СЕМАНТИЧЕСКИЕ ФИЛЬТРЫ ДЛЯ РАЗРЕШЕНИЯ МНОГОЗНАЧНОСТИ В НАЦИОНАЛЬНОМ КОРПУСЕ РУССКОГО ЯЗЫКА: ПРИЛАГАТЕЛЬНЫЕ[1]

SEMANTIC FILTERS FOR THE WORD SENSE DISAMBIGUATION IN RNC: ADJECTIVES

Шеманаева О.Ю. *(shemanaeva@yandex.ru),* ***Кустова Г.И.*** *(galina03@mtu-net.ru),*
Ляшевская О.Н. *(olesar@mail.ru),* ***Рахилина Е.В.*** *(katia1@mail.ru),*
ВИНИТИ РАН

В статье представлена система семантических фильтров имен прилагательных, которая используется для разрешения неоднозначности лексико-семантической разметки в Национальном корпусе русского языка. Большинство значений многозначных прилагательных (как и других многозначных слов) в словаре Корпуса снабжено пометой семантического класса. В тексте каждое вхождение слова автоматически получает все словарные семантические пометы. С помощью семантических фильтров лишние пометы автоматически удаляются.

Мы уже писали в наших предыдущих публикациях о том, какова идеология семантических фильтров для Национального корпуса русского языка и какова технология их создания (см. [Кустова, Ляшевская, Падучева, Рахилина 2005], [Рахилина, Кобрицов, Кустова, Ляшевская, Шеманаева 2006]; о различных идеологиях семантической классификации лексики и семантической разметки см. также [Atkins 1993], [Fellbaum, Grabowski, Landes 1998], [Dolan, Vanderwende, Richardson 2002]). Поэтому сейчас мы ограничимся лишь короткой справкой. В Корпусе наряду с морфологической разметкой существует семантическая разметка: слову в электронном словаре Корпуса приписываются различные семантические пометы – таксономический класс (*кинжал*: «оружие», *печаль*: «эмоция», *кривляться*: «поведение»), мереология (*каблук*: «части одежды и обуви»), топология (*ниша*: «вместилище»), оценка (*благоухание*: «положительная оценка», *пресмыкаться* (перед кем): «отрицательная оценка»), некоторые словообразовательные пометы, важные для семантики (*проверка*: «отглагольное», *здешний*: «отадвербиальное»), и т.п. Помимо очевидной лингвистической ценности такой разметки, ее можно использовать еще и как средство разрешения многозначности слов в текстах Корпуса.

В обычных бумажных словарях разные значения идут под разными номерами. Это неудобно для пользователя Корпуса (поскольку он не помнит, сколько значений выделяет тот или иной словарь и под какими номерами они идут) и нерационально с точки зрения возможных лингвистических исследований, которые могут проводиться на Корпусе и для которых он (в числе многих других важных задач) был задуман и создан. В электронном словаре Корпуса вместо номеров (точнее, наряду с номерами) значениям приписываются семантические пометы, т.е. семантический класс, например: *обивка* (*закончена*) – «физическое воздействие», (*шелковая*) *обивка* – «вещества и материалы».

Основная проблема перенесения помет из словаря в тексты Корпуса состоит в следующем. В словаре Корпуса у многозначного слова обычно имеется несколько семантических помет (поскольку разные значения слова обычно, хотя и не всегда, принадлежат к разным семантическим классам). Когда программа автоматически расставляет пометы в тексте, то она каждому вхождению слова приписывает все пометы, которые есть у слова в словаре, поскольку программа не знает, в каком значении выступает слово в данном тексте, и не может выбрать единственную правильную помету, удалив все лишние. Это должны сделать семантические фильтры. Фильтр основан на принципе контекстной однозначности. В контексте слово выступает в определенном значении. Если сформулировать параметры контекста или контекстов, соответствующих данному значению, и сделать соответствующий поисковый запрос, Корпус выдаст примеры употребления слова в данном значении – которому соответствует определенная семантическая помета. Остальные пометы удаляются. И так для каждого значения. Таким образом, многозначность снимается с точностью до семантического класса (т.е. с точностью до семантической пометы). Тем самым пользователь получает возможность формулировать поисковые запросы и получать

[1] Данная работа выполняется при поддержке РФФИ, проект № 05-06-80396, и РГНФ, проект № 05-04-04008а

примеры, условно говоря, не на глагол *идти* в 5-ом (или 7-ом?) значении, а на глагол *идти* в значении ‘движение’ (*поезд идет*) или в значении ‘иметь место’ (*идет урок*).

