Правительство Российской Федерации

Федеральное государственное автономное образовательное учреждение высшего профессионального образования

Национальный исследовательский университет "Высшая школа экономики"

Факультет филологии

Направление подготовки 035800.68 «Фундаментальная и прикладная лингвистика»

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

На тему «Построение лексико-типологической анкеты с помощью моделей дистрибутивной семантики»

Студент группы №72л
Д.А. Рыжова
(Ф.И.О.)

Научный руководитель
Профессор факультета филологии НИУ ВШЭ, д.ф.н. Е.В. Рахилина
(должность, звание, Ф.И.О.)

Москва, 2014 г.

Оглавление

Введение

Лексическая типология – молодая и бурно развивающаяся отрасль лингвистики. Первой работой в этой области считается проект по анализу цветообозначений в различных языках, начатый Берлином и Кеем (см. Berlin&Kay 1969) и подхваченный впоследствии множеством последователей и оппонентов из разных стран мира (см., например, Фрумкина 1984, Wierzbicka 1990, Corbett&Davies 1995, Levinson 2000). Методология, разработанная в рамках этого проекта, легла в основу самого распространенного на сегодняшний день подхода к лексико-типологическому анализу – психолингвистического метода, практикуемого исследовательской группой Института имени Макса Планка в Неймегене (см. Majid&Bowerman 2007, Majid&Levinson 2008 и др.). Этот метод базируется на эксперименте. Для каждого изучаемого семантического поля подготавливается набор стимулов: карточек, выкрашенных в разные цвета, - для исследования цветообозначений; клипов, представляющих ситуации разделения объектов на части, - для глаголов разбиения, разрезания и т.п.; пузырьков со смесями различного вкуса – для изучения вкусовых прилагательных и т.д. Затем эти «интерактивные» анкеты предъявляются испытуемым, которые должны описывать стимулы словами своего родного языка.

Этот метод обладает целым рядом достоинств. Прежде всего, единая анкета, по которой опрашиваются информанты, становится базой для сравнения и позволяет проводить параллели между лексическими составами разных, в том числе и неродственных, языков, формулировать правила, выводить типологически значимые закономерности. Кроме того, такой метод позволяет быстро получить необходимую информацию, причём не только о хорошо изученных языках с богатой письменной традицией, но и о малых бесписьменных языках, для которых нет ни корпусов, ни хороших и достаточно полных толковых и переводных словарей.

Однако, несмотря на все преимущества, этот метод не лишен недостатков. Во-первых, он накладывает серьёзные ограничения на область исследования: если прилагательные со значением ‘сладкий’ или ‘солёный’ можно изучать, предлагая испытуемым пробовать воду с сахаром или с солью, то как должна выглядеть анкета для анализа предикатов боли (ср. русск. *колет в боку, першит в горле*), оценочных прилагательных (*хороший, ужасный, божественный, отвратительный* и т.п.) или метафорических сдвигов (‘сладкая вода’ vs. ‘сладкая жизнь’) остаётся неясным.

Другая слабая сторона психолингвистического метода состоит в том, что он предполагает изучение наименований тех или иных физических характеристик или

процессов в отрыве от контекста, от реальной жизненной ситуации, в которой искомые слова могут использоваться. Так, например, один и тот же с физической точки зрения цвет может описываться в русском языке разными прилагательными, в зависимости от того, атрибутом какого объекта он является, ср.: *карие глаза* vs. *каштановые волосы* vs. *коричневое платье* vs. *бурый медведь* (подробнее об этом см. Рахилина 2008).

Альтернативный, фреймовый, подход[1] к лексической типологии (см. Рахилина, Резникова 2013), которого придерживаемся мы, призван преодолевать подобного рода сложности. Он восходит к традициям Московской Семантической Школы (см. Апресян 1974) и предполагает анализ лексемы через призму её сочетаемостных свойств. Иными словами, мы утверждаем, что семантика лексемы может быть представлена не в виде традиционного толкования, набора сем или возможных денотатов, а в виде правил сочетаемости. Проиллюстрируем это положение на примере прилагательного *тонкий*.

Признак ‘тонкий’ описывает ситуацию с одним обязательным участником – носителем признака. В роли этого участника могут выступать разнообразные объекты: верёвки, шнурки, тетрадки и слои пыли, слух и нюх, голос или какой-нибудь механизм или сложный прибор. Однако легко заметить, что в сочетании с объектами разного типа прилагательное *тонкий* принимает разные значения: в сочетании с существительными, обозначающими длинные вытянутые (такие, как верёвка, хвост, палец, дерево и т.п.) и плоские предметы (книга, доска, лента, ткань) оно обозначает их небольшую толщину, причем в первом случае толщина – это диаметр поперечного среза, а во втором – просто расстояние от одной плоскости до другой (например, от одного форзаца книги до другого); в сочетании со словами, обозначающими различные звуки, – их небольшую громкость и высокий частотный диапазон; со словами *прибор* или *механизм*, а также с названиями процессов восприятия – способность обнаруживать малейшие изменения, реагировать на легчайшие воздействия. Получается, что семантику прилагательного можно представить в виде набора правил его сочетаемости:

- *тонкий* + названия длинных вытянутых предметов = малый диаметр поперечного среза;
- *тонкий* + обозначения звука = слабая громкость и высокий частотный диапазон и т.п.[2]

[1] Этот подход в настоящее время активно развивает Московская лексико-типологическая группа MLexT во главе со своим основателем Е.В. Рахилиной (см. Рахилина 2013). В состав этой группы входит и автор настоящей работы.

[2] Ср., например, попытку использовать такого рода правила для снятия семантической омонимии в Национальном корпусе русского языка (см. Шеманаева и др. 2007)

Более того, разным значениям прилагательного будут соответствовать разные переводные эквиваленты, причём процесс выбора нужного слова можно также представить в виде правил сочетаемости. Ср., например, перевод прилагательного *тонкий* на китайский:

- ‘тонкий’ + название длинного вытянутого предмета => *xì* (*xì gùnzi* – ‘тонкая палка’);
- ‘тонкий’ + название плоского предмета => *báo* (*báo zhǐ* – ‘тонкая бумага’) и т.д. (подробнее см. Кюсева и др. 2013).

Такой метод изучения семантики слов был опробован на обширном лексическом материале, ср. Брицын и др. 2009, Майсак, Рахилина 2007, Круглякова 2010, Кюсева 2012. В ходе проведенных исследований доказано, что такой подход действительно позволяет аккуратно анализировать лексемы как в рамках одного языка, так и на типологическом уровне, составляя простые правила соответствия переводных эквивалентов друг другу. Полученные результаты дают основание ставить перед собой и такие амбициозные цели, как составление подробных мультиязычных словарей, которые давали бы представление не только о примерном значении, но и о вполне конкретных правилах употребления слов[3]. Этот метод кажется перспективным и с точки зрения машинного перевода, поскольку он не имеет дела ни с какими абстрактными определениями и толкованиями, а вместо этого опирается на простые списки возможных контекстов употребления лексемы.

Однако все эти важные задачи в настоящее время невозможно решить, даже несмотря на то, что сам алгоритм кажется нам простым и понятным. Сложность заключается в том, что составление для каждой лексемы правил сочетаемости и правил перевода на другие языки – это очень трудоёмкий процесс, требующий месяцы упорной работы сразу нескольких специалистов.

Задача настоящего проекта – разработать метод, который позволит частично устранить этот недостаток, т.е. автоматизировать хотя бы один из этапов лексико-типологического исследования. В первой части Главы 1 нашей дипломной работы мы расскажем подробнее об алгоритме составления правил сочетаемости и, в особенности, правил перевода и покажем, какую именно его стадию мы предполагаем автоматизировать; во второй части мы представим метод автоматического анализа лексики (модели дистрибутивной семантики), который мы будем использовать для решения наших задач. В главе 2 будут описаны все стадии работы предлагаемого нами

[3] Современные словари не очень хорошо справляются с этой задачей. В большинстве случаев, в качестве перевода для одного русского слова предлагается список слов другого языка, а правила их распределения остаются за кадром (подробнее об этом см. Кюсева и др. 2013)

алгоритма. И, наконец, в последней главе мы предложим ряд выводов и соображений относительно дальнейшей работы в этом направлении.

