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

СЕМАНТИЧЕСКАЯ РАЗМЕТКА И СЕМАНТИЧЕСКИЕ ФИЛЬТРЫ ДЛЯ НАЦИОНАЛЬНОГО КОРПУСА РУССКОГО ЯЗЫКА[1]

Национальный корпус русского языка снабжен разными видами разметки, в том числе – морфологической и семантической[2]. Семантическая разметка включает пометы таксономического класса, мереологию, оценку и др. типы информации. В разных видах разметки есть разные виды неоднозначности. Что касается семантической разметки, то связанная с ней проблема неоднозначности состоит в следующем.

В словаре у каждого значения многозначного слова есть своя собственная семантическая помета. Однако когда программа автоматически расставляет пометы в тексте, то она каждому вхождению слова приписывает все пометы, которые есть в словаре, потому что программа не знает, в каком значении выступает слово в данном тексте. Теперь нам нужно различить эти значения. Одним из эффективных путей решения этой проблемы являются семантические фильтры, т.е. семантические правила, которые позволяют оставлять при каждом вхождении слова только одну помету. Таким образом, многозначность снимается с точностью до семантического класса (т.е. с точностью до семантической пометы).

[1] Работа выполнена при поддержке РФФИ, проект № 05-06-80396.

[2] Мы подробно рассказывали об идеологии и видах разметки в наших докладах на предыдущих конференциях; см. также: *Кустова Г.И., Ляшевская О.Н., Падучева Е.В., Рахилина Е.В.* Семантическая разметка лексики в Национальном корпусе русского языка: принципы, проблемы, перспективы // Национальный корпус русского языка: 2003–2005. Результаты и перспективы. М., 2005.

Для снятия многозначности с помощью фильтров используется принцип контекстной однозначности. В словаре у слова может быть несколько значений, а в тексте (кроме специальных случаев языковой игры) – одно значение. Поскольку слово обычно не выступает в тексте изолированно, а включено в определенный контекст (в другой терминологии – в конструкцию), то этот контекст (конструкция) и является, в самом грубом и общем виде, фильтром. Этот подход – поиск ограничений на употребление слова в составе конструкций – связан, в частности (но не только), с постулатами Грамматики конструкций Ч. Филлмора[1].

Фильтр работает следующим образом: формулируются семантические, морфологические и синтаксические условия, в которых реализуется некоторое значение слова; эти условия записываются в виде контекста; осуществляется поиск слова в заданном контексте; результатом этого поиска будет корпус примеров, где слово выступает в заданном значении, соответствующем определенной семантической помете. Остальные пометы стираются. Далее процедура повторяется для остальных семантически размеченных значений слова.

Проиллюстрируем работу фильтра на простейшем примере. Слово *стопка-1* имеет значение ‘множество предметов, сложенных друг на друга’, *стопка-2* – значение ‘маленькая рюмка, маленький стаканчик’ (в данном случае не важно, что это разные слова, а не разные значения; значения различаются по тому же принципу) . В словаре у *стопки*-«множества» будет помета «совокупность», у *стопки*-«стаканчика» – помета «посуда». Чтобы их различить, нужно два фильтра:

стопка книг

стопка + сущ.: род.п.: предметы

[1] *Fillmore Ch, Kay P., O'Connor K.Th.* Regularity and Idiomaticity in Grammatical Constructions: The Case of Let Alone // Language. 64. 1988; *Goldberg A.* Constructions: A Construction Grammar Approach to Argument Structure. Univ. of Chicago Press, 1995.

стопка водки

стопка + сущ.: род.п.: еда и напитки

В примерах, отобранных первым фильтром, будет уничтожена помета «посуда» и оставлена помета «множество», в примерах, отобранных вторым фильтром, наоборот, останется помета «посуда».

Конечно, как и при любых статистических процедурах, здесь нельзя достичь полного охвата. Например, если слово употреблено не в этой конструкции, не в этом контексте, то оно не обрабатывается данным фильтром. Кроме того, всегда бывает какой-то «шум». Например, семантический класс может быть слишком крупным. Отдельного класса «напитки» нет, есть класс «еда и напитки»; следовательно, этот фильтр отберет, кроме *стопки водки, абсента, спирта, самогона, коньяка*, еще и *стопку блинов, лепешек, лаваша, печенья* (реальные примеры из корпуса). Однако сами эти ошибки имеют не меньшую ценность для совершенствования семантической разметки, а также для дальнейшего анализа и формулирования лингвистически интересных закономерностей, чем позитивный результат. В данном случае ошибки показывают, что для каких-то целей класс «еда и напитки» желательно разделить на «еду» и «напитки». Но практически такое разделение имеет смысл вводить в том случае, если оно будет работать для большой группы фильтров и существенно улучшать результаты поиска.

