (а значит, всегда априорная) их классификация, понадобится снятие омонимии, а главное—сама табличка выбора признаков неизбежно потеряет системность. Действительно, рядом с именами лиц, инструментов, веществ, пространств и т.д. в ней обнаружится класс имен собственных как таковых, не попавших ни в какой из дробных разрядов (например, *«Марсельеза»*, *ГТО* и прочие аббревиатуры). Хорошо бы, наверное, придумать в этой зоне что-то третье.

3.2. Базовость признаков

Значения многих важных классов («поведение», «мероприятие», «возраст», «изменение состояния или признака» и др.) со строго семантической точки зрения не являются элементарными. Но и разработчикам, и пользователям важно иметь именно такие классы для поиска—в частности потому, что они активно участвуют в конструкциях, задавая семантические ограничения на лексическое наполнение последних. Тем самым нужно, чтобы эти классы оставались в поисковой табличке как целостные единицы.

Другой вопрос, насколько удобно использовать внутри самой лексической базы данных Корпуса их разложение на более элементарные компоненты: иными словами, можно ли попробовать в базе заменить помету *behave* на ее составляющие—*hum:act:neg* (см. статью А. А. Кретова), при том что в поисковой табличке все равно

останется признак «поведение»? Или для глаголов *взрослеть, твердеть, богатеть* заменить помету *changest* («изменение состояния») на «составную» *incep:be:diff*?

Нельзя. Базовый класс на то и базовый, чтобы существовать особняком, не смешиваясь с другими. Разложение базового «гештальта» на составляющие пересечет его со всеми теми классами, признаки которых входят в его состав. Тогда глаголы изменения состояния будут искаться на запрос о бытийных, а поведение или возраст—на запрос о человеке и т.д. и т.п. Это сразу нарушит принцип «не порождать лишнего шума» и существенно затруднит работу пользователей.

3.3. Принцип крупных классов

В корпусе есть помета «физические свойства» (*t:physq*). Она введена ради противопоставления классу «свойства человека» (*t:humq*), которое нужно, в частности, для снятия неоднозначности в случаях переноса признаков с предмета на человека (*мягкий хлеб* → *мягкий человек*). Обе пометы должны присутствовать и в прилагательных (ср. *крепкий* VS. *добрый*), и в отпредикатных именах (*крепость* <*чая*> VS. *добродетель*), но пока в полном объеме они применяются только к адъективной лексике. Конечно, как и всякая помета, *t:physq* достаточно условна, так что если говорить об обозначаемых ею свойствах, то они не столько физические, сколько эмпирически наблюдаемые, воспринимаемые органами чувств—ср. ‘вкус’ или ‘запах’ (хотя, разумеется, органами чувств они воспринимаются потому, что имеют в конечном счете именно физическую природу). Условность этой пометы проявляется и в том, что к физическим относятся и «потенциальные» качества типа *растворимый*, которые важно противопоставить тоже потенциальным, но «нефизическим» прилагательным—таким как *неотвратимый* или *непредсказуемый*.

Представить *physq* и *humq* как составные пометы с общей частью (*q*) и противопоставленными *phys* и *hum* не удастся по только что указанным в разделе 3.2 причинам: тогда человеческие качества получат отдельную помету *hum* как часть *hum:q* и пересекутся с классом людей в целом, а значит, будут выдаваться по запросу об именах лиц. Это неудобно для пользователей. Но и для разработчиков тоже:

выясняется, что различие между *hum* и *humq* может использоваться для снятия неоднозначности в глаголе, ср. *Добродетель* (*humq*) *украшает человека* vs. *Девочка* (*hum*) *украшает елку*. Таким образом, эти классы как раз очень хорошо противопоставлены семантически и, по нашему мнению, просто не нужны как объединение.

