ИСПОЛЬЗОВАНИЕ СТАТИСТИКО-КОМБИНАТОРНЫХ СВОЙСТВ КОРПУСА СОВРЕМЕННЫХ ТЕКСТОВ ДЛЯ ФОРМИРОВАНИЯ СТРУКТУРЫ КОМПЬЮТЕРНОГО ТЕЗАУРУСА RUSSNET

Азарова И.В., Синопальникова А.А.

Введение

Основной целью компьютерного тезауруса RussNet[1] является представление ядра лексической системы современного русского языка. Основные принципы построения такого типа словарей были сформулированы в рамках проектов WordNet[2] и EuroWordNet[3] и развиваются в современном проекте BalkaNet[4].

Особенностями компьютерного словаря RussNet являются следующие: словарь представляет собой модель организации лексического пространства некоторого языка, отдельной точкой лексического пространства является лексикализованное понятие, которое задается синонимическим рядом, так называемым «синсетом»; между синсетами установлены семантические отношения, среди которых наиболее важными являются родовидовые, а также отношения антонимии, меронимии и проч.

Стандартная методика построения wordnet-словарей предполагает использование двух типов данных: источников первого порядка (корпусов текстов, результатов психолингвистических экспериментов), которые содержат эмпирические данные о функционировании слов в текстах (речи), и источников второго порядка – готовых словарей: толковых, синонимических и проч.

В рамках проекта RussNet была сформулирована методика[5] построения тезауруса, которая опирается на корпус современных текстов**.** Основные положения этой методики следующие. (1) Данные источников второго порядка необходимо верифицировать

---

[1] 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.

[2] Fellbaum, 1998 — WordNet: an electronic lexical database / Fellbaum Ch. (ed.). Massachusetts, 1998.

[3] EuroWordNet: A Multilingual Database with Lexical Semantic Network / Vossen, P. (ed.) Dodrecht, Kluwer, 1998.

[4] Romanian Journal of Information Science and Technology. Vol. 7. N. 1–2. Special Issue on the BalkaNet Project. Bucharest, 2004.

[5] Общие положения этой методкии и обсуждение отдельных пунктов были представлены в работах: Azarova I., Sinopalnikova A. Adjectives in Russnet // International WordNet Conference, GWC 2004 Brno, Czech Republic, January 20-23, 2004. P. 251-259; Азарова И.В., Синопальникова А.А., Яворская М.В. Принципы построения wordnet-тезауруса RussNet // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции Диалог'2004 («Верхневолжский», 2-7 июня 2004 г.) М., 2004. С. 542-547.

относительно имеющегося корпуса текстов. (2) Дифференциация лексико-семантических вариантов (значений) отдельной лексемы производится на основании регулярных различий в структуре контекстов, которые описываются через понятие валентности. (3) Порядок перечисления лексико-семантических вариантов лексемы и порядок представления доминанты синонимического ряда определяются частотными характеристиками единиц в имеющемся корпусе.

Особенности нашего понимания ***корпуса современных текстов***, ориентированного на наполнение компьютерного тезауруса типа wordnet, следующие. Тексты в корпусе относятся к периоду «постсоветской эпохи» (с середины 80-х годов XX в. до настоящего времени). В жанровом отношении в корпусе преобладают тексты, имеющие усредненные параметры информативности, экспрессивности и стереотипности, поэтому в корпусе 40% газетных текстов; 30% научно-популярных текстов, описывающих реалии не только обыденной жизни, но и других сфер; 20% отрывков из произведений художественной литературы, причем важным является ограничение на объем фрагментов (не более 5 тысяч словоупотреблений), поскольку в противном случае большие объемы авторского текста могут создавать идиолектные «флуктуации» употребления значений слов в корпусе; 10% текстов законов, договоров, инструкций и проч., обеспечивающих набором современных клише употребления слов.

Структура значений толкового словаря используется как первоначальная схема разграничения смыслов в представительной выборке контекстов слова из корпуса, которая размечается вручную, при этом описываются параметры контекстов. В тезаурус заносятся лексико-семантические варианты, узуально представленные в корпусе, т.е. встретившиеся более, чем в 1% контекстов нашего корпуса, состоящего из 21 млн. словоупотреблений. Выявленные лексико-семантические варианты слов упорядочиваются в соответствии с ***частотой употребления*** слов в размеченных контекстах. Была сформулирована гипотеза, которая проходит в настоящее время проверку, что частота употребления первого (основного) значения лексемы значительно превышает частоту употребления следующих вариантов. Случаи примерно равного числа контекстов, представляющих разные лексико-семантические варианты, вызваны либо относительно низкой частотой данного значения, либо перед нами конкурирующие (омонимичные) употребления слов, находящиеся на одном уровне лексической иерархии, т.е. являющиеся согипонимами или входящие в разные структуры родовидовой иерархии тезауруса.

