# Принципы построения wordnet-тезауруса RussNet

Азарова И.В., Синопальникова А.А., Яворская М.В.
Кафедра математической лингвистики СпбГУ

Настоящий доклад посвящен уточнению методики построения RussNet, wordnet-тезауруса для русского языка, разрабатываемого сотрудниками кафедры математической лингвистики Санкт-Петербургского государственного университета.
В первой части доклада приводится перечень принципов, которые были использованы при построении wordnet-тезаурусов в широко известных проектах Принстонского WordNet, EuroWordNet и BalkaNet. Приводятся результаты апробации данных методик на материале русского языка в рамках проекта RussNet. Полученные данные имеют не только частное практическое значение при построении компьютерного тезауруса русского языка, но и общетеоретическое значение, поскольку показывают, как извлекать релевантную информацию из имеющиеся лингвистических ресурсов: толковых, идеографических и семантических словарей русского языка.
Во второй части доклада приводится схема использования лингвистических источников разного типа: результатов анализа корпуса текстов, дефиниций толковых словарей, данных ассоциативных словарей и частотных распределений лексем при построении тезауруса RussNet. Предлагаемая схема позволяет адекватно отобразить специфические для лексической системы русского языка связи лексикализованных понятий и минимизировать субъективность представления данных, т. е. верифицировать результаты проекта.
В докладе обсуждается методика присоединения RussNet к имеющимся системам национальных тезаурусов типа EuroWordNet и BalkaNet посредством связей с межъязыковым перечнем связующих концептов (ILI).

## Введение

Основной целью проекта RussNet является построение компьютерного словаря типа WordNet для лексики русского языка. В современной лингвистике термин wordnet (ставший нарицательным) употребляется применительно к особой разновидности лингвистических ресурсов (лексиконов, лексико-семантических баз данных, компьютерных тезаурусов), построенных по модели, которая была разработана в Принстонском Университете в 1985 г. Основными структурными единицами словарей типа WordNet являются синонимический ряд (иначе, синсет) и слово. Слова (точнее лексико-семантические варианты слов) и синсеты связаны между собой различными семантическими отношениями:

- на лексико-семантические вариантах слов задаются синонимические отношения,
- на синсетах – различные парадигматические и синтагматические отношения такие, как антонимия, гипонимия (т. е. родовидовые отношения), меронимия (отношения типа часть–целое) и различные виды лексического вывода – каузация, пресуппозиция, и др.

Несмотря на то, что существует определенная традиция и стандарты построения словарей типа WN, зачастую мы не можем заимствовать опыт наших коллег, и напрямую воспроизводить их методику. Это обусловлено, во-первых, спецификой русского языка (синтетического, флективного языка с развитой морфологической системой), во-вторых, отсутствием некоторых источников лексической информации (например, больших корпусов текстов), служащих отправной точкой исследования, в-третьих, недостатками и недочетами стандартных методик, ставшими очевидными при практическом использовании wordnet-тезаурусов. Так, например, сейчас все чаще говорят о том, что в Принстонском WN – самом первом wordnet – приводится гораздо больше значений слов, чем реально различается носителями языка.

## Типы источников и их применение в различных wordnet-проектах

Необходимость представлять в рамках wordnet разнообразные данные о лексических единицах в отдельности и лексической системе языка в целом требует привлечения различных источников лингвистической информации. Условно их можно разделить на две большие группы.

**Источники первого порядка** (корпусы текстов, результаты психолингвистических экспериментов) содержат эмпирические данные о реальном функционировании слов в языке (речи). Затрудняет использование этих источников то, что лексико-семантическая информация в них представлена в неявной форме, и ее извлечение требует применения специальных процедур и средств.

**Источники второго порядка** (разного рода лексикографические источники: словари, тезаурусы и т.п.) строятся на основе источников первого порядка с привлечением интуитивных знаний, интроспекции экспертов-лингвистов. Они предоставляют информацию в эксплицитной форме, что упрощает процесс ее извлечения. Однако зачастую составители словарей придерживаются иных установок, действуют в рамках различных концепций, что заставляет прибегать к дополнительной проверке или уточнению данных[1].