В принципе, для аналогии с прилагательными, можно снабдить класс непредметных имен ‘цвет’, как предлагает А. А. Кретов, дополнительной пометой *physq*. Поиск это не ускорит, но, безусловно, добавит системности в разметку. Однако нужно понимать, что в любом случае в зоне прилагательных мы не можем полностью распределить все ‘физические свойства’ по классам, поскольку для них нет общеизвестных помет. Например, более спорным выглядит решение о присвоении словам *мягкий, вязкий* необщепринятого признака *plast*—такой класс (в отличие от ‘цвет’ или ‘форма’) пользователю незнаком. Но даже если согласиться и принять это решение, оно, что называется, не спасет положения, потому что в класс ‘физические свойства’ входят еще и такие прилагательные, как *слабый, сильный, пористый, слоистый, пуленепробиваемый, растворимый, горючий, прозрачный, жидкий, глинистый, песчаный, каменистый* и т.д., для которых уж точно не найдется общепонятных помет. Мелкие классы из одного-двух слов неудобны, плохо воспринимаются, загромождают поисковую форму и по всем этим причинам не годятся для корпусной разметки. И наоборот, общий класс ‘физические свойства’ оказывается и психолингвистически, и технически релевантным.

Другой интересный случай касается глаголов восприятия, которые, безусловно, являются базовыми в любом естественном языке—просто в силу его антропоцентричности.

В словаре Корпуса таких глаголов порядка двух сотен, однако бóльшая часть этого списка—глаголы ***зрительного*** восприятия (*смотреть, глядеть, любоваться, глазеть* и др., а также их приставочные корреляты) и лишь небольшая—все остальные. Поэтому если приписывать пометы *smell, taste* глаголам обоняния, вкусового восприятия и др., мы получим крайне маленькие и – как всегда в таких случаях—сомнительные классы. Например, глагол *нюхать*, на базе которого строилась бы вся группа запаха (*нанюхаться, понюхать, принюхиваться, разнюхать*), строго говоря, не является гла-

голом запаха. Еще хуже дело обстоит с осязанием: единого класса осязания обычно не выделяется, потому что прототипического глагола осязания нет, а свойства, воспринимаемые осязанием, очень разные (ср. перечисляемые в статье А. А. Кретова *мягкий, вязкий, тяжёлый, лёгкий и горячий, ледяной*).

В такой ситуации для пользователя, конечно, проще составлять не семантические, а «лексические» запросы с конкретными глаголами, т.е. вместо семантического запроса с признаками «восприятие: обоняние» формулировать запрос, в котором фигурирует непосредственно глагол *нюхать* и его приставочные корреляты.

Что касается глаголов зрительного восприятия, то, поскольку это достаточно мощный и единый класс, странно было бы его делить (как предлагает А. А. Кретов), сопоставляя с пометами прилагательных *light* и *color*. Да и как делить? Тем более что с помощью зрения человек оценивает не только свет и цвет, но также и форму, которая связана, в частности, еще и с осязанием, а помимо этого—местоположение предметов, расстояние, размер и многое другое! Не говоря уже о том, что зрительно восприниматься могут не только предметы, но и ситуации (*Видел, как они входили в подъезд*). Так что, пожалуй, тут все правильно: пусть класс глаголов восприятия остается базовым, а нужные уточнения пользователь в каждом конкретном случае легко сделает сам.

Итак, с практической точки зрения, в Корпусе должны использоваться пометы, которые достаточны или просто удобны для поиска—а это имена больших таксономических классов, в которых один признак определяет и семантические характеристики, и совокупность синтаксических свойств.

4. Семантическая разметка и снятие лексической неоднозначности

Надо сказать, что семантическая разметка, принятая в Корпусе, проходит еще одну апробацию: она задействована в правилах снятия семантической омонимии (подробнее см. Рахилина и др. 2006, Шеманаева и др. 2007, Толдова и др. 2008). Проект снятия семантической омонимии все еще находится в стадии разработки, однако для этой цели уже создано много «фильтров»—в основном для качественных прилагательных, предметных имен и глаголов, и в них

задействованы те самые семантические признаки, по которым строится поиск. Надо сказать, что некоторые принципы работы с этими тремя классами слов различаются, поэтому все дальнейшее касается только качественных прилагательных[4].