При формировании «синсетов» (синонимических рядов) частотность употребления лексических единиц используется для задания своеобразного «кортежа»: выделяется

«доминанта» – наиболее часто используемое нейтральное слово для выражения лексикализованного понятия – и второстепенные элементы синсета, которые существенно уступают доминанте по частоте использования в корпусе, но могут устойчиво употребляться для номинации понятия в определенных функциональных сферах, применительно к определенной области понятий. В данном пункте методики проверяется гипотеза, что ***частотное распределение элементов синсета*** дает столь же четкую схему распределения ЛСВ слов в кортеже, как и в при частотном распределении лексико-семантических вариантов отдельного слова.

Статистико-комбинаторные характеристики контекстов используются для выявления типичных для данного ЛСВ слова схем сочетаемости, они заносятся в RussNet в виде перечней валентностей, которые задаются в формально-грамматическом, смысловом и синтаксическом планах. Характеристика обязательной, факультативной и окказиональной валентностей задается на базе частоты реализации связи в наборе контекстов корпуса (более 60%, в интервале 30-60%, менее 30% вхождений). В данной работе мы рассмотрим, как статистико-комбинаторные характеристики контекстов используются при выявлении парадигматических семантических отношений между синсетами, и их верификации.

Валентности и отношения глаголов *злиться* и *сердиться*

Для прояснения предложенных гипотез и пунктов нашей методики рассмотрим анализ значений глаголов *злиться* и *сердиться*. Значение глагола *злиться* в МАС[6] определено как "испытывать злость на кого-то, что-то; сердиться на кого-то, что-то". Это определение задает хотя бы в части значения отношение эквивалентности с глаголом *сердиться*. И отсылает нас к определению слова *злость, злоба* "раздраженно-враждебное чувство". В свою очередь *сердиться* определяется "испытывать раздражение, гнев на кого-, что-л.", как если бы эти два состояния были весьма близки, что не вполне соответствует интуиции носителя языка (под гневом понимается "состояние **сильного** негодования, раздражения"). В ТСРГ[7] для двух глаголов даются взаимные отсылки синонимичности. В ОСС[8] среди ряда факторов, разделяющих эти синонимы, можно выделить два: то, что ***злиться*** могут не только люди, но и животные (в отличие от ***сердиться***), кроме того, статус субъекта действия ***сердиться*** более высок или равен статусу объекта, вызвавшего эмоции. В случае действия, обозначенного словом ***злиться***,

---

[6] Словарь русского языка / В 4-х томахю М., 1981.

[7] Толковый словарь русских глаголов: Идеографическое описание / Под ред. Л. Г. Бабенко. М., 1999.

[8] Новый объяснительный словарь синонимов русского языка. 1 вып. / Под рук. Ю. Д. Апресяна. М., 1999.

статус не определен. Это различие выглядит как вполне правдоподобное с интуитивной точки зрения.

Корпусный анализ контекстов употребления данных глаголов с использованием методики валентностей, предложенной нами в предыдущих работах, позволяет прийти к следующим результатам. Во-первых, различие в частоте употребления ***злиться/сердиться*** (142 и 180) этих глаголов в нашем корпусе не согласуется с гипотезой соотношения частот при синонимии (доминанта синсета обычно имеет большее преобладание). На материале НКРЯ[9] соотношение частот 271 и 335, что дает примерно такое же соотношение.

Проверяя факт употребления этих глаголов с субъектом-животным, обнаруживаем, что и тот, и другой употребляются таким образом в незначительном числе случаев (*злиться* – 7 раз, *сердиться* – 4 раза, в процентном отношении 4,8% и 2,2%), что по нашей методике входит в диапазон случайных вариаций.

Более подробно схема валентностей глагола *злиться* выглядит следующим образом. Субъектом состояния, обозначенного глаголом *злиться*, узуально является человек (96%), в частном случае (3%) ребенок, другие возможные субъекты: 2% – животное, 2% – явление природы (эти вхождения можно расценивать как метафорический перенос).

Причина появления «раздраженно-враждебного чувства» в контексте глагола *злиться* указана окказионально (34%), причем в подавляющем числе случаев (23%) указан объект негативных эмоций (иногда сам субъект состояния), который одновременно является причиной, в остальных случаях указана собственно причина: 6% – явления природы, 2% – описание ситуации в придаточном предложении, 3% – названия предметов и животных. В 66% случаев причина не конкретизируется, при этом *злиться* приводится в конструкции перечисления с названиями других состояний или действий:

> *он не договаривает, всегда нервничает и <злится>; …не мешало Борису <злиться> и обижаться теперь на очередной несправедливый разнос; <злясь> и веселясь…; <злиться> и комплексовать…; <злясь> больше всего на себя самого и одновременно удивляясь…; вы на литконсультантов <злитесь> и себя гением мните…; почему <злится> и отчего страдает в действительности современный японец; молчит и <злится>…; <злитесь>, приходите в ярость…; начинаю <злиться>, раздражаться…; она бывало покрикивала на него, <злилась>, краснела…*

В одном контексте рассматриваемые глаголы упомянуты вместе: *Ну что, не сердишься, не <злишься> больше?* Квалификация состояния дается также окказионально: всего 4% случаев (*всерьез, сильно* и *безутешно*).