Автор благодарит своих научных руководителей Екатерину Владимировну Рахилину и Дениса Паперно за постоянную поддержку, неусыпный контроль, ценные и своевременные советы и замечания, а также Светлану Юрьевну Толдову, любезно согласившуюся прочесть и оценить настоящую работу.

Глава 1. Постановка задачи

1.1. Лексико-типологическая анкета

Фреймовый подход к лексической типологии основан на анализе контекстов, в которых может употребляться исследуемая лексема. Для этого очень хорошо приспособлены корпусные технологии. Так, Национальный корпус русского языка, располагающий морфологической и семантической разметкой, позволяет получить, например, все словосочетания изучаемого прилагательного с существительными, зарегистрированные в корпусе, а также просмотреть отдельно все возможные сочетания признакового слова с существительными разных таксономических типов.

Однако только корпусных данных, к сожалению, недостаточно. Прежде всего, потому, что корпуса не предоставляют никакой отрицательной информации. Исследователь не может предполагать, что если словосочетание не встретилось в корпусе, то оно действительно невозможно в данном языке. Корпуса, какими бы большими и аккуратно собранными они ни были, часто оказываются недостаточно сбалансированными и объёмными. Это особенно релевантно для задач, связанных с исследованием лексики: лексический состав языка меняется очень быстро, а представлять в корпусе все новейшие сленговые употребления слов практически невозможно. К тому же, анализ лексики требует очень больших объёмов текстовых данных, так как слова, а тем более словосочетания, встречаются значительно реже, чем, например, грамматические показатели.

Помимо этого, представительные национальные корпуса созданы лишь для очень небольшого количества языков, большая часть которых – крупные европейские, а для типологии, несомненно, большой интерес представляют и малые «экзотические» языки.

В этой ситуации основным инструментом исследования становится единая анкета-опросник, которая одинаково применима для сбора материала как крупных языков с богатой традицией, так и малых языков с неразвитой или полностью отсутствующей письменностью. Эта же анкета становится базой для сравнения данных разных языков между собой.

Для успешного выполнения своих функций анкета должна удовлетворять ряду требований. Во-первых, она должна быть представительной: в ней должны быть учтены все основные типы контекстов, в которых могут употребляться лексемы изучаемого семантического поля. Это нетривиальная задача, так как известно, что слова разных языков обычно не совпадают по объёму значений, и поэтому простого перевода всех словосочетаний, допустимых для некоторой лексемы в одном языке, недостаточно для

полного анализа её переводного эквивалента в другом. Так, например, английский эквивалент русского прилагательного *толстый – thick* – может сочетаться со словами ‘туман’, ‘дым’ и т.п., выражая при этом значение ‘густой’, которое в русском языке не только покрывается другой лексемой, но и – на интуитивном уровне – не имеет никакого отношения к признакам размера. Между тем, такое объединение очень частотно: оно наблюдается во французском, немецком, кабардинском и других языках.

С другой стороны, анкета должна быть не очень большой: в противном случае её будет очень сложно предъявлять информантам. Поэтому важно не просто собрать список всех возможных словосочетаний, в которых могут фигурировать лексемы изучаемого семантического поля, но и расклассифицировать их, выделив набор различных групп употреблений, релевантных для данной семантической области. Под релевантными мы подразумеваем такие группы контекстов, которые, согласно нашим предположениям и эмпирическим данным, никогда не разделяются, т.е. во всех языках покрываются одним прилагательным, а не разными. Такие минимальные ситуации, на которые реагирует язык, мы и называем фреймами[4].

Так, например, наши исследования признакового поля ‘тяжёлый’[5] показали, что для него важны следующие фреймы:

1. ситуация транспортировки объекта;
2. ситуация определения веса объекта, не предполагающая его дальнейшую транспортировку;
3. зрительное восприятие объекта и др.

Эти ситуации в каждом языке реализуются в виде набора словосочетаний. Так, например, в русском языке к описанию ситуаций первого типа относятся словосочетания *тяжёлая сумка*, *тяжёлый рюкзак*, *тяжёлый шкаф*; второго типа – *увесистая дубинка*, *увесистый том*; третьего – *массивная люстра*, *тяжёлые своды*, *тяжеловесные шторы*.

Анкета должна включать в себя лишь несколько примеров реализации каждого типа ситуаций. При этом важно, чтобы примеры были достаточно представительными. Предполагается, что анкета будет предъявляться носителям очень разных языков и культур, а значит, она должна быть построена на основе максимально универсальных понятий и реалий. Например, в качестве иллюстрации фрейма тяжелого веса объекта, ощущаемого при его транспортировке, правильнее использовать формулировку ‘тяжёлая сумка’, а не ‘тяжёлая авоська’ или ‘тяжелая барсетка’.

[4] Подробнее о понятии фрейма см. Рахилина, Резникова 2013 и Резникова и др. 2013
[5] См. Кюсева и др. 2012, а также Холкина 2014

Таким образом, разработка анкеты – очень важная и трудоёмкая часть лексико-типологического исследования, включающая в себя несколько этапов.

Первый вариант анкеты создаётся на основе русского языка. Этот язык является для нас родным, что позволяет нам собрать по нему надежный и достаточно полный материал, а затем подробно и адекватно его проанализировать. В первую очередь с помощью НКРЯ собираются и изучаются возможные контексты употребления лексем рассматриваемого семантического поля. Спорные места уточняются с помощью опросов носителей. Далее полученные контексты делятся на смысловые группы – будущие фреймы. На материале одного языка выделение фреймов происходит на основе теоретических рассуждений. Так, например, для анализа признаковых семантических полей в анкету включаются в основном сочетания прилагательных изучаемого поля с различными существительными (носителями признака), при этом общий набор контекстов делится на классы в зависимости от семантики существительного. При самом грубом делении в каждый класс попадают сочетания с именами одного таксономического класса. Например, в анкете для поля ‘острый’ в один класс попадут названия инструментов (‘нож’, ‘серп’, ‘игла’, ‘копьё’), в другой класс – обозначения различных видов пищи (‘суп’, ‘соус’, ‘приправа’), в третий – обозначения разного рода взаимодействий (‘конфликт’, ‘конфронтация’, ‘дебаты’). Дальнейшее деление происходит на основе других параметров, которые, как нам известно из нашего предыдущего опыта, часто бывают релевантны для признаковых семантических полей. Так, нам известно, что часто важную роль играет не только таксономическая, но и топологическая классификация существительных, т.е. классификация с точки зрения формы обозначаемых этими словами предметов (ср. пример выше про признак ‘тонкий’). В соответствии с этим параметром, первый класс анкеты для семантического поля ‘острый’ разделится на две части: объекты с режущим краем (‘нож’, ‘серп’) и колющим концом (‘игла’, ‘копьё’), т.е. «линия» vs. «точка». Однако важно понимать, что теоретических рассуждений для выделения фреймов в любом случае недостаточно, поэтому анкета, составленная на материале одного языка, создаётся по возможности с некоторой избыточностью: каждый выделенный класс иллюстрируется достаточно большим количеством примеров.

На следующем этапе анкета «тестируется» на материале другого языка, причём желательно, чтобы этот язык был либо близкородственным русскому[6], либо крупным языком, располагающим хорошими корпусами и выгодно отличающимся от других большим количеством доступных исследователю информантов. Всё это необходимо для

[6] Подробнее о том, что лексико-типологическое исследование удобно начинать с анализа близкородственных языков, см. Кюсева, Рыжова 2012.