Вообще, любая работа с корпусом всегда имеет какие-то теоретико-лингвистические выходы. Это относится, конечно, и к неоднозначности. Неоднозначность – не только помеха при поиске, которая требует разрешения с помощью фильтров. Типы неоднозначности – важный раздел лингвистической теории, который имеет, в свою очередь, серьезное значение и для лексикографической практики. И хотя на эту тему есть много специальных работ, обычно они касаются типов многозначности у слов какой-то одной части речи (глаголов, прилагательных), или же

преимущественно освещают какой-то один тип многозначности (метафорические или метонимические значения). Но системно, массово и более или менее исчерпывающе такую работу можно проделать только на корпусе.

Разумеется, фильтры пишутся не для всех слов (не по алфавитному принципу). Один из важнейших приоритетов – частотность. В первую очередь будет сниматься многозначность для слов, которые встретились в корпусе 20 тыс., 15 тыс., 10 тыс. или хотя бы тысячу раз. Если же слово встретилось 200–300 раз, то в статистическом смысле оно не представляет особого интереса. Однако при создании фильтров учитывается и другой критерий – лингвистическая релевантность: слово может быть не очень частотным, однако лингвистически интересным.

Пока наибольшее количество фильтров сделано для существительных и прилагательных, поскольку они являются контекстом друг для друга.

Обратимся к фильтрам прилагательных. Вот фрагмент фильтра для прилагательного *тупой* (в настоящий момент насчитывается около 3000 вхождений; поскольку это прилагательное в числе самых многозначных, в качестве иллюстрации приводятся лишь некоторые значения, в частности, не включены метонимические значения *тупой взгляд*, *тупой ум*):

тупой

1. примеры: *тупой нож*, *тупая игла*, *тупое копье*, *тупой носок* (ботинка)

пометы значения прилагательного:

r:qual t:physq (= качественное + физическое свойство)

Примечание: в разметке есть помета «форма» (t:physq:form) – это подкласс физического свойства. Однако, чтобы не исключать из поиска инструменты (нож, игла), у которых свойство связано с функцией, оставлено более общее семантическое ограничение.

фильтр:

тупой + S r:concr t:tool:instr/t:tool:weapon/pt:part

(= *тупой* + сущ.: предметные имена: инструменты/оружие/часть (целого))
2. примеры: *тупой ученик*, *тупой бюрократ*, *тупой осел*
пометы:
r:qual t:humq ev:neg
(= качественное + качества человека + оценка: отрицательная)
фильтр:
тупой + S r:concr t:hum/ t:animal
(=*тупой* + сущ.: предметные имена: лица/животные)
3. примеры: *тупое отчаяние*, *тупая ненависть*, *тупое спокойствие*, *тупое упрямство*, *тупое безразличие*, *тупое равнодушие*, *тупая самоуверенность*, *тупое самодовольство*
пометы:
r:qual shift dt:humq ev:neg
(= кач. + сдвиг от «качества человека» + оценка: отрицательная)
фильтр:
тупой + S r:abstr t:psych:emot/t:humq/t:behav
(= *тупой* + сущ.: непредметные имена: психическая сфера (в том числе «эмоция»)/свойство человека/поведение и поступки человека)
Обороты: *тупая боль*

В фильтре есть две логические части. Одна – собственно фильтры, соответствующие разным значениям данного слова, вторая – обороты (коллокации). Обороты – тоже очень важная часть фильтра. У них есть как техническая функция, так и теоретическая. Техническая состоит в следующем. Статистически оборотов может быть очень много, и при запросе на какое-то слово вместе с обычными свободными сочетаниями будут выдаваться и обороты. Например, на запрос «прилагательное + *обязанность*» будет, в числе прочего, выдаваться *воинская обязанность*, на запрос «*дом* + сущ. род.» – *дом обуви*, *дом обоев*, *дом престарелых*, *дом отдыха* и т.д. Поэтому обороты должны размечаться хотя бы для того, чтобы при необходимости исключить их из поиска. С

другой стороны, пользователю могут быть нужны именно обороты, и тогда ему придется вручную выбирать их из сотен или тысяч примеров либо делать специальный запрос на каждый оборот в отдельности. Очевидно, что для пользователя гораздо удобнее, когда хотя бы самые частотные обороты выдаются автоматически.