Глагол *сердиться* столь же узуально имеет субъектом человека (98%), среди этих вхождений встречаются и дети (2%) и упоминания родителей, редко – влиятельных лиц

(*Сталин, Ельцин*), но в общем какая-либо социальная стратификация явно не обозначена (хотя есть единичные примеры, где эта коннотация четко выражена в контексте: *Не сердись, парень. – Рабам не положено сердиться*). Кроме человека, в 2% контекстов представлены обозначения животных (эти примеры можно расценивать как метафорический перенос, однако они вполне сходны с употреблениями глагола *злиться*).

Объект-человек, на который направлено эмоциональное состояние *сердиться*, указан в 24% случаев, причина состояния указывается гораздо чаще (16% вхождений). Кроме того, большая часть контекстов, в которых почти никогда не указана причина, представляет собой конструкцию с отрицанием (32%[10]): либо собственно отрицание состояния[11] (*не сердился*), либо модальной оценки состояния (*не надо сердиться, приходится сердиться, не любил сердиться*), и очень часто мягкой просьбы (*не сердись, не сердитесь*). Таким образом, относительно ясное противопоставление глаголов *сердиться* и *злиться* с опорой на данные контекстного анализа проходит, в первую очередь, по параметру обоснованности перехода к негативному отношению к кому-, чему-л. или необоснованности, спонтанности перехода к такому состоянию.

Злиться		***Сердиться***
95%	Субъект – человек	98%
23%	Объект – человек	24%
11%	Причина – ситуация	18%
6%	Конструкция с отрицанием	32%
60%	Необоснованная, спонтанная реакция	24%

Таблица 1. Схема реализации валентностей *злиться* и *сердиться*.

Квалификаторы состояния, обозначенного глаголом *сердиться*: *всегда, вечно, еще, беззлобно, только, тоже*, встречаются столь же нерегулярно, как и квалификаторы *злиться*, но имеют коннотацию понижения интенсивности негативного чувства (что отчасти присутствовало в определении МАС) и соотносятся с интервалом времени. Отчасти эта характеристика связывает глаголы между собой, поскольку они оба сочетаются с фазовыми глаголами (*начал злиться/сердиться*), но для *сердиться* чаще используется лексическая форма инхоатива *рассердиться*[12], в отличие от *разозлиться*, что, вероятно, указывает на то, что состояние *злиться* имеет большую амплитуду, и

[9] Национальный корпус русского языка / URL: www.ruscorpora.ru.

[10] По данным НКРЯ эта доля еще больше – 46%.

[11] В ОСС указано, что в сочетании с *не* оба глагола *злиться* и *сердиться* употребляются со значением мягкой просьбы, но это не согласуется с данными корпусного анализа. Для *злиться* такое употребление было в несколько раз меньше – 6% контекстов.

[12] Соотношение частот глаголов *рассердиться-сердиться* 125/180, *разозлиться-злиться* 72/142.

начать злиться менее интенсивно, чем *разозлиться*. Интенсивность чувства, выражаемого глаголом ***злиться***, подчеркивается еще одним параллельным источником – РАС[13], который мы используем наряду с контекстной информацией корпуса. Реакции глагола *злиться* (*сердиться* нет в числе стимулов) *сильно, очень, ужасно* указывают на интенсивность чувства, *напрасно, зря* – на беспричинность состояния; самая частотная реакция *на себя* – менее заметное расхождение: выражение недовольства по отношению к себе передается глаголом *злиться* естественнее, чем *сердиться*.

Периферийные глаголы

К глаголам *сердиться* и *злиться* примыкают 2 группы менее частотных глаголов. Первая: *негодовать* (56), *беситься* (41), *раздражаться* (49), причем в корпусе для последнего представлено 2 значения: частотное значение негативного эмоционального состояния (94% вхождений), низкочастотное – чувствовать болезненное ощущение на коже (6%). Вторая группа: *гневаться* (15), к которой примыкает *разъяриться* (9), указывающий на переход в эмоциональное состояние.