того, чтобы новый язык можно было также изучить достаточно подробно, опираясь не только на анкету, но и на другие источники информации. Привлечение нового материала всегда влечет за собой уточнение и расширение анкеты: обнаруживаются новые противопоставления, новые метафорические сдвиги.

Уточняется при этом и материал русского языка. Это происходит за счёт того, что в новом языке слова рассматриваемого семантического поля покрывают, как правило, и контексты, которые в русском языке обслуживаются другими лексемами, не включенными в исследование изначально на основании чисто теоретических соображений (ср. пример выше про прилагательное *густой*).

Возврат к русскому после анализа каждого нового языка и последующее уточнение анкеты мы называем челночным методом (см. Рахилина, Резникова 2013). Очень важно, что этот процесс продолжается не до бесконечности: надежная анкета, способная играть роль основного инструмента исследования, формируется после последовательного анализа 3-5 языков. В дальнейшем она ещё может подвергаться разного рода модификациям, но уже лишь в деталях.

Таким образом, в результате мы получаем анкету, состоящую из набора контекстов, разделенных на группы и представляющих ситуации, которые в разных языках могут покрываться лексемами рассматриваемого семантического поля (см. анкету для поля ‘острый’ в приложении, а также её фрагмент в таблице 1). Возможно ли построить такую анкету автоматически, причём на материале только одного языка, например, русского? На этот вопрос и призвана ответить настоящая работа.

Фрейм	**Словосочетание-иллюстрация**
‘острый инструмент с режущим краем’	*Острый нож*
	Острый меч
‘острый инструмент с колющим концом’	*Острая игла*
	Острый гвоздь
‘четкая линия’	*Резкий контраст*
	Резкая фотография
‘колющаяся поверхность’	*Колючее одеяло*
	Колючий свитер
…	

Таблица 1. Фрагмент анкеты для семантического поля ‘острый’

1.2. Векторные модели

Модели дистрибутивной семантики (или DSM – Distributional Semantics Models, см. Baroni et al. 2013) лучше всего подходят для решения наших задач, поскольку они основываются на тех же самых теоретических предпосылках: на гипотезе о том, что значение слова можно представить через его сочетаемость. При этом дистрибутивная семантика предлагает алгоритмы автоматической оценки контекстов, в которых употребляется изучаемая лексема, на базе статистики, собранной по обширным корпусам текстов, а такого рода данные должны быть существенно полнее и точнее результатов ручного анализа корпусных материалов.

Ключевое понятие DSM – семантический вектор, с помощью которого представляется сочетаемость лексемы. Каждый такой вектор соответствует одной лексеме и состоит из некоторого числа измерений. В качестве измерений выступают, как правило, слова, в контексте которых лексема может употребляться. Таких измерений, в зависимости от задачи, может быть несколько десятков, несколько сотен или несколько тысяч. Чем больше у вектора измерений, тем точнее он моделирует сочетаемость лексемы. В качестве измерений обычно выбираются слова знаменательных, а не служебных частей речи, и по возможности достаточно частотные, чтобы можно было собрать больше статистики.

Значение каждого измерения – количество употреблений лексемы в контексте данного слова. Понятие контекста также определяется строго: он может быть синтаксическим (см., например, Padó&Lapata 2007) или простым контактным. В случае ориентации на синтаксический контекст вектор строится по корпусу с синтаксической разметкой, и контекстом опорного слова считаются лексемы, встретившиеся на некотором синтаксическом расстоянии от искомого (например, не далее, чем в трёх шагах по дереву зависимостей). Однако чаще всего векторные модели строятся на основе простого понятия совместной встречаемости (co-occurrence): устанавливается фиксированный размер окна, и это окно считается контекстом рассматриваемого слова. Таким образом, вектор – это набор значений всех измерений данного векторного пространства.

Так, например, в модели с двумя измерениями (ось 1: *пить*, ось 2: *есть*) и размером окна ±1 знаменательное слово семантические вектора для существительных *чай*, *кофе* и *мороженое* могли бы иметь следующий вид:

чай: <348, 13>

кофе: <303, 2>

мороженое: <1, 297>

Особое преимущество такого способа представления информации заключается в том, что с геометрическими объектами и числовыми данными можно производить разного рода математические операции. Прежде всего, вектора разных слов можно сравнивать между собой, определяя степень близости между ними и, следовательно, делая выводы о степени сходства языковых единиц. Самая распространенная мера близости – косинус угла между векторами. Часто используется также мера Евклидова расстояния между конечными точками векторов, но применение этой метрики даёт содержательный результат только в том случае, если сравниваемые вектора имеют равную длину. На рисунке 1 продемонстрирован случай, когда два вектора расположены достаточно близко друг другу (о чём и говорит косинус угла между ними), но при этом Евклидово расстояние довольно велико. Такое расхождение часто происходит в случае сравнения между собой слов, обладающих сходными дистрибутивными свойствами, но очень разных с точки зрения общей частотности (ср. *машина* и *автомобиль*). Исходно различные по длине вектора можно нормализовывать, приводя их к общему виду.

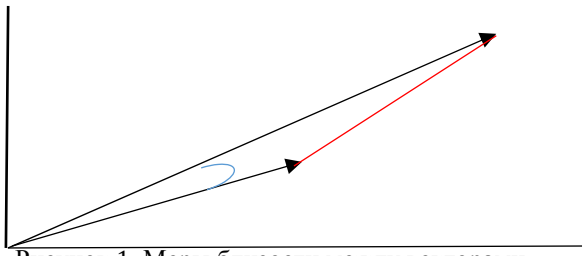

Рисунок 1. Меры близости между векторами

Для наиболее точного вычисления степени близости между векторами, помимо нормализации, используются разного рода другие вспомогательные операции. В частности, распространенной проблемой является недостаток статистических данных, который преодолевается разными способами. Во-первых, с высокой разреженностью векторов (т.е. большим количеством нулевых значений измерений) можно бороться с помощью уменьшения размерности пространства. Эта процедура подразумевает сокращение количества измерений за счет объединения в один класс похожих осей. Например, если в наборе измерений есть слова *яблоко* и *груша*, то скорее всего их значения всегда будут взаимосвязаны: возможно, лексема *груша* менее частотна, чем *яблоко*, но при этом они появляются в большинстве случаев в похожих контекстах, поэтому есть основания предполагать, что значение измерения *яблоко* будет всегда во сколько-то раз больше, чем значение измерения *груша*. Аналогичным образом, предположительно, должны быть устроены и другие подобные оси, такие как *апельсин*, *лимон*, *банан* и т.д. В таком случае все эти измерения можно объединить в одно более общее (грубо говоря, измерение «названия фруктов»). Иными словами, уменьшение

размерности пространства – это операция, симметричная кластеризации векторов, которая способствует не только уменьшению разреженности данных, но и сокращению времени, требуемого для их последующей обработки.

Ещё один тип операций полезен для представления информации о сочетаемости единиц, больших, чем слово, в первую очередь, двухсловных словосочетаний. Модели дистрибутивной семантики позволяют собирать вектора и для такого рода опорных элементов: в таком случае словосочетание просто считается единой языковой единицей, и значение измерений вычисляется в зависимости от того, в каких контекстах этот юнит употребляется. Однако очевидно, что частотность словосочетания значительно ниже, чем частотность каждой из его составляющих в отдельности, поэтому для сбора такого рода статистики требуются корпуса очень больших объемов, которыми исследователь, как правило, не располагает. В такой ситуации применяется алгоритм композиции: вектор словосочетания вычисляется на основе векторов составляющих его элементов. Для пересчета данных используются разные алгоритмы: сумма или произведение значений каждого измерения; или же на основе небольшого обучающего корпуса составляются формулы более сложных зависимостей. Эта процедура называется композицией (см. Mitchell&Lapata 2010). Как и операция по уменьшению размерности пространства, она позволяет решить проблему недостатка данных и серьёзно сэкономить время: вместо того, чтобы для каждого нового словосочетания собирать вектор сочетаемости, можно построить его с помощью композиции из уже обсчитанных элементов.