Вкратце суть этого проекта такова. Большая доля слов русского языка многозначна, ср. знаменитые *лук* ('растение') и *лук* ('оружие'), *худой* ('не толстый') и *худой* ('плохой'), *колоть* ('воздействовать иглой') и *колоть* ('болеть') и мн. др. Однако в потоке речи эта омонимия «снимается» более широким контекстом, так что говорящий и слушающий ее не замечают. Задача состоит в том, чтобы «научить» машину реагировать на релевантный контекст и, таким образом, разрешить семантическую многозначность хотя бы для самых частотных случаев. Вот тут как раз и оказываются задействованы семантические признаки—и самих многозначных слов, потому что разные значения нужно как-то отличить друг от друга, а значит, разметить семантически,—и тех слов, которые составляют их ближайшее окружение, так как часто выбор значения многозначного слова зависит именно от семантического класса соседней лексемы.

Разрабатывая правила снятия многозначности, или фильтры, разметчики пользуются уже имеющимися в Корпусе признаками, тем самым составление фильтров оказывается своеобразной «экспертизой» для наших семантических помет. Оказывается, что в основном для снятия омонимии классов, уже имеющихся в корпусе, достаточно, и необходимость добавления новых возникает крайне редко. Значит, говорящие на естественном языке, выбирая значения, опираются на довольно простые и общие свойства слов, и к тем же самым свойствам обращается пользователь Корпуса при поиске, пытаясь «угадать конструкцию».

Конечно, на эту тему можно сказать еще многое в теоретическом плане—и о том, как соотносится эта идея с современными семантическими теориями, и о том, что дает такой прикладной эксперимент для лексической типологии, для теории построения универсального метаязыка, или насколько подобная практика ин-

[4] Ср. статью Г. И. Кустовой и С. Ю. Толдовой в настоящем сборнике, касающуюся снятия многозначности в глаголах.

тересна с психолингвистической точки зрения—но мы хотели бы в данной статье всего лишь проиллюстрировать сказанное несколькими ясными примерами.

Первый пример демонстрирует важность разряда соседнего с прилагательным существительного, т. е. его принадлежности к предметным или непредметным именам. Это одно из базовых противопоставлений, крайне существенных для развития многозначности адъективной лексики. Так, прилагательное *легкий* означает физическое свойство (‘нетяжелый’) ровно в тех случаях, когда оно относится к предметному имени; дальнейшее разграничение его значений ведется с использованием таксономических классов непредметных имен. Поэтому один из его фильтров будет выглядеть так:

Слово	**Контекст**	**Итоговое значение**
легкий	+«предметное»	SEM= разряд: «качественное», таксономический класс: «физическое свойство: вес»

Понятно, что в правилах учитывается и более дробная классификация, прежде всего, таксономическая. Так, среди значений прилагательного *голый* принято различать по крайней мере следующие:

- ‘неодетый’, ср. *голый человек,*
- ‘неприкрытый’, ср. *на голом полу*,
- ‘чистый, без примесей’, ср. *голый спирт*,

и у каждого из этих значений есть свои ограничения на таксономический класс существительного. Их можно сформулировать в терминах наших семантических признаков:

Слово	**Контекст**	**Итоговое значение**
голый	+ «лица»	SEM = разряд: «качественное», таксономический класс: «физическое состояние»
голый	+ «пространство и место»	SEM2 = разряд: «качественное», таксономический класс: «внешний вид»
голый	+ «вещество	SEM2= разряд: «качественное», таксономический класс: «физическое свойство»

Хороший пример использования непредметных классов дает прилагательное *холодный*. Среди его значений есть следующие:

- ‘низкий (о температуре)’ ср. *холодный ветер*,
- ‘оттенок цвета’, ср. *холодные цвета*,
- ‘относящийся к человеку—его ментальной / эмоциональной / психологической сфере или поведению’, ср. *холодный взгляд*.