Конечно, здесь есть ряд вопросов и трудностей.

Прежде всего, сами пометы должны быть относительно простыми и общепонятными (общепринятыми в лингвистическом обиходе). Это связано с общей идеологией Корпуса – он рассчитан на максимально широкий круг пользователей разного уровня подготовки: это и школьники, и студенты, и профессиональные исследователи языка, и программисты, и переводчики, и иностранцы, и мн. др. «Рядовой» пользователь должен представлять, какие слова стоят за той или иной семантической пометой и что он может получить на свой запрос. Например, есть помета «время». Естественно ожидать, что на запрос «сущ.: время» будут получены примеры со словами *час*, *пора*, *утро* и под., на запрос «прилаг.: время» – примеры со словами *вчерашний*, *ежегодный*, *довоенный*, *прошлый*.

В силу этого не все значения многозначных слов могут быть снабжены пометами. Например, значение глагола *копать*, реализуемое в предложениях вида *Он копает под меня*, не получает пометы. Не говоря уже о том, что трудно придумать помету, по которой пользователь догадался бы, что речь идет именно о таких значениях, трудно также найти достаточно большой и однородный класс производных значений, ради которых эту помету стоило бы вводить. В таких случаях мы вынуждены ограничиваться пометой «метафорическое значение». Т.е. пользователь может получить контексты с глаголом *копать* в значении физического действия (*копать траншею*; *копать огород*) и в метафорическом значении, куда войдут, помимо указанного значения *копать под кого*, также другие метафорические значения (ср. *Надо копать глубже* – в ситуации какого-либо расследования, ср. производное от него *раскопать новые факты*).

Таким образом, пометы (как уже существующие, так и вновь вводимые) должны отвечать двум главным требованиям: быть возможно более понятными и охватывать достаточно большие и лингвистически релевантные классы слов.

Все сказанное относится и к пометам прилагательных, а также к фильтрам, использующим эти пометы.

По данным на конец 2006 г., имена прилагательные составляют около 10,7% словоупотреблений Корпуса, половина из которых многозначны (имеют несколько несовпадающих наборов семантических тэгов). Созданная система семантических фильтров охватывает порядка 500000 употреблений наиболее частотных имен прилагательных и обеспечивает снятие лексико-семантической неоднозначности в среднем на 70%.

Наибольшую результативность как в количественном, так и в качественном отношении дают фильтры, которые можно записывать в виде конструкций «прилагательное + существительное» (с помощью таких фильтров можно снимать многозначность как прилагательных, так и существительных, о которых шла речь в нашей предыдущей публикации [Рахилина, Кобрицов, Кустова, Ляшевская, Шеманаева 2006]).

Поскольку свойства и структура многозначности качественных и относительных прилагательных существенно различаются, то и стратегии составления фильтров для этих двух групп различны.

Фильтры для качественных прилагательных

Основные качественные прилагательные, как правило, многозначны. Теоретически есть две стратегии расстановки семантических помет: либо попытаться приписать свою помету каждому значению (тогда существующих помет, хотя их и не так мало, все-таки не хватит, и придется придумывать новые), либо пренебречь какими-то различиями в пользу простоты и системности. Как ясно из сказанного выше, стратегической установкой разработчиков Корпуса является второй путь.

Есть типичные контексты, в которых значения качественных прилагательных меняются системно. Например, исходное значение у прилагательных *мягкий, жесткий, твердый, острый, тупой, горячий, холодный* и т.п. описывает признак физического объекта и реализуется, соответственно, в контексте существительных со значением физического объекта (с дальнейшей детализацией: предмет vs. вещество; твердое vs. сыпучее vs. жидкое vs. газообразное вещество): *мягкий диван*, *острая игла*, *горячий песок*, *холодная вода*. Основное условие сдвига значения – изменение таксономического класса определяемого существительного: человек; различные сферы внутреннего мира человека (интеллект, эмоции, воля); природные и техногенные процессы и явления; различные типы объектов и ситуаций ментальной, социальной, культурной сферы, ср.: *мягкий человек, тупой доцент, твердое решение, острый ум, холодный взгляд, мягкий климат, горячее копчение, острые противоречия, горячий прием, жесткий контроль* и т.п.

Поскольку перечисленные прилагательные чрезвычайно многозначны и все их значения различить невозможно, некоторые значения придется объединить. Как это сделать наиболее оптимальным способом? Рассмотрим прилагательное *мягкий* (для простоты значения будут задаваться не номерами, а минимальными контекстами).