Векторные модели дистрибутивной семантики уже показали свою состоятельность в различных сферах NLP. Так, например, они успешно используются для решения задач семантической дизамбигуации (см. Schütze 1998, Agirre&Edmonds 2006), кластеризации текстов и нахождения документов по запросам пользователей (см., например, Salton 1971), извлечения отношений (см. Lin&Pantel 2001) и т.д. Однако в области типологических исследований такого рода модели, насколько нам известно, ещё никем не применялись.

Первым шагом в направлении внедрения алгоритмов дистрибутивной семантики в лексическую типологию стала выпускная квалификационная работа М.В. Кюсевой (см. Кюсева 2014). В этой работе представлен следующий эксперимент. По нескольким корпусам текстов на русском языке (основному и газетному подкорпусам НКРЯ, а также корпусу интернет-текстов РУВАК Сергея Шарова[7]) были собраны вектора для словосочетаний, вошедших в анкету, лежавшую в основе типологического анализа семантического поля качественного признака ‘острый’. Этот признак уже был изучен ранее на материале 20 языков (см. Кюсева 2012), поэтому для него существует надежная и

[7] О корпусе НКРЯ см.: http://ruscorpora.ru; РУВАК: http://corpus.leeds.ac.uk/internet.html

неоднократно проверенная типологическая анкета, а также собран обширный языковой материал. Основной целью эксперимента было определение степени соответствия данных, полученных на основе анализа векторной модели, результатам типологического исследования, проведенного вручную по традиционной методологии группы MLexT.

Данные различных языков, полученные в рамках работы над проектом, посвященном качественному признаку ‘острый’, включены в типологически ориентированную Базу данных признаковой лексики. В этой базе данных материал представлен в следующем виде: единицей входа является строка анкеты, и для каждого прилагательного, относящегося к рассматриваемому полю, указано, покрывает оно данную строку анкеты или нет[8]. Такая форма организации материала позволяет численно представить степень «типологической близости» одних строк анкеты к другим. В работе Кюсева 2014 эта мера вычисляется на основе того, насколько часто каждая пара строк покрывается одним и тем же словом: для каждой пары словосочетаний из анкеты подсчитывается количество лексем, которые охватывают либо обе эти ситуации, либо ни одной из них, и затем полученное число делится на общее количество прилагательных поля ‘острый’, зарегистрированных в Базе. Так, если в Базе данных нет ни одной лексемы, которая описывает *острый гвоздь*, но не употребляется в сочетании с существительным, обозначающим иглу, то значение метрики для контекстов ‘острый гвоздь’ и ‘острая игла’ будет равняться единице. А если в Базе есть 10 слов, которых сочетаются с существительным ‘игла’, но не сочетаются с существительным ‘нож’ при том, что всего в Базу введено 25 лексем, то пара контекстов ‘острая игла’ и ‘острый нож’ получит значение 0,6.

С другой стороны, для вектора каждого словосочетания была посчитана косинусная мера его близости с векторами всех остальных словосочетаний из анкеты. Таким образом, все строки анкеты были попарно сопоставлены друг с другом, причём двумя разными способами, поэтому для каждой пары было получено две разных оценки степени близости. Затем для этих двух мер была подсчитана корреляция, с целью определить, насколько похожие результаты они показывают.

Полученная величина оказалась очень высокой[9], что лишний раз подтверждает адекватность фреймового подхода к лексической типологии. Проведенный эксперимент доказывает, что фрейм – это не теоретический конструкт и не плод фантазии исследователя, а естественное объединение похожих ситуаций, воспроизводимое от языка

[8] Подробнее о Типологически ориентированной базе данных признаковой лексики см. Резникова и др. 2013.

[9] 0,726 при том, что максимально возможное значение равно 1.

к языку и проявляющееся в том, что в ареально и генетически разных языках для описания ситуаций, входящих в один фрейм, используется одна и та же лексема. Мы предполагаем, что фреймовая структура каждого поля универсальна, т.е. представляет собой решётку, накладываемую на каждый язык. При этом каждый конкретный язык заполняет этот каркас по-своему: одни противопоставления выделяются особенно ярко (например, маркируются разными лексемами), другие, наоборот, смягчаются (к примеру, проявляются только в разнородных переносных значениях). Таким образом, фреймовая структура поля угадывается уже на материале одного языка, что и доказывается в работе Кюсева 2014: вектора сочетаемости словосочетаний из одного фрейма ближе друг другу, чем вектора словосочетаний из разных фреймов.

С другой стороны, полученный результат показывает, что аппарат дистрибутивной семантики действительно в некотором смысле имитирует ручную работу лексического типолога, проводящуюся по методологии группы MLexT. Это означает, что векторные модели могут быть использованы для решения задач автоматизации тех или иных этапов лексико-типологического исследования, в частности, построения анкеты-опросника. Причем, что особенно важно, попытка автоматического сбора списка релевантных для изучаемого семантического поля ситуаций и разделение их на фреймы может базироваться на материале одного языка.

1.3. Задачи настоящего исследования

Задача настоящего исследования – разработать алгоритм автоматического (или полуавтоматического) составления лексико-типологической анкеты на основе данных одного языка – русского. В качестве экспериментального лексического материала мы используем признаковое поле ‘острый’. Такой выбор продиктован сразу несколькими соображениями. Во-первых, это поле уже хорошо изучено по нашей традиционной методологии, поэтому мы сможем легко оценить результаты, полученные новым способом. Во-вторых, именно на этом материале проводились первые опыты моделирования векторных пространств для лексико-типологических исследований (см. Кюсева 2014), благодаря чему мы сможем быстрее подготовить техническую базу для новых экспериментов.

Для того, чтобы составить полноценную анкету-опросник, удовлетворяющую всем требованиям, представленным в первой части Главы 1, нам необходимо решить несколько задач. А именно:

1) определить, какие прилагательные русского языка могут иметь отношение к изучаемому полю;
2) для каждого прилагательного составить список коллокаций, т.е. существительных, в сочетании с которыми оно может употребляться;
3) подготовить векторное пространство, т.е. для каждого отдельного словосочетания посчитать вектор его сочетаемости;
4) разделить словосочетания на группы (потенциальные фреймы), проведя кластеризацию собранных векторов;
5) уменьшить размер анкеты, отобрав из каждого кластера несколько наиболее представительных элементов;
6) оценить полученные результаты.

В следующей главе мы рассмотрим последовательно каждую из этих задач и подробно обсудим возможные методы их решения.

Глава 2. Разработка алгоритма автоматического составления анкеты-опросника

2.1. Определение круга прилагательных, относящихся к изучаемому полю

Задача составления списка прилагательных, относящихся к тому или иному семантическому полю, при своей кажущейся простоте связана с большим количеством сложностей. Обычно, начиная анализ нового признакового поля в рамках методологии MLexT, мы отталкиваемся от одного русского прилагательного и, как правило, даже не предпринимаем попыток оценить, какие ещё лексемы могут быть отнесены к той же семантической зоне, что и слово, оказавшееся в центре нашего внимания.

Такой шаг вполне оправдан и не ведет к потере информации. Наш опыт показывает, что русские признаковые поля, как правило, устроены в соответствии с доминантной стратегией лексикализации значений: есть одно прилагательное, которое покрывает почти все релевантные для данного поля фреймы, и несколько периферийных слов, редко употребляемых и обладающих очень узким набором значений, ср. исследования полей ‘острый’ (Кюсева 2012), ‘полный’ (Тагабилева, Холкина 2010), ‘тяжёлый’ (Кюсева и др. 2012), ‘мягкий’ (Павлова 2014). Таким образом, подробный анализ одного русского прилагательного в большинстве случаев позволяет исследователю составить достаточно полный список ситуаций, относящихся к данной семантической зоне.