Несмотря на существование общих принципов построения wordnet-тезаурусов, в зависимости от установки разработчиков и того, входит ли тезаурус в объединенную систему типа EuroWordNet или BalkaNet, набор источников и методики их использования существенно варьируются в рамках различных проектов. Так например, в рамках Принстонского WordNet на первом этапе работы использовались источники первого порядка, а именно результаты психолингвистических экспериментов, в дальнейшем методика была расширена за счет применения контекстного и дефиниционного анализа. С возникновением проекта EuroWordNet особую роль стали играть корпусы текстов как наиболее достоверные источники лингвистической информации, и методы их анализа, кроме того были разработаны

---

[1] Например, в ходе работы с лексикографическими источниками было обнаружено, что в ТСРГ (Бабенко, 1999) отношения синонимии и антонимии не являются симметричными. Для глагола *беситься* в качестве синонимов указываются *злиться, неистовствовать*, при этом *беситься* не включен в синонимические ряды для каждого из них: *злиться* – син. *сердиться;* а *неистовствовать* – син. *бесноваться, буйствовать, бушевать*. В ходе анализа материалов РСС (Шведова, 1998) нами была выявлена неравномерность в членении семантических классов, одни из них оказались описаны излишне подробно, тогда как для других степень детализации была явно недостаточной. В эксперименте это было продемонстрировано тем, что на долю одного класса пришлось около 80% существительных, извлеченных из текстов, тогда как для ряда других не было найдено ни одного представителя.

процедуры автоматической обработки электронных толковых словарей, тезаурусов и онтологий.
Таким образом, на сегодняшний день к стандартным методам построении национальных wordnet-тезаурусов относятся дефиниционный, контекстный и словообразовательный методы анализа значений. В рамках Принстонского WordNet также учитываются психолингвистические данные, полученные при проведении собственных экспериментов или представленные в списках ассоциативных норм английского языка. Каждый из перечисленных методов имеет определенные рамки и ограничения. В частности, в процессе дефиниционного анализа предполагается обращение к традиционным толковым словарям, которые были разработаны для иных целей и в рамках совершенно иной парадигмы. Об этом свидетельствует и структура словарной статьи, в которой выделение значений и оттенков значений, порядок упорядочения значений в словарной статье, выделение первого или основного значения слова носит в большой степени субъективный характер и меняется от словаря к словарю. То, что статья толкового словаря является вариантом разметки структуры значений слова и объективно отражает в некоторой степени функционирование слова в языке, безусловно. Однако, какова степень огрубления или, наоборот, чрезмерного подразделения значений, их иерархизации, остается непонятным, часто обусловлено составом картотеки примеров и установками группы лексикографов.
Среди базовых принципов построения wordnet-тезауруса в рамках Принстонского WordNet был сформулирован принцип, по которому перечисление значений слова должно соответствовать частотному распределению значений в текстах, то есть первым должно являться значение, наиболее часто встречающееся в текстах. Этот принцип не всегда выполнялся в wordnet-проектах в силу сложности его реализации. Во-первых, следует сформулировать временную и тематическую перспективу текстов. Во-вторых, разметка значений в корпусах трудоемка, нет готовых корпусов с размеченными значениями, следовательно, проводить такую разметку практически невозможно. В-третьих, словарные картотеки зачастую задают неравноценное представительство для разных значений: количество частых значений уменьшено в силу их обычности, "тривиальности", в то время как редкие значения, возможно окказионально встретившиеся, включаются наряду с узуальными, утвердившимися в языке. Все эти факторы, на наш взгляд, значимы, их необходимо учесть при построении компьютерного тезауруса, однако следует принять четкие решения по каждому пункту, даже если это идет в разрез с традиционными лексикографическими принципами.