Основные группы значений (группы контекстов) следующие: (а) *мягкая подушка, мягкий хлеб*; (б) *мягкий человек* → *характер, взгляд* (проявления мягкости человека); (в) *мягкий климат*; (г) *мягкий приговор* (мы берем лишь самые характерные контексты, поскольку в словаре у *мягкий* выделяется 7 значений, а с подзначениями – 16). У нас имеются две пометы: «физическое свойство» и «свойство человека». Исходное значение (*мягкая подушка*) имеет помету «физическое свойство». Для *мягкого человека* подходит помета «свойство человека». Для оставшихся значений нужно выбрать одно из двух решений: либо приписать им помету «метафора» (*мягкий человек*, разумеется, тоже метафора, но поскольку для этого значения помета есть, естественно ее использовать), либо придумать специальные семантические пометы для *мягкого климата*, *мягкого приговора*, *мягкого упрека* и т.д. Ясно, что такие пометы будут слишком дробными и слишком специальными. Поэтому в таких случаях мы ограничиваемся пометой «метафорический сдвиг».

Другим типичным (и системным) для качественных прилагательных семантическим процессом является метонимический сдвиг: человек → свойства, признаки или проявления человека (*веселый человек – веселый взгляд*; *холодный человек – холодная улыбка*; *тяжелый человек – тяжелый взгляд*; *добрый человек – доброе лицо*), ср. [Апресян 1974], [Кустова 2002], например:

1. *осторожный человек*

Признак значения: «качественное»; «качество человека» (r:qual t:humq).

Фильтр:

осторожный + сущ.: лицо

(= *осторожный* + t:hum/t:hum:kin/t:hum:etn/t:animal)

2. *осторожное обращение*

метонимический сдвиг от 1-го значения (1+shift_meton r:qual der:shift dt:humq)

Фильтр:

осторожный + сущ.: абстр.

(= *осторожный* + r:abstr)

Рассмотрим вкратце «технологическую цепочку» снятия многозначности качественного прилагательного на примере слова *круглый*.

КРУГЛЫЙ

Круглый имеет качественное значение формы (*круглое лицо, круглая луна*; помета «форма»), качественное значение высокой степени проявления какого-либо признака (*круглый дурак*, *круглый отличник*) и относительное значение, сочетающееся с мерой времени и числами (*круглые сутки, круглое число, круглая дата*).

Важное значение для пользователя имеет отделение коллокаций (устойчивых сочетаний) от свободных сочетаний. Специальная разметка коллокаций нужна, в частности, для того, чтобы они не выдавались пользователю при поиске свободных сочетаний. Так, например, на запрос «прилагательное цвета + *дом*» выдается несколько тысяч примеров, большая часть которых – коллокация *Белый дом*. Есть также небольшое количество коллокаций *желтый дом* и лишь незначительный остаток – материал, интересовавший пользователя. Естественно, что для изучения свободных значений коллокации должны быть заранее отделены от свободных сочетаний. Это нужно и в противоположном случае – если пользователь хочет специально изучать именно коллокации.

Для отделения коллокаций от свободных сочетаний нужны специальные фильтры.

Рассмотрим весьма частотную коллокацию *круглый стол*. Это сочетание нельзя автоматически разметить как коллокацию во всех случаях, поскольку оно может быть и свободным, ср. *на круглом столе лежали круглые яблоки* vs. *на круглом столе обсуждались проблемы корпусной лингвистики*. Но есть более простые случаи, когда коллокация однозначно отличается от свободного сочетания. Для этих случаев мы можем задать синтаксические фильтры, например: *участники/цели/задачи круглого стола*; *в круглом столе*; *круглый стол по*, и др. В таких контекстах *круглый стол* – коллокация.

Далее пишутся фильтры для каждого значения.

При работе фильтров важен порядок их следования. Удобнее сначала задать фильтры для более специфичных, лексически связанных значений (*круглый дурак/отличник* и *круглые сутки*).

Значение высокой (полной) степени (*круглый дурак*) встречается только в контексте лица. В словаре (МАС) *круглый дурак* (*глупец*, *невежда*) считается особым значением, а *круглый отличник* и *круглый сирота* – фразеологизмами (в нашей терминологии – оборотами). Однако на них пишется общий фильтр, поскольку все эти сочетания предсказуемы, воспроизводимы (и, в конечном счете, перечислимы), имеют общую семантику полной степени и общее семантическое ограничение (контекст лица).