Однако возможность начинать исследование с анализа одного-единственного слова обеспечивается в первую очередь даже не доминантным устройством русских признаковых полей, а особенностями нашей методологии. Недостаток данных на первом шаге исследования компенсируется на следующих этапах, а именно в процессе анализа переводных эквивалентов русского прилагательного. Анализ наборов их значений позволяет расширить предполагаемые границы поля, дополнить типологическую анкету и пересмотреть материал русского языка, включив в исследование новые лексемы. Этот челночный метод, о котором мы уже говорили в Главе 1, позволяет ничего не упустить даже в том случае, если в русском языке в изучаемом поле нет доминанты. Так, например, в уже обсуждавшемся выше семантическом поле признаков размера в русском языке, как кажется, нельзя выделить доминантное прилагательное: по крайней мере, лексемы *узкий* и *тонкий* в зоне малых размеров (или *широкий* и *толстый* – в зоне больших) кажутся равноправными. При этом исследование было начато с лексемы *тонкий*, а затем, благодаря материалам других языков (прежде всего, английского, французского, китайского и хантыйского), в которых наборы значений переводных эквивалентов этого слова охватывали зоны, характерные для других русских прилагательных, стало ясно, что

к этому полю следует также отнести лексемы *узкий*, *тесный* и, вероятно, *жидкий* (ср. пример про прилагательное *густой* выше).

Компьютерный подход к составлению лексико-типологической анкеты, который мы разрабатываем, не может опираться на материал сразу нескольких языков (прежде всего потому, что в нашем распоряжении имеются корпуса текстов только на русском языке). Соответственно, в нашу задачу входит поиск простого и быстрого способа определения лексического состава семантического поля, не предполагающего предварительного подробного анализа обширного типологического материала.

Ещё одно обстоятельство, затрудняющее процесс составления списка лексем, относящихся к тому или иному семантическому полю, связано с самим понятием поля. В рамках нашего подхода под семантическим полем подразумевается набор связанных между собой фреймов[10]. Алгоритм составления списка фреймов мы неоднократно упоминали выше: в его основе лежит челночный метод, согласно которому мы отталкиваемся от одной русской лексемы, составляем список её значений (т.е. групп контекстов, в которых она употребляется), затем анализируем её переводные эквиваленты, находим среди их значений новые, не входящие в сферу действия русского слова, с которого мы начали, возвращаемся к русскому языку, ищем лексемы, покрывающие добавленные контексты, и повторяем для них всю процедуру.

Этот алгоритм схематически представлен в таблице 2. Предположим, что в качестве отправной точки исследования выбрано прилагательное lex 1 языка L1. У этого прилагательного три значения, т.е. три явно выделяемые группы контекстов употребления (Sem 1, Sem 2 и Sem 3). При этом в некотором другом языке, L2, эти значения выражаются с помощью двух разных прилагательных: первое, lex 2, охватывает значение 1 и часть контекстов значения 2, а также некоторое новое значение 4, которого нет у исходной лексемы lex 1; второе, lex 3, покрывает группу контекстов Sem 3 и другую часть группы контекстов Sem 2 и имеет при этом ещё два значения, ранее нам не встречавшихся (Sem 5 и 6). Таким образом, язык L2 позволяет выделить три новых группы контекстов и одно новое противопоставление внутри группы, которая изначально казалась нам однородной (Sem 2). Далее, на третьем этапе исследования, оказывается, что в языке L1, с которого мы начинали, для каждого из выявленных значений (Sem 4, 5 и 6) есть отдельная лексема, выражающая, в свою очередь, и некоторые другие значения, которые не фигурировали на предыдущих этапах исследования. К концу третьего этапа

[10] Ср. более традиционное определение из учебника по семантике М.А. Кронгауза: «Семантическим полем называется множество слов, объединенных общностью содержания, или, говоря более конкретно, имеющих общую нетривиальную часть в толковании» (см. Кронгауз 2005, стр.130)

работы исследователь получает список из десяти потенциальных фреймов (так как группа контекстов Sem 2 разбилась на две части) и 6 лексем. Для новых значений 7, 8 и 9 снова ищутся переводные эквиваленты в языке L2 и т.д.

Step 1
L1, lex 1
Sem 1
Sem 2
Sem 3

Step 2	
L2, lex 2	L2, lex 3
Sem 1	Sem 2.2
Sem 2.1	Sem 3
Sem 4	Sem 5
	Sem 6

Step 3		
L1, lex 4	L1, lex 5	L1, lex 6
Sem 4	Sem 5	Sem 6
Sem 7	Sem 8	
	Sem 9	

Таблица 2. Алгоритм составления списка фреймов, входящих в состав изучаемого семантического поля. Условные обозначения: L1, L2 – языки; lex 1-6 – слова, относящиеся к данному полю; Sem 1-9 – их значения. Курсивом выделены новые значения, не фигурировавшие на предыдущих этапах.

В Главе 1 мы отметили, что процесс составления списка фреймов конечен. Однако на самом деле это верно лишь отчасти. Практика показывает, что новые противопоставления (т.е. разбиения групп контекстов на части) действительно вскоре перестают появляться. Например, русское прилагательное *острый* не проводит различия между колющими и режущими предметами, поэтому изначально они все попадают в общий фрейм хорошо функционирующих, заточенных инструментов. Но уже материалы крупных европейских языков показывают, что это противопоставление необходимо. Так, во французском режущие инструменты описываются лексемой *tranchant,* а колющие – прилагательным *aigu*. Этот параметр выявляется на очень ранних этапах и воспроизводится в большом количестве языков. Однако никаких новых противопоставлений внутри каждой из этих групп даже после анализа 20 языков так и не появилось.

Иначе обстоит дело с появлением совсем новых значений, ранее в анкету не входивших. Процесс добавления новых контекстов употребления слов, видимо, потенциально бесконечен. Здесь необходимо отметить важную черту нашего алгоритма составления списка фреймов, которую мы до сих пор не упоминали: несмотря на то, что для каждого прилагательного в каждом языке анализируется полный спектр его употреблений, включая метафорические, новыми контекстами, требующими пересмотра материалов уже исследованных языков считаются только такие, в которых реализуются прямые значения лексем. Иными словами, если в сербском языке переводной эквивалент русского *острый* прилагательное *оштар* имеет значение ‘строгий (о людях)’, русскому прилагательному несвойственное, это не означает, что заполняющая эту нишу лексема *строгий* будет включена в исследование. Прилагательное *строгий* считалось бы относящимся к настоящему семантическому полю только в том случае, если бы, помимо

значения качества человека, оно могло выражать, например, значение ‘имеющий хорошо заточенное лезвие’ и сочетаться с существительными *нож*, *пила*, *бритва* и т.п.[11]

Такое ограничение заметно сокращает область поиска новых фреймов и, казалось бы, гарантирует, что полученный в результате список лексем будет состоять исключительно из представителей одной семантической области. Однако практика показывает, что этот идеал, видимо, недостижим, поскольку семантические поля (по крайне мере, признаковые), устроены принципиально иным образом: они никогда не бывают замкнутыми. Для каждого поля можно определить круг «соседей», смежных с ним семантических зон, с которыми оно неразрывно связано. Эта связь проявляется, как правило, в том, что в наборе фреймов прямых значений поля выделяется периферийный фрагмент, который в одних языках заполняется лексемами одного поля, а в других – другого.