## Типы источников и методы их применение в проекте RussNet

RussNet является тезаурусом типа WordNet для русского языка, который разрабатывается сотрудниками кафедры математической лингвистики Санкт-Петербургского государственного университета. Основные принципы построения компьютерного словаря представлены на сайте филологического факультета СПбГУ[2] и в ряде статей (Азарова и др. 2003; Материалы к компьютерному тезаурусу… 2002; *Azarova* et al. 2002).
В рамках проекта RussNet было принято решение о построении ядра компьютерного словаря на **базе корпуса современных текстов**. Этот период, на наш взгляд, начинается с середины 80-х годов (конца "советской эпохи") до настоящего времени. И хотя, наверно, и этот период имеет внутреннюю неоднородность, но ею можно пренебречь. В отношении тематического распределения текстов была выбрана достаточно стандартная схема преобладания газетных текстов (40%) как жанра, наиболее быстро откликающегося на изменения в языке, достаточно

[2] http://www.phil.pu.ru/depts/12/RN/

экспрессивного и вариативного; большой доли (30%) научно-популярных текстов как экспрессивно нейтральных и описывающих реалии не только обыденной жизни, но и других сфер; небольшая часть (20%) отрывков из художественной литературы, причем важным является отсутствие произведений, взятых целиком, а также больших фрагментов текстов (свыше 5 тысяч словоупотреблений), которые могли бы создавать идиолектные «флуктуации» употребления значений слов в корпусе; небольшая часть (10%) текстов законов, договоров, инструкций и проч., обеспечивающая конструкциями современных клише делового употребления слов.

Имеющийся корпус текстов, состоящий из 21 миллиона словоупотреблений, используется для отбора единиц, которые соответствуют ядру общеупотребительной лексики русского языка. Предполагается, что эти слова задают верхние уровни гипонимической иерархии и вершины деревьев в RussNet. Первоначально были отобраны слова с частотой более 120 вхождений на 1 млн. словоупотреблений (Vossen, 1998). В их число входят около 500 существительных, 200 глаголов, 200 прилагательных, и 100 наречий. Полученную совокупность была дополнена словами, соответствующими так называемому «**ядру языкового сознания русских**» (Уфимцева, 2002), т. е. словами, появляющимися в ответах испытуемых при ассоциативном эксперименте наиболее часто, и следовательно, связанными с наибольшим количеством других слов (более 100 обратных ассоциаций), например, *человек, дом, жизнь, вода, день, лес, работа, книга, стол, город, друг, любовь, радость; есть, идти, думать, жить, большой, красивый, хороший; плохо, быстро, много* и др.

Разбивая на классы полученную совокупность слов, мы получаем представление о количестве родовидовых деревьев в RussNet, однако выполнение этой задачи осложняется тем, что наиболее частотные слова русского языка являются и наиболее многозначными. Поэтому далее необходимо выделить наиболее употребительные значения этих слов. Для этой цели нами используется корпус, из которого при помощи программы Бонито[3], разработанной сотрудниками Университета им. Масарика, извлекаются контексты употребления рассматриваемых лексем. Менеджер текстов Бонито позволяет осуществлять поиск контекстов для отдельной словоформы, ряда словоформ или лексемы целиком, сортировать контексты употребления относительно левой или правой частей контекста, создавать частотные словари и извлекать статистические характеристики совокупности контекстов.

Набор извлеченных контекстов для каждой лексемы размечается относительно схемы значений, представленных в толковом словаре (например, МАС). Нами были проведены исследования разметки полного набора контекстов и его подмножеств с тем, чтобы выяснить, насколько четко сохраняется схема распределения частотности значений лексемы. Опытным путем было установлено, что выборочная разметка случайным образом взятых 100-150 контекстов из разных произведений дает ту же схему распределения контекстов, что и полная совокупность, включающая 1500-2000 контекстов. Доля контекстов, являющихся реализацией наиболее частотного значения, колебалась не более, чем в интервале ±1%, при этом соотношение долей контекстов с наиболее частым значением и следующим за ним по частотности значением регулярно различались на 50%. Таким образом, при иерархизации значений по частотности достаточно разметки части контекстов, выбранных случайным образом.