Для значения также чрезвычайно важен порядок следования (порядок работы) фильтров. Значение *круглые сутки*, вообще говоря, выделяется фильтром «*круглый* + сущ.: время». Однако одним этим фильтром обойтись нельзя: кроме *круглых суток* на такой запрос будут также выдаваться *круглые часы* (в предметном значе-

нии – с круглым корпусом, поскольку у *часов* в словаре две пометы – «механизм» и «время») и *круглый месяц* (в смысле *круглая луна*). Временных значений сочетания *круглые часы* и *круглый месяц* как раз не имеют, так что омонимии здесь нет. Их нужно просто исключить из поиска.

Поэтому сначала запускаются фильтры для сочетаний со словами *часы* и *месяц*:

круглый + *час*&pl

круглый + *месяц*

В результате у существительных *часы* и *месяц* зачеркивается помета «время» и остается соответствующая «предметная» помета, а у *круглого* в таком контексте оставляется помета «форма» (*круглый*: SEM='r:qual t:physq:form').

А затем, когда *месяц* и *часы* в контексте *круглый* уже не имеют пометы «время», т.е. исключены из общего поиска, запускается общий фильтр с существительными класса «время», и мы получаем все контексты типа *круглые сутки, круглый год*:

круглый + S&t:time

В этом контексте у *круглого* зачеркиваются «качественные» пометы и остается помета «относительное» (*круглый*: SEM2='r:rel').

Оставшиеся случаи употребления *круглого* в контексте существительных, обозначающих физические объекты, их части, а также части тела (*круглое лицо*, *круглый шар*, *круглое колесо*, *круглая башня*, *круглая шляпа*, *круглая лепешка*), будут относиться к первому значению («форма»).

Фильтры для относительных прилагательных

У относительных прилагательных другая структура многозначности, отсюда другая стратегия создания фильтров.

Относительные прилагательные имеют общую помету «относительное» ('r:rel'). Кроме того, есть некоторое количество прилагательных, которые имеют фиксированный семантический класс (и, следовательно, специальную помету в словаре): это, прежде всего, «место» (*загородный, повсеместный, нагорный* и под.) и «время» (*вчерашний, ежедневный, довоенный* и под.).

В сфере относительных прилагательных нас интересуют два основных (наиболее массовых) процесса: (1) образование относительных значений по типовым, системным семантическим моделям и (2) образование качественных значений.

В случае (1) имеются в виду модели, по которым интерпретируются сочетания отсубстантивных относительных прилагательных с существительными. Они, как известно, интерпретируются с участием определяемого существительного, из которого извлекается некоторый семантический предикат: *обувная коробка* ('где хранится обувь': *коробка* → 'хранить'), *обувной магазин* ('где продают обувь': *магазин* → 'продавать'), *обувная фабрика* ('где производят обувь': *фабрика* → 'производить'). При этом есть какие-то характерные и частотные семантические модели, которые нередко включаются в словари в качестве самостоятельных значений и для которых можно написать фильтры.

Например, прилагательные от названий природных объектов (*лесной, морской* и под.) обычно обозначают место (*лесная хижина* = 'хижина находится в лесу'), хотя могут иметь и другую интерпретацию (*лесная промышленность*). Если будут найдены контексты (классы существительных), в которых *лесной, морской, речной*, *полевой* и т.д. интерпретируются по локативной формуле ('находящийся в лесу' / 'находящийся в (на) море' и т.д.), прилагательным в таких контекстах можно будет приписать помету «место». Таким образом, здесь можно использовать уже имеющуюся помету.

В других случаях можно ввести новые пометы. Например, артефакты обычно изготавливаются из каких-то исходных материалов путем их обработки, переработки. Поэтому можно ввести помету «материал изготовления». Ее будут получать относительные прилагательные, которые в контексте существительных со значением артефактов интерпретируются по модели: *X-овый Y* (*железная цепь*) = «изготовленный из X, с добавлением X, путем переработки X» (где X – производящее существительное). Например, для еды и напитков материалом изготовления могут быть растения, плоды, природные вещества (*пшеничная мука, гороховый суп, вишневый пирог, медовый пряник, молочный коктейль*), для оружия – металлы (*стальной клинок*), для украшений – металлы и камни (*золотые серьги, изумрудное колье, жемчужное ожерелье*) и т.д.

(2) Другой важный и по лингвистическим соображениям, и по охвату материала класс фильтров – фильтры для качественных значений относительных прилагательных.