Однако сам сбор оборотов – это еще и важная лингвистическая задача. И корпус принципиально облегчает ее решение. По существу, список оборотов – это вид фразеологического словаря. Как мы знаем, в языке есть сравнительно постоянный состав оборотов (вроде *железной дороги*) и есть, так сказать, мобильный: одни устаревают, исчезают, другие появляются (ср. *оранжевая революция, пищевая добавка, автогражданское страхование*). И это важный срез истории языка – то, чего еще нет в словарях и, возможно, никогда в них не войдет, потому что при отборе оборотов в бумажные словари учитывается и степень закрепленности в узусе, и стилистический регистр.

Другая интересная лингвистическая проблема, связанная с оборотами, – это их разметка. Должно ли слово в составе оборота сохранять свои исходные пометы? Это, в свою очередь, зависит от того, правомерно ли относить его к тому же семантическому классу, что и свободное значение, стоит ли вообще размечать обороты? Известно, например, что качественные прилагательные в составе оборотов становятся относительными (*мягкий вагон; тяжелая атлетика; тяжелая артиллерия; горячие блюда; холодные закуски; вредные привычки* и т.д.). Очевидно, что проблема разметки оборотов должна решаться параллельно с написанием фильтров.

Но вернемся к собственно фильтрам, т.е. семантическим правилам. Как уже было сказано, фильтры не только имеют очевидное практическое значение, помогая улучшать характеристики корпуса и повышать качество результатов поиска. Важный теоретико-лингвистический выход, который мы получаем при

работе с корпусом, – это уточнение существующих лингвистических представлений и усовершенствование лингвистической теории. В нашем случае это развитие теории многозначности, типов производных значений и моделей их образования.

Семантику традиционно интересовали типы производных значений и типы регулярных семантических сдвигов. В лингвистической литературе можно найти представительные списки таких регулярных переходов для слов разных частей речи (в том числе и для прилагательных)[1]. Корпус позволяет не только пополнить эти списки новыми типами семантических переходов, но и уточнить и систематизировать уже описанные и известные. Кроме того, поскольку эти переходы описываются на языке семантических классов, то в конечном счете наша задача – создавать фильтры не для отдельных слов и значений, а для целых семантических классов[2].

У прилагательных есть несколько типов помет, которые используются и в поисковых запросах, и в фильтрах.

Можно выделить 3 основных типа, если угодно – три уровня. Первый уровень – самый простой вид помет прилагательного – лексико-грамматический разряд (качественные и относительные). Второй уровень – наиболее общее семантическое подразделение: признаки предмета (физические свойства) и качества человека. Наконец, третий уровень – собственно семантические

[1] *Апресян Ю.Д.* Лексическая семантика. Синонимические средства языка. М., 1974; *Падучева Е.В.* Динамические модели в семантике лексики. М., 2004; *Рахилина Е.В.* Когнитивный анализ предметных имен: семантика и сочетаемость. М., 2000; *Шрамм А.Н.* Очерки по семантике качественных прилагательных. Л., 1979; и мн. др.

[2] В основе этого подхода – одна из главных идей системы «Лексикограф»: представителям одного семантического класса свойственны одинаковые типы семантических сдвигов; см., например: *Падучева Е.В.* Указ. соч.

(таксономические) классы внутри этих общих разрядов (цвет, размер, температура, вес, форма)[1].

Сама номенклатура признаков тесно связана как с техническими, так и с теоретическими проблемами. Например, есть ли какие-то специфичные типы переходов у прилагательных группы «цвет», «размер», «температура», или они все должны рассматриваться как «физические качества»? Сразу можно сказать, что по крайней мере у цвета и размера есть свои собственные характерные переходы, причем как от качественных к относительным, так и от относительных к качественным значениям. Например, качественные прилагательные цвета используются для политических и социальных характеристик в относительных значениях (*белая гвардия, красный командир*), но и наоборот, у многих относительных прилагательных (*вишневый, лимонный* и под.) развиваются качественные «цветовые» значения. Следовательно, и в семантических правилах (фильтрах) должен фигурировать не просто «физический признак предмета», а именно «цвет».