Рассмотрим в качестве примера семантическое поле ‘тяжелый’. Центральным для этого поля является фрейм ‘предметы, которые трудно поднять или перенести’ (ср. в русском прототипические контексты *тяжелый чемодан*, *тяжелый рюкзак*, *тяжелая сумка*). Однако в японском, наряду с этими контекстами, лексема *omoi* (‘тяжелый’) может описывать ещё и педали, которые трудно крутить, и кнопки, которые трудно нажимать. В русском для кнопок и педалей используется прилагательное *тугой*, относящееся к другому полю. Его ядерным фреймом является ситуация ‘трудно растягивать’ (ср. *тугой лук*, *тугая резинка*). В свою очередь, в японском для этих ситуаций предпочитается лексема *katai*, описывающая тугой лук, но крайне редко – кнопки и педали. В результате фрейм ‘кнопка, которую трудно нажимать/ педаль, которую трудно крутить’ является в русском языке периферийным для поля ‘тугой’, а в японском – для поля ‘тяжелый’. Соответственно поля ‘тяжелый’ и ‘тугой’ оказываются смежными, а кнопки/педали образуют своего рода «мост» между ними.

Таким образом, представленный нами алгоритм определения набора фреймов может работать бесконечно, поскольку между разными семантическими полями есть переходные фрагменты, позволяющие переходить из одной зоны в другую, захватывая всё новые и новые контексты. Поэтому границы поля обычно определяются не собственно устройством языка, а решением исследователя, который выбирает, анализом сколь

[11] Отдельную проблему представляет задача определения того, какие значения можно считать прямыми, а какие – переносными. Мы исследуем преимущественно семантические поля качественных признаков, исходно обозначающие физические свойства предметов, поэтому в большинстве случаев прямыми значениями считаются физические, а переносными – абстрактные (ср. ‘острый нож’/ ‘острый нос’ vs. ‘острый ветер’/ ‘строгий человек’). Такое решение не всегда кажется удовлетворительным (например, фрейм ‘острый/резкий звук’, подразумевающий физическую характеристику объекта, явно является результатом метафорического сдвига), однако в рамках данной работы у нас нет возможности рассматривать этот сюжет более подробно.

объёмного набора фреймов он хочет ограничиться. Это означает, что у задачи составления списка прилагательных, относящихся к тому или иному семантическому полю, заведомо нет правильного ответа: результат зависит отчасти от субъективного решения, принимаемого исследователем.

Наконец, третья проблема, связанная с определением круга прилагательных, включаемых в область исследования, не играет роли при ручном сборе материала, но становится ощутимой при переходе на компьютерную основу анализа данных: проблема переносных значений. Как уже было сказано выше, при челночном методе отбора лексем не учитываются слова, обслуживающие только переносные значения интересующего нас поля. Однако это не означает, что метафорические сдвиги вообще исключены из анкеты: если у слова, описывающего хотя бы один из фреймов прямых значений, есть и переносные употребления, то все они учитываются в опроснике. Для нас эта информация очень важна: все наши исследования (ср. Кюсева и др. 2012, Кашкин 2013, Павлова 2014) показывают, что метафоры мотивированы исходными значениями слов, и поэтому модели семантических сдвигов воспроизводятся от языка к языку столь же регулярно, что и принципы лексикализации прямых значений. С другой стороны, между прямыми и переносными значениями есть очень существенные различия:

1) Набор фреймов прямых значений мы называем потенциально неограниченным потому, что семантические поля имеют проницаемые границы: если исследователь не очертит себе область изучения самостоятельно, то вполне возможно, что постепенно с помощью челночного метода в список фреймов попадут все физические значения, которые могут быть лексикализованы с помощью качественного прилагательного в естественном человеческом языке[12]. Мы предполагаем, что число таких фреймов может оказаться конечным, однако охватить их все в рамках одного исследования крайне затруднительно. Ситуация с переносными значениями принципиально иная: если анализ новых языков при ограничении области исследования очень быстро перестаёт приносить новые фреймы прямых значений (10-15 языков обычно уже достаточно), то новые метафоры, пусть единичные при наличии большого пласта повторяющихся переносов, но всё же продолжают появляться и в двадцатом, и в двадцать пятом языке.
2) Новые метафорические значения, появляющиеся практически в каждом новом для исследователя языке, также добавляются в анкету. Однако эти новые пункты

[12] См. аналогичное рассуждение в статье François 2008.

для большинства языков остаются незаполненными по указанной выше причине: в исследование включаются только прилагательные, охватывающие фреймы прямых (в нашем случае – физических) значений. Таким образом, строки прямых значений в анкете всегда заполняются, в то время как многие строки переносных значений оказываются заполнены лишь для одного-двух языков.

В рамках ручного метода работы проблема большого количества «метафорических лакун» решается просто. Так как фрейм – это ситуация, которая может реализовываться с помощью тех или иных словосочетаний, в самой анкете фреймы переносных значений представляются описательно (ср., например, одно из метафорических значений в анкете для поля ‘острый’: ‘суровый, строгий (о человеке)’) и иллюстрируются несколькими близкими примерами из русского языка, пусть и с использованием прилагательных, не относящихся к данному полю (ср.: *строгий человек*, *суровый взгляд* и т.п.). Для ручной работы ввод слов из другой семантической зоны не так страшен: важно, чтобы исследователь просто понимал, какое значение подразумевается под данным конкретным фреймом, чтобы материал нового языка можно было корректно отразить в заполненной анкете.

Для автоматического метода работы эта проблема, напротив, очень существенна. Разрабатываемый нами алгоритм определения структуры поля опирается на анализ словосочетаний вида «прилагательное + существительное», поэтому каждой строке анкеты нужно обязательно поставить в соответствие признаковую лексему. В рамках работы Кюсева 2014 это затруднение преодолевалось следующим путем: все строки анкеты были заполнены подходящими русскими прилагательными, независимо от того, к какому семантическому полю эти слова относятся, но для лексем из других зон в исследование были включены только контексты, затрагивающие поле ‘острый’, и никакие другие.

Однако и этот метод не решает проблему до конца. Природа метафорических значений такова, что во многих случаях для одного и того же фрейма можно подобрать несколько очень близких по смыслу прилагательных. Так, например, взгляд человека, обладающего ‘острым умом’ (т.е. умного), можно описать словами *умный* или *проницательный*, а в бесленеевском диалекте кабардинского языка сочетание прилагательного ‘острый’ с существительным ‘человек’ даёт значение ‘активный, подвижный, энергичный, деятельный’, и выбрать в качестве представителя фрейма один из этих русских эпитетов крайне затруднительно[13]. Всё это в очередной раз

[13] В большинстве случаев такой широкий набор возможностей возникает в зоне качеств человека, а также в области интенсификаторов и оценочных прилагательных. Вследствие антропоцентричности языка

свидетельствует, что единственного правильного варианта анкеты, как и единственного верного списка прилагательных, которые должны быть включены в исследование, не существует.

Анкета, созданная вручную и использовавшаяся при проведении экспериментов М.В. Кюсевой, включает 15 прилагательных: *острый, резкий, крутой, колючий, четкий, быстрый, яркий, умный, проницательный, сильный, непоседливый, хороший, высокий, грубый, газированный*. Невооруженным глазом видно, что многие слова попали сюда из-за «метафорических лакун»: ср. *непоседливый, газированный, сильный, высокий…*

В рамках нашего исследования мы опробовали несколько алгоритмов создания аналогичного списка автоматически, на материале только одного языка:

1. метод ближайших соседей;
2. метод анализа синонимов;
3. метод обратных переводов (полуавтоматический, с привлечением словарей других языков).

Рассмотрим последовательно все три метода и выделим их достоинства и недостатки.

2.1.1. Метод ближайших соседей

Метод ближайших соседей подразумевает анализ векторного пространства и определение нескольких векторов, наиболее близких к заданному. Иными словами, он позволяет получить список лексем, вектора которых находятся ближе всего (по косинусной мере расстояния) к вектору опорного слова (в нашем случае – прилагательного *острый*). Логично предположить, что именно такие слова и являются синонимами интересующей нас лексемы.