Здесь можно сформулировать следующие контекстные правила:

Слово	**Контекст**	**Итоговое значение**
холодный	+«природное явление» +«время»	SEM=разряд: «качественное», таксономический класс: «физическое свойство: температура»
холодный	+«цвет»	SEM2=разряд: «качественное», таксономический класс: «физическое свойство: цвет»
холодный	+«ментальная сфера» +«психическая сфера» +«свойство человека» +«поведение и поступки человека»	SEM2=разряд: «качественное», таксономический класс: «свойство человека»

Надо сказать, что параметр таксономического класса, каким бы эффективным он ни был, все же не покрывает всех тонкостей и различий в семантике прилагательных. Так, два разных значения лексемы *редкий* используются с существительными одного и того же таксономического класса «растения», ср. *редкая трава* (‘растет на большом расстоянии друг от друга’) и *редкое растение* (‘то, которое редко встречается’). Здесь «помогает» мереологическая разметка: в контексте существительных класса «множества и совокупности объектов» прилагательное *редкий* может выступать только в значении расстояния:

Слово	**Контекст**	**Итоговое значение**
редкий	+«растение»&«совокупности объектов»	SEM=разряд: «качественное», таксономический класс: «расстояние»

Полезной в плане различения значений прилагательных может оказаться и топология предметных имен (т.е. их геометрические характеристики). Например, прилагательное *тугой* в сочетании с существительными, представляющими класс «вместилища», имеет значение большого размера (*тугой кошелек*), тогда как в контексте имен, называющих неодушевленные объекты других топологических классов, оно отсылает к физическому свойству, не связанному с размером (что-то вроде ‘крепкий’), ср. *тугой узел*.

Слово	**Контекст**	**Итоговое значение**
тугой	+«вместилища»	SEM2=разряд: «качественное», таксономический класс: «размер: большой»
тугой	+«предметные»	SEM=разряд: «качественное», таксономический класс: «физическое свойство»

Конечно, сказать, что выделенных в Корпусе семантических классов для правиловых фильтров хватает всегда (с учетом топологии и мереологии), все-таки было бы преувеличением. Система семантических помет постоянно совершенствуется—в том числе благодаря фильтрам. Например, практика составления контекстных правил показала, что класс «профессии» релевантен не только с энциклопедической, но и с лингвистической точки зрения. Так, у слов *старший* и *младший* конкурируют два значения: ‘старший по возрасту’ и ‘старший по иерархии’. Оба значения представлены в контексте существительных класса «лица», однако второе значение оказывается возможным только при лексемах, образующих особый подкласс среди имен лиц—существительных, называющих профессии. Соответственно, добавив класс «профессии» в систему семантических помет корпуса, мы сможем отфильтровать контексты, в которых слова *старший/младший* используются во втором значении:

- *старший* + «профессии»: *старший* ‘иерархия’;
- *младший* + «профессии»: *младший* ‘иерархия’;
  (ср. *старший / младший научный сотрудник, лаборант, офицер* и др.)

В сочетании с другими существительными класса «лица» описываемые прилагательные получают первое значение:

- *старший* + «лица»: *старший* ‘возраст’;
- *младший* + «лица»: *младший* ‘возраст’;
  (ср. *старший / младший брат*)