Другой важный как в теоретическом, так и в практическом плане вопрос: снабжать ли специальными семантическими пометами производные метафорические и метонимические значения. Пока мы в фильтрах пишем условно: *благородный человек* → *благородные намерения*: 1+shift_meton, т.е. метонимический сдвиг от первого значения; *умный человек* → *умная машина*: 1+shift_metaph, т.е. метафорический сдвиг от первого значения.

В процессе написания фильтров не только выясняется необходимость введения новых семантических признаков, но и обнаруживаются различные общие закономерности семантических переходов.

Например, есть два общеизвестных типа сдвига значения прилагательного: (1) признак предмета → признак человека (*тяжелый мешок – тяжелый человек*); (2) признак человека →

[1] Не все признаки сейчас доступны для поиска.

признак предмета/субстанции/ситуации (*веселый человек – веселый праздник*).

Первый тип сдвига (признак предмета → признак человека) преимущественно метафорический. То, что физический признак предмета применяется к нефизической сфере человека (*мягкий характер*; *жесткий человек*; *горячий джигит*; *широкая натура*) в режиме метафоры, имеет ясные когнитивные основания. Но есть ли в этом типе, где очевидным образом преобладает метафора, метонимические значения? Речь не идет о «вторичных» метонимических сдвигах (когда признак метафорически распространяется на сферу человека, а затем метонимически может применяться к различным характеристикам и проявлениям человека, ср. *мягкий человек* → *мягкий характер*; *мягкий упрек*), а именно о прямых метонимических переносах признака предмета на нефизическую сферу человека. На большом массиве примеров этого никто не проверял, поскольку большой массив примеров просто нельзя было получить. Сейчас появилась возможность проверить это если не на всем корпусе, то на весьма представительной его части – наиболее частотных словах.

Второй тип сдвига (признак человека → признак предмета /субстанции/ситуации) – преимущественно метонимический, и это тоже имеет когнитивную мотивированность[1].

Метонимический перенос признаков человека на внешние объекты и ситуации лучше изучен и описан, поскольку метонимический перенос более системный и регулярный. Однако и здесь есть большое поле для дальнейшей детализации. Например, можно выделять подклассы прилагательных, дающих метонимические сдвиги на разных, так сказать, основаниях.

В частности, здесь есть два характерных класса – условно говоря, эмоциональный и агентивный. Метонимический перенос

[1] См., в частности: *Кустова Г.И.* Типы производных значений и механизмы языкового расширения. М., 2004. Часть II.

признака в этих классах происходит на совершенно разные типы объектов.

В эмоциональном классе происходит метонимический перенос признаков и состояний человека (*грустный*, *веселый*, *радостный*, *печальный* и т.п.), во-первых, на внешние проявления этих состояний (*грустный голос; печальный взгляд*, *веселый смех*) и, во-вторых, на явления и события внешнего мира, вызывающие данные состояния (*грустный пейзаж*; *радостная новость*; *веселый рассказ*)[1].

Агентивный класс (*аккуратный*, *внимательный*, *вдумчивый*, *осторожный*, *энергичный*) – это признаки человека, характеризующие его во время деятельности. Метонимически они характеризуют саму деятельность и ее результаты, в первую очередь – ситуации, но иногда и предметы, ср.: *аккуратный ученик*; *аккуратное исполнение*; *аккуратный газон*. Этот перенос происходит, скорее всего, через стадию наречия: *аккуратно исполнил – аккуратное исполнение*; *аккуратно подстриженный газон – аккуратный газон*, *аккуратно сложенная стопка книг – аккуратная стопка книг*. Во втором типе (признак человека → признак предмета/субстанции) наряду с метонимическими значениями есть метафорические (*благородный человек – благородный металл*). Здесь опять-таки важно понять, сколько таких переходов в процентном отношении и можно ли среди них выделить какие-то однородные группы. И то и другое возможно (и имеет смысл) делать на большом корпусе. Национальный корпус русского языка, объем которого постоянно увеличивается за счет текстов разных жанров и разных исторических периодов развития русского языка, позволяет проделывать такую работу и получать статистически надежные результаты.

[1] *Кустова Г.И.* Указ. соч. Глава 9.