Анализ контекстов позволяет также выявить набор значений, которые следует представлять в компьютерном тезаурусе. В частности, единичные случаи реализации значений считаются окказиональными. Для разделения значений на окказиональные и узуальные вводится

---

[3] http://nlp.muni.cz./projects/bonito

пороговое значение (1%) от общего числа контекстов, которое должна составлять доля контекстов, реализующих значение в корпусе, для включения его в структуру лексикализованных понятий компьютерного тезауруса. Для разграничения значений также используется параметр частотного представительства в совокупности контекстов, помимо которого используется еще рамка валентностей. При этом считается, что отдельное значение должно иметь отдельную схему валентностей или сочетаемости с контекстом.

**Сочетаемость** предикативных и признаковых слов определяется набором обязательных и факультативных активных валентностей, причем обязательной считается валентность, реализующаяся с частотой более 70-85% в контекстах рассматриваемого слова в корпусе современных текстов, а факультативной – та, которая реализуется с частотой более 15-30%. Окказиональные валентности представлены, как правило, менее, чем в 15% контекстов рассматриваемого слова. Выделение валентностей осуществляется на основе функционально-синтаксических позиций при слове, которые фиксируются тремя параметрами: (1) функцией, определяемой вопросом, на который отвечает заполняющая форма; (2) формой поверхностного выражения валентности; (3) семантическим типом слова, занимающего валентную позицию. Например, для глагола *направится* в нашем корпусе из 21 млн. словоупотреблений было найдено 358 контекстов употреблений в значении «двинуться в каком-л направлении», контексты составили практически 100% общего числа контекстов употребления данного слова, поскольку в другом значении это слово было употреблено лишь один раз. Употребление в этом значении предполагает 2 обязательные валентности: (1) упоминание ***лица*** (группы лиц), которое совершает движение, причем, как правило, конкретный способ передвижения указан в непосредственной близости от данного (часто в составе того же самого предложения); (2) ***направления движения***, которое представлено конструкцией "к + $N_3$" (44%) (*к дивану, к другу, к спуску, к нему…*), называющей чаще (36%) место локализации, а реже (8%) – лицо (лиц), по направлению к которым ориентировано движение; в небольшом числе случаев происходит сочетание этих частотных поверхностных структур (локализация + лицо); вторая частотная конструкция "в + $N_4$" (27%) указывает на направление пространственной локализации движения (*в комнату, в деревню, в угол гостиной…*); окказионально встречаются конструкции "в сторону + $N_2$", "на + $N_4$", "по + $N_3$". Словарная дефиниция МАС «двинуться куда-л, в какую-л сторону, в каком-л направлении» перечисляет и частотные, и низкочастотные типы реализации валентности направления. Помимо лица, позицию первой валентности может занимать название транспортного средства и даже неодушевленного объекта, однако, такие примеры составляют 1% от общего числа контекстов. Окказиональные валентности (менее 10%) представлены также способом действия (*решительно, прямо, напрямик* и т.п.), указанием целевого действия (*курить, изучать* и т.п.), местом действия (*по берегу, через парк, по суше* и т.п.), начальной точкой движения (*из Вифании*). Набор обязательных и факультативных валентностей составляет описание валентностной структуры значения слова, которая может непосредственно использоваться в синтаксических правилах формальной грамматики.