Таким образом, процесс изготовления фильтров интересен для нас не только как прикладная задача—снятие омонимии в Корпусе, но одновременно и как задача теоретическая. На этом материале становится ясно, какие семантические классы слов одного лексико-грамматического разряда обуславливают семантическую многозначность слов другого лексико-грамматического разряда. Очевидно, что в зоне прилагательных ключевыми являются противопоставления «одушевленных» (включая «лица») и «неодушевленных», а также «предметных» и «непредметных» имен: мена между этими классами существительных всегда ведет к сдвигу семантики прилагательного. Существенным, однако, представляется вопрос, какие еще классы имен релевантны для различения значений в адъективной семантической зоне и—более того—как они связаны с типом семантического перехода в прилагательном, т. е. в каких случаях изменение одного таксономического класса на другой влечет за собой метонимический, а в каких—метафорический сдвиг. Такое исследование требует большого языкового материала—и в этом отношении Корпус и реализованная в нем семантическая разметка оказываются идеальной источниковой базой. В свою очередь, проведение такого теоретического исследования будет способствовать уточнению таксономических классификаций, принятых в Корпусе, и тем самым—совершенствованию системы семантической разметки НКРЯ.

Литература

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

Десятова А. В., Ляшевская О. Н., Махова А. А. Конструкция с творительным формы «X Y-ом» // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 4–8 июня 2008 г.). Вып. 7 (14). М.: РГГУ, 2008. С. 113–139.

Иомдин Л. Л. Большие проблемы малого синтаксиса // Труды международной конференции по компьютерной лингвистике и интеллектуальным технологиям Диалог'2003. М.: Наука, 2003.—С.216–222.

Копотев М. В., Мустайоки А. Современная корпусная русистика // Инструментарий русистики: корпусные подходы. Slavica Helsingiensia—34. Helsinki: Helsinki Univ.Press, 2008.

Копчевская-Тамм М., Рахилина Е. В. С самыми теплыми чувствами (по горячим следам Стокгольмской экспедиции) // Тестелец Я. Г., Рахилина Е. В. (ред.) Типология и теория языка: от описания к объяснению. Сб. к 60-летию А. Е. Кибрика. М.: Языки русской культуры, 1999.

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

Кретов А. А. Анализ семантических помет в национальном корпусе русского языка. Статья в наст. сборнике.

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

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.: Индрик, 2005.

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

Майсак Т. А., Рахилина Е. В. (ред.) Глаголы движения в воде: лексическая типология. М.: «Индрик», 2007.

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

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.

Резникова Т. И., Бонч-Осмоловская А. А., Рахилина Е. В. Глаголы боли в свете Грамматики конструкций // НТИ, сер. 2.—2008.—№ 4.—С. 7–15.

Толдова С. Ю., Кустова Г. И., Ляшевская О. Н. Семантические фильтры для разрешения многозначности в Национальном корпусе русского языка: глаголы // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 4–8 июня 2008 г.). Вып. 7 (14).– М.: РГГУ, 2008.

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

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.

Atkins S., Fillmore Ch. Describing polysemy: the case of *crawl* // Ravin Y. Leacock C. (eds.) Polysemy: Linguistic and computational approaches. Oxford: Oxford University Press, 2000.

Fillmore Ch. J., Kay P. & O'Connor K.T. Regularity and Idiomaticity in Grammatical Constructions: the Case of LET ALONE. In: *Language* 64, 1988, pp. 501–538.

Goddard, C. 2003. 'Thinking' across languages and cultures: Six dimensions of variation. *Cognitive Linguistics* 14(2/3), 2003, pp. 109–140.

Goldberg A. E. (1995) Constructions: A Construction Grammar Approach

to Argument Structure. Chicago: Chicago University Press, 1995.
Lakoff G. Women, fire and dangerous things: What categories reveal about the mind. Chicago: University of Chicago, 1987.
Majid, A., Bowerman, M. (eds.): Cutting and breaking events: A cross-linguistic perspective. Special issue of Cognitive Linguistics, 18(2) (2007)
Talmy, L. How language structures space. In: H. Pick and L. Acredolo (eds.), *Spatial Orientation: Theory, Research, and Application*. New York: Plenum Press, 1983, pp. 225–282.
Viberg Å. The verbs of perception // Haspelmath M. et al. (eds.) language typology and language universals: an international handbook. Berlin: de Gruyter, 2001.