Мы протестировали этот метод на базе двух векторных пространств. Первое пространство состояло из векторов 10 000 наиболее частотных лексем знаменательных частей речи. Частотность определялась по имевшемуся в нашем распоряжении основному подкорпусу НКРЯ; по этому же текстовому материалу собирались и все вектора. Вектор каждого слова состоял из значений 10 000 измерений, причём в качестве набора измерений также использовались самые частотные знаменательные слова основного подкорпуса НКРЯ. Значение измерения показывало, сколько раз лексема, для которой строится вектор, встретилась в корпусе в контексте слова-измерения в окне ±5 знаменательных слов. Второе пространство отличалось от первого только размером окна:

средств для описания человека всегда очень много, а значения оценки и степени связаны с экспрессивной функцией, что также требует разнообразия способов их выражения (ср. Фрей 2006).

учитывались только контактные (с точностью до служебных слов) употребления в окне ±1 знаменательная лексема. В обоих случаях результатом эксперимента являлся список из 50 ближайших соседей вектора лексемы *острый*, из которого затем удалялись все не-прилагательные.

Эксперимент, проводившийся на базе пространства векторов, посчитанных по широкому окну, дал очень шумные результаты: ближайшим соседом прилагательного *острый* является его антоним *тупой*, а за ним следуют в беспорядке имена различных качественных признаков. Узкое окно, как и ожидалось, позволяет улавливать чуть более тонкие различия в употреблениях слов: в результирующем списке больше прилагательных, чем при широком окне. Однако различие, кажется, исключительно количественное: эксперимент с узким окном позволяет убрать несколько ненужных слов, которые появлялись в первом варианте списка, но зато добавляет пару десятков ничуть не более полезных лексем (см. таблицу 3).

Широкое окно	**Узкое окно**
Тупой	Тупой
Тяжелый	Жгучий
Тонкий	**Неприятный**
Легкий	Тяжелый
Жуткий	**Мучительный**
Блестящий	**Тоскливый**
Глубокий	**Любопытный**
Жгучий	**Выразительный**
Печальный	Сильный
Похожий	Жуткий
Злобный	**Темный**
Жесткий	Жесткий
Непривычный	**Серьезный**
Видимый	**Приятный**
Страшный	Тонкий
Маленький	**Грустный**
Сильный	**Внимательный**
Странный	**Сложный**
Беспокойный	Беспокойный
Уродливый	Злобный
Мрачный	**Нежный**
Сухой	**Нестерпимый**
Отвратительный	**Разнообразный**
Наглый	Печальный
Мягкий	Похожий
	Странный
	Жадный
	Привлекательный
	Схожий
	Интересный
	Осторожный
	Непривычный

	Стойкий
	Уродливый
	Длинный
	Страшный
	Тревожный
	Блестящий

Таблица 3. Ближайшие соседи прилагательного *острый* в широком и узком окне (из 50 соседей отобраны имена прилагательные). Жирным шрифтом выделены слова, которые встречаются только в одном из списков.

Между тем, неудовлетворительность полученных результатов неслучайна и легко объяснима. Метод ближайших соседей лучше приспособлен к определению когипонимов, чем синонимов. Ближе всего друг к другу находятся слова, которые употребляются в одних и тех же контекстах, т.е. прежде всего лексемы, обозначающие различных представителей одного и того же класса. Ср. первые шесть ближайших соседей существительного *яблоко*: *груша, виноград, апельсин, орех, яблоня, арбуз*. В нашем же случае мы предполагаем, что разные лексемы одного поля покрывают разные фрагменты, т.е. разные фреймы, этой семантической зоны. А разные фреймы, в свою очередь – это и есть группы контекстов. Тем самым, наше исходное предположение о том, что разные значения признаковых лексем реализуются в разных контекстах, заранее противоречит идее составления списка элементов поля по методу ближайших соседей.

Из этих наблюдений следует любопытный вывод о том, что семантические поля разных пластов лексики имеют различную внутреннюю организацию. Так, поля предметных слов в большинстве случаев строятся на когипонимических взаимоотношениях между элементами, т.е. они чаще всего в языке играют роль переменных, вставляемых в различные конструкции, причем существительные из одной семантической группы, как правило, появляются в одних и тех же контекстах (см. Рахилина 2008). Поля качественных признаков, напротив, строятся на основе принципа дополнительного распределения по разным типам переменных. Интересно, что относительные прилагательные в этом отношении ведут себя как существительные, ср. несколько ближайших соседей лексемы *синий*: *зеленый, желтый, голубой, лиловый, фиолетовый, серый, черный, розовый, коричневый, белый* и т.д.

2.1.2. Метод анализа синонимов

Второй метод составления списка элементов поля основан на извлечении близких друг другу прилагательных из словарей синонимов. Он даёт более приемлемые результаты, однако связан со своими сложностями.

В словаре синонимов Ю.Д. Апресяна (Апресян 2004), являющемся, на данный момент, наиболее авторитетным источником такого рода информации, для многих

качественных прилагательных, в том числе и слова *острый*, отдельных статей нет. Но зато нужный нам раздел есть в словаре под редакцией З.Е. Александровой (Александрова 2001), и в нём приведены следующие синонимы интересующего нас прилагательного:

Заостренный	Островерхий	Остроумный	Напряженный
Наточенный	Пряный	Мучительный	Резкий
Вострый (прост.)	Проницательный	Актуальный	

Таблица 4. Синонимы прилагательного *острый* по словарю З.Е. Александровой

Поскольку этот список составлялся вручную, в нём нет ничего «лишнего»: каждое слово синонимично какому-нибудь из значений лексемы *острый*. Однако не все эти слова могут быть включены в наше исследование: во-первых, к анализу признаковых зон мы стараемся не привлекать причастия (т.е. *заостренный* и *наточенный* будут отброшены); а во-вторых, слишком редкие слова, которые будет сложно анализировать на корпусном материале (т.е. *вострый* и *островерхий* также не войдут в окончательный список). Из оставшихся семи слов только два (*пряный* и *резкий*) охватывают некоторые физические значения, ср. *пряный вкус, резкое движение*. Остальные пять являются синонимами к метафорическим значениям прилагательного *острый*.

Опасаясь упустить что-нибудь важное, мы обратились к электронным базам с более широкими синонимическими рядами synonymizer.ru и synonymonline.ru. Эти ресурсы создавались с особыми целями: они призваны помогать составлять для контента сайтов уникальные тексты, не нарушающие авторских прав, формируя из одной статьи сразу несколько путём замены слов и конструкций на синонимичные.

Оба ресурса выдают для слова *острый* около сотни синонимов, что, конечно, для нас неприемлемо. Мы провели несколько этапов фильтрации: сначала в каждом из списков мы оставили только те слова, которые входят в 10 000 частотных знаменательных лексем по основному подкорпусу НКРЯ, а затем объединили два списка, оставив только те прилагательные, которые встречаются в обоих. Полученный в результате набор представлен в таблице 5.

В этой таблице результирующий список довольно грубо разбит на две части: прилагательные, охватывающие физические компоненты vs. слова, покрывающие переносные значения поля ‘острый’. Количество слов, относящихся к фреймам физических значений, по-прежнему мало (по сравнению со списком прилагательных, собранным вручную, куда входили слова *крутой* (ср. *крутой берег*), *колючий* (*колючие шипы*), *четкий* (*четкая линия, четкая фотография*), *высокий* (*высокая нота*)). Синонимов метафорических значений, напротив, остаётся очень много даже после всех уровней

фильтрации. Между тем, включение таких слов в исследование противоречит методологии MLexT (см. выше) и порождает большое количество фреймов, не имеющих отношения к полю *острый*. К примеру, прилагательное *современный* пересекается со словом *острый* лишь в одной группе контекстов, причём выражает значение, довольно далекое от семантики базовой лексемы, ср.: *современная проблема*, *современный недостаток* (= ‘нашего времени’, чаще во множественном числе) vs. *острая проблема*, *острый недостаток* (= ‘требующий немедленного разрешения’). При этом лексема *современный* очень частотная и обладает широкой сочетаемостью, так что, включая её в список прилагательных, относящихся к полю ‘острый’, исследователь получает множество словосочетаний, лишь портящих статистику (ср. *современные дети*, *современная Россия*, *современный стиль* и т.п.).