Еще один важный вопрос состоит в том, насколько валетностная схема признакового слова совпадает с собственно языковой структурой, например, со структурой статьи словаря ассоциативных реакций РАС. **Ассоциативные словарь** также предоставляет данные о сочетаемости слов, но в гораздо менее развернутой форме. В РАС информация о потенциальных или реальных контекстах слова, ограничена рамками словосочетания, многословные ассоциации достаточно редки, они составляют менее 2% от общего числа ответов. Кроме того, поскольку в ассоциативном эксперименте симулируется ситуация речевого общения, левый контекст слов (*большой* ***дом****, взять* ***за руку***) воспроизводится чаще, чем правый (***начать*** *работать,* ***хочу*** *петь*). Несмотря на эти различия и ограничения, результаты контекстного анализа текстов и материалов РАС во многом согласуются друг с

другом, позволяют выделить основные и периферийные значения слова. Например, на четыре выделенных в RussNet значения глагола *чувствовать* приходится около 84% ассоциаций в РАС и более 94% вхождений в текстах корпуса. Однако, ассоциативный словарь не дает преимуществ на уровне выявления особенностей сочетаемости слов, но облегчает установления семантических отношений между словами.

Количество различаемых в компьютерном тезаурусе значений слова определяется набором схем сочетаемости. В отдельных случаях сложно определить, насколько детально следует их описывать. Рассмотрим эту проблему на примере схем сочетаемости прилагательного *большой*. Три первых значения для этого слова в МАС сформулированы следующим образом: Значительный по величине, размерам; *противоп.* Малый, маленький ‖ Значительный по количеству, многочисленный ‖ Появляющийся, находимый или производимый в большом количестве ‖ Продолжительный по времени, охватывающий значительный промежуток времени.

Значительный по силе, интенсивности, глубине ‖ Важный по значению.

*при существительных, характеризующих качество человека, имеет усилительный смысл:* В высокой степени, чрезвычайный ‖ Замечательный в каком-то отношении, выдающийся.

Просмотр контекстов употребления в корпусе прилагательного *большой* позволяет определить частотность реализации значений: основное значение «значительный по величине, размерам» является самым частотным (38%). Среди существительных, сочетающихся с прилагательным в этом значения подавляющая часть (19%) обозначает артефакты (то, что создано человеком), среди которых есть и бытовые предметы (*матрац, пульт, печка*), и контейнеры различного рода (*кувшин, коробка, резервуар*), и помещения (*дом, зал, ресторан*), и ряд объектов, имеющих не столько пространственные измерения, сколько плоскостные (*карта, снимок, атлас*), причем выделение артефактов, как наиболее частотных определяемых объектов, не носит характер противопоставления естественным или природным объектам (*камень, залив, океан*), которых все-таки меньше (13%) и среди которых на удивление мало (3%) названий животных и растений (*птица, кошка, ромашка*). Две другие группы существительных, обладающих достаточно четким значением, являются небольшими по объему (по 4%). Одна обозначает части тела человека или животного (*голова, рука, лапа*), а также другие части: текста (*параграф*), вещества (*капля*). Вторая – собственно измерения (*размеры, расстояние, рост, высота*).

Оттенок первого значения «значительный по количеству» (24%) реализуется у прилагательного в сочетаниях с существительными обозначающими **совокупности** людей (*семья, совет, оркестр*), предметов (*коллекция, ряд*), финансов (*деньги, выигрыш, потери*), причем синоним в определении «многочисленный» сочетается не со всеми существительными. По частотности употребления в контекстах и особенностям сочетаемости с существительными это значение прилагательного является самостоятельным.

Следующим по частотности (20%) значением является «значительный по силе, интенсивности, глубине». Основная группа существительных, сочетающихся с указанным прилагательным в этом значении, как правило, представлена транспозитами от глаголов или прилагательных. Причем прилагательное *большой* является своеобразным трансформом наречий-адъюнктов признаковых слов, например, ***сильно*** *давить* => ***большое*** *давление*, ***сильно*** *разочароваться* => ***большое*** *разочарование*, ***очень*** *редкий* => ***большая*** *редкость*, которые в основном выражают значение интенсификатора признака или действия. Помимо этого значения, встречаются трансформы количественных значений, повторяемости действий: *меняться* ***часто*** => ***большая*** *изменчивость,* ***много*** *потратить* => ***большие*** *траты*. Это значение прилагательного совпадает с формулировкой 3-го значения, за исключением того, что в последнем случае определяются качественные характеристики людей: ***очень*** *демократичный человек* => ***большой*** *демократ.* В таком случае, возможно ли