Здесь главным критерием является не столько частотность (качественные значения могут иметь и относительно низкую частоту в Корпусе), сколько системность. Поэтому нас в первую очередь будут интересовать не

«индивидуальные» метафоры (*железная воля, золотое сердце*), которые с трудом поддаются обобщению, а более системные семантические переходы, одним из которых является образование цветового значения.

Качественные цветовые значения относительных прилагательных (*шоколадный*, *молочный, песочный* и под.) имеют сравнительно небольшую частоту, но большую регулярность, т.к. образуются от основ определенных семантических классов:

от названий растений и плодов растений: *вишневый, малиновый, лимонный, брусничный, абрикосовый, гороховый, пшеничный*;

от названий некоторых веществ и материалов: *шоколадный, молочный, песочный, снежный, асфальтовый, кирпичный*;

от названий драгоценных и полудрагоценных камней: *янтарный, изумрудный, бирюзовый*;

от названий металлов: *золотой, серебряный, медный, бронзовый, стальной*.

С другой стороны, классы существительных, в сочетании с которыми реализуются «производные» цветовые значения, тоже вполне определенные. «Производный» цвет бывает обычно у артефактов. Типичным контекстом для цветовых прилагательных являются существительные классов: «одежда и обувь», из «веществ и материалов» – ткани, краски и лаки, из «транспорта» – автомобили и вагоны поездов. А, например, здания и сооружения обычно имеют «натуральные» (первичные) цвета: *белые домики, зеленая беседка, серый забор*. И дело вовсе не в том, что у них не может быть «сложных» цветов. Просто их не принято называть «производными цветами». Если нормально сказать *гороховый сюртук* или *брусничный фрак*, то в случае «*гороховой стены*» говорящий скорее скажет *светло-коричневая* или *темно-желтая*.

В качестве иллюстрации рассмотрим прилагательное *вишневый*, у которого есть и значение «материал изготовления», и значение «цвет».

ВИШНЕВЫЙ

1) *вишневый 1* ‘материал изготовления’
вишневый + S r:concr t:food (еда и напитки)
вишневое варенье, в. сироп, в. сок, в. вино, в. ликер

2) *вишневый 2* ‘цвет’

а) *вишневый* + S r:concr t:tool:cloth (одежда и обувь)
вишневый фрак, в. пилотка, в. смокинг, в. полушалок, в. шарф, в. хитон, в. ботинки

б) *вишневый* + S r:concr t:tool:transp (транспортные средства)
вишневая «девятка», в. «пятерка», в. «иномарка», в. автомобиль

в) *вишневый* + S r:concr t:tool:furn (мебель)
вишневые стенды; в. полки

г) *вишневый* + S r:concr t:stuff (вещества и материалы)
бытовая химия, косметика: *вишневый лак, в. помада, в. раствор*
ткани: *вишневый шелк, в. бархат, в. атлас, в. штоф* (обивка), *в. плюш*
камень: *вишневый мрамор*.

Эти фильтры также запускаются в определенном порядке: сначала «*вишневый* + еда и напитки», затем «*вишневый* + вещества и материалы» (в противном случае в «вещества и материалы» попадут *вишневая мякоть, наливка, вишневый сок, морс, кисель*, которые, наряду с пометой «еда и напитки», имеют также помету «вещества»).

Список литературы

1. Апресян Ю.Д. Лексическая семантика: Синонимические средства языка. М.: 1974.
2. Кустова Г.И. О типах производных значений слов с экспериенциальной семантикой // Вопросы языкознания. 2002. № 2.
3. 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.

4. Рахилина Е.В., Кобрицов Б.П., Кустова Г.И., Ляшевская О.Н., Шеманаева О.Ю. Многозначность как прикладная проблема: лексико-семантическая разметка в Национальном корпусе русского языка // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог 2006». М.: 2006.
5. Atkins S. Tools for computer-aided corpus lexicography: the Hector project // Acta linguistica Hungarica 41, 1993.
6. Dolan W, Vanderwende L., Richardson S. Polysemy in a broad-coverage natural languigeprocessing system // Ravin Y., Leacock C. (eds.) Polysemy: Theoretical and Computational Approaches. N.-Y.: Oxford University Press, 2002.
7. Fellbaum C., Grabowski J., Landes S. Performance and confidence in a semantic annotation task // Fellbaum C. (ed.) WordNet: An Electronic Lexical Database. Cambridge (Mass.): The MIT Press, 1998.