Синонимы физических значений	**Синонимы метафорических значений**
Острый	Мучительный
Жгучий	Лютый
Резкий	Напряженный
Пронзительный	Актуальный
	Современный
	Насущный
	Сильный
	Критический
	Драматический
	Своевременный
	Крайний
	Остроумный
	Чуткий
	Жестокий
	Нежный
	Тонкий

Таблица 5. Список синонимов прилагательного *острый*, полученный в результате обработки данных, предоставленных ресурсами synonymizer.ru и synonymonline.ru

Аналогичная ситуация происходит с признаком *тонкий*, который синонимичен слову *острый* в небольшом круге контекстов, ср.: *тонкое зрение*, *тонкий слух* vs. *острое зрение*, *острый слух*. При этом само прилагательное *тонкий* в своём исходном значении также обозначает качественный признак и относится к другому семантическому полю, со своей фреймовой структурой и широким набором значений. Связь с полисемией прилагательного *острый* таких слов, как *своевременный* или *нежный* и вовсе установить довольно сложно.

Таким образом, метод составления списка прилагательных, относящихся к изучаемому полю, путем анализа синонимов опорного слова связан с двумя проблемами: с

одной стороны, лексем, покрывающих зону физических фреймов поля, набирается слишком мало, с другой стороны, синонимов для метафорических значений, напротив, чересчур много.

2.1.3. Метод обратных переводов

Метод обратных переводов подразумевает выход за рамки русского языка и имитирует, хотя и в сильно упрощенном виде, ручную работу лексического типолога. Этот механизм так же, как и предыдущий, основан на словарном материале, но источником данных уже являются не словари синонимов, а двуязычные переводные словари.

Мы опробовали два варианта этой методологии. В первом случае алгоритм составления списка прилагательных был следующим: для опорной лексемы *острый* извлекались все варианты однословных переводов её прямых значений на английский, немецкий и французский языки, которые затем переводились обратно на русский[14]. Такая процедура наиболее точно воспроизводит челночный метод сбора прилагательных: в других языках находятся слова, способные покрывать фреймы ядерных (прямых[15]) значений изучаемого поля, а затем для всех групп их употреблений, в том числе метафорических, находится русский аналог для анкеты. Результирующий набор русских лексем, как и список синонимов, был отфильтрован по частотности: в итоговый перечень вошли только те прилагательные, которые попадают в список из 10 000 наиболее частотных знаменательных слов по основному подкорпусу НКРЯ. Полученный набор представлен в таблице 6.

Из таблицы видно, что метод обратных переводов позволяет сформировать широкий набор прилагательных, в который попадает немного больше обозначений физических свойств объектов, чем в списки предыдущих версий, однако лексем, описывающих абстрактные характеристики объектов, по-прежнему очень много. Особенно это касается прилагательных, обозначающих качества человека (см. сноску 13), которые мы для наглядности выделили в отдельный столбик таблицы.

---

[14] В качестве источника для перевода использовалась коллекция электронных словарей Яндекс (slovari.yandex.ru). Этот ресурс группирует переводные эквиваленты по значениям лексемы, для которой ищется перевод (в отличие, например, от словаря «Мультитран» - www.multitran.ru), что позволяет хоть и вручную, но достаточно быстро определить, какие переводы следует учесть, а какие исключить.

[15] В данном случае множество прямых значений не совпадает с множеством физических употреблений. С целью наиболее точно сымитировать традиционную методологию, мы опирались на решение, принятое на первом этапе исследования признакового поля ‘острый’, проведенного вручную на материале русского языка, согласно которому прямыми значениями считались только три: ‘режущий (об инструменте с режущим краем)’, ‘колющий (об инструменте с колющим концом)’ и ‘вытянутый (о форме объекта)’. Остальные физические значения (‘острый (о вкусе)’, ‘высокий, пронзительный (о звуке)’ и некоторые другие) трактовались как метафоры (подробнее см. Кюсева 2012).

Физические признаки	**Абстрактные признаки (не качества человека)**	**Качества человека**
Острый	Резкий	Остроумный
Колючий	Пронзительный	Умный
Крутой	Быстрый	Наблюдательный
Едкий	Точный	Внимательный
Отчетливый	Яркий	Решительный
Высокий	Обидный	Суровый
	Крепкий	Жестокий
	Неприятный	Хитрый
	Сильный	Коварный
	Определенный	Смелый
	Направленный	Энергичный
	Боевой	Ловкий
	Тонкий	Искусный
	Глубокий	Строгий
		Ожесточенный
		Злой

Таблица 6. Список прилагательных, полученный методом обратных переводов (вариант 1: *острый* => переводы прямых значений на английский, немецкий и французский => переводы всех значений на русский).

Поскольку этот список содержал максимальное количество слов, которые казались нам полезными, мы предприняли попытку провести его дополнительную фильтрацию с целью сократить количество абстрактных признаков. Для этого мы посчитали близость (по косинусной мере) вектора каждого из прилагательных последнего набора к вектору лексемы *острый*[16], в надежде, что слова, являющиеся переводными эквивалентами лишь для некоторых переносных употреблений опорного прилагательного, покажут низкое значение косинуса. Однако, как видно из таблицы 7, физические признаки, которые нам хотелось бы сохранить, также демонстрируют низкий уровень близости к лексеме *острый* в векторном пространстве. Что, учитывая методику их сбора, неудивительно: преимущество метода обратных переводов как раз в том, что он позволяет охватить те фрагменты изучаемого поля, которые в русском языке не покрываются центральным, доминантным прилагательным.

[16] В пространстве векторов сочетаемости, собранных на основе узкого окна ±1 знаменательное слово.

Прилагательное	**Значение косинуса**
Наблюдательный	0,050
Боевой	0,071
Колючий	0,084
Ожесточенный	0,158
Крутой	0,177
Направленный	0,180
Едкий	0,198
Решительный	0,213
Коварный	0,230
Пронзительный	0,238
Яркий	0,244
Искусный	0,250
Быстрый	0,252
Определенный	0,255
Умный	0,269
Ловкий	0,273
Точный	0,283
Обидный	0,284
Строгий	0,285
Отчетливый	0,286
Остроумный	0,297
Злой	0,298
Энергичный	0,302
Крепкий	0,313
Резкий	0,325
Высокий	0,328
Смелый	0,329
Жестокий	0,341
Суровый	0,347
Глубокий	0,352
Хитрый	0,353
Внимательный	0,398
Тонкий	0,399
Сильный	0,412
Неприятный	0,440
Острый	1,000

Таблица 7. Близость векторов прилагательных, полученных методом обратных переводов, к вектору лексемы *острый* (косинус).

Получив недостаточно удовлетворительный результат, мы модифицировали алгоритм сбора релевантных лексем по методу обратных переводов. Изменения были направлены на уменьшение количества абстрактных признаков, имеющих лишь опосредованное отношение к изучаемому семантическому полю. Новая версия алгоритма устроена следующим образом: в русско-английском, русско-французском и русско-

немецком словарях ищутся переводные эквиваленты для физических значений лексемы *острый*; затем полученные прилагательные переводятся обратно на русский язык, причём в конечный список попадают только слова, которые соответствуют физическим значениям английских, французских и немецких лексем. Таким образом, новая процедура отличается от предыдущей по двум параметрам:

1) в качестве отправной точки используются все физические значения прилагательного *острый*