Проверяя эту группу на соотношение частот обозначений собственно эмоциональных состояний и их инхоативов, замечаем, что по типу *сердиться/рассердиться* соотносится пара глаголов *беситься/взбеситься* (41/42), остальные пары характеризуются понижением частоты инхоатива: *раздражаться/раздражиться* (46/5), *негодовать/вознегодовать* (59/11), *гневаться/разгневаться* (15/7).

Тест на сочетаемость глаголов с *не* весьма характерен: по типу *сердиться* ведет себя глагол *гневаться* (27% контекстов с *не*), по типу *злиться* – глаголы *беситься, негодовать*; глагол *раздражаться* занимает промежуточное положение (14% контекстов).

Глагол *беситься* (особенно его инхоатив *взбеситься*) дает наибольшее разнообразие типов субъекта состояния, среди которых не только отдельные индивиды, но и группы индивидов (*толпа, низы, актив*), животные, предметы, явления (*механизмы, прогресс, вода*). Этот же глагол обладает минимальным количеством контекстов (10%), в которых есть указание причины негативного эмоционального состояния. То есть глагол *беситься* является гипонимом-интенсивом для глагола *злиться*.

Глагол *гневаться* примыкает к глаголу *сердиться*, выражая сходное обоснованное, но более интенсивное негативное чувство, то есть является его интенсивом, что также

[13] Караулов Ю.Н., Черкасова Г.А., Уфимцева Н.В., Сорокин Ю.А., Тарасов Е.Ф. Русский ассоциативный словарь. Т. I. М., 2002.

подтверждается соотношением по частотности с инхоативом. Одним из подтверждений синонимичности/квазисинонимичности глаголов могут быть замещения в контекстах (хотя они встречаются довольно редко): *Когда они гневаются, то я спокоен. Иными словами, когда я <сержусь>, они успокаивают меня, а когда они, я успокаиваю их.*

Глагол *раздражаться* частично похож на глагол *злиться* свой необоснованностью перехода в негативное эмоциональное состояние, что иногда в контекстах подчеркивается лексически (*без причины, по любому поводу, по пустякам*). Кроме того, регулярно входит в конструкции перечисления с названиями других состояний и действий, как было характерно для глагола *злиться*:

<раздражается>, нервничает…; и никто не удивляется, не <раздражается>;
…мучительной необходимости любить, глядеть в глаза и <раздражаться>.

Это позволяет нам предположить, что глагол *злиться* является доминантой синонимического ряда, а *раздражаться* входит в него в качестве периферийного элемента.

Глагол *негодовать* является для этой группы периферийным в силу того, что чаще предполагает словесную реакцию (около 30%), которая у базовых глаголов представлена в окказиональном числе контекстов (3-5%).

Заключение

И наконец, каковы парадигматические отношения глаголов *сердиться/злиться*? Их противопоставление по двум аспектам значений (обоснованность/необоснованность перехода в состояние; амплитуда значений интенсивности негативных эмоций), разные схемы валентностей и частотность одного уровня в корпусе, вероятнее всего, показывают, что они являются согипонимами, занимая центральное положение в ядре этой группы глаголов. Пересечение значений двух глаголов конечно же довольно значительно, в традиции EuroWordNet синсеты могли бы соединяться семантическим отношением NEAR_SYNONYMY (квазисинонимии). В части контекстов они выступают как синонимы, но различаются рамками валентностей.

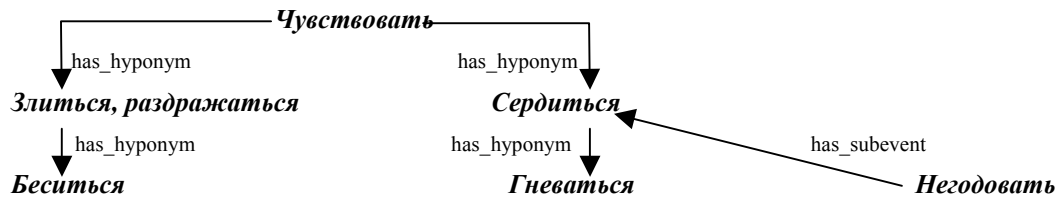

Рис. 1. Схема семантических отношений глаголов *злиться* и *сердиться*

Таким образом, на примере глаголов *сердиться* и *злиться* нами было продемонстрировано, как статистико-комбинаторные характеристики контекстов могут

использоваться для выявлении семантических отношений между синсетами, и их верификации. Пример, рассмотренный нами в данной статье, является довольно сложным с точки зрения методов традиционной лексической семантики, поэтому был выбран нами не случайно. Как было показано во втором разделе статьи, адекватное описание эмоциональной лексики в значительной мере затруднено спецификой ее семантики; методы логического, дефиниционного или деривационного анализа в данном случае оказываются практически неприменимы. Поэтому описанная нами методика несомненно предоставляет особый интерес как дополнение к традиционным методам установления семантических отношений между лексическими единицами.