объединение этих значений? Учитывая, что для отглагольных существительных возможно сочетание с антонимичными прилагательными (ср. *маленькие радости, маленькое давление*), а для обозначений качеств такое сочетание невозможно (**маленькая редкость, *маленький демократ, *маленький мастер*), очевидно, что значения должны быть сформулированы как два лексикализованных понятия с четким указанием семантических типов существительных, сочетающихся с прилагательными в данных значениях.

При корректировке методики построения синсетов была установлена следующая закономерность: элементы синсета (синонимы) обладают однотипной сочетаемостью (совпадение обязательных и/или факультативных валетностей) в корпусе современных текстов, при этом форма поверхностного выражения валентностей у синонимов может различаться, окказиональные валентности также могут быть различны.

Для подключения русского wordnet-тезауруса к структуре межъязыкового индекса (ILI) используются специальные отношения, которые были предложены в рамках EuroWordNet: EQ-синонимия, EQ-гипонимия и проч. Собственно элементы ILI представляют собой набор понятий, распределенных по областям, но не упорядоченных полностью в структуры деревьев. Если устанавливается отношение тождества между синсетом RussNet и элементом ILI, то синсет присоединяется к элементу одиночной связью EQ-синоним. Например, *{продукт1}* EQ-синоним *{artefact, artifact}*. В противном случае, синсет присоединяется как минимум двумя связями, например, EQ-гипоним и EQ-мероним. Например, для русского синсета *{запеканка1}* не существует прямого эквивалента в ILI, поэтому он связывается отношением EQ-гипоним с элементом *{baked goods}* и EQ-мероним-субстанция *{dairy product}*.

## Литература


1. Vossen, 1998 – Vossen, P. (ed.): EuroWordNet: A Multilingual Database with Lexical Semantic Network. Dodrecht, Kluwer.
2. Уфимцева, 2002 – Уфимцева Н.В. Ядро языкового сознания русских (по данным массовых ассоциативных экспериментов) // Доклады научной конференции «Корпусная лингвистика и лингвистические базы данных». СПб, 2002. С. 157–164.
3. Азарова и др. 2002 – *Азарова И.В., Митрофанова О.А., Синопальникова А.А.* Компьютерный тезаурус русского языка типа WordNet // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции Диалог 2003 (Протвино, 11-16 июня 2003 г.) М., 2003. С. 43–50.
4. Материалы к компьютерному тезаурусу… 2002 – Материалы к компьютерному тезаурусу лексики русского языка / Сост. И.В. Азарова, О. А. Митрофанова. СПб., 2002. 232 с.;
5. *Azarova et al. 2002 – Azarova I., Mitrofanova O., Sinopalnikova A., Yavorskaya M., Oparin I.* RussNet: Building a Lexical Database for the Russian Language // Workshop Proceedings: Workshop on WordNet Structures and Standardisation, and how these affect Wordnet Application and Evaluation. 28th May 2002. Las Palmas de Gran Canaria, 2002. P. 60–64.


## Источники


6. МАС – Словарь русского языка / Под. ред. А.П. Евгеньевой. Т. 1–4. М., 1985–88.
7. РАС – Караулов Ю. Н. и др. Русский ассоциативный словарь. Т. 1–6. М., 1994, 1996, 1998.
8. РСС – Русский семантический словарь: Толковый словарь, систематизированный по классам слов и значений / Российская академия наук. Ин-т рус. яз. им. В. В. Виноградова; Под общей ред. Н. Ю. Шведовой. – М., 1998.
9. ТСРГ – Толковый словарь русских глаголов: Идеографическое описание. Английские эквиваленты. Синонимы. Антонимы / Под ред. Л. Г. Бабенко. М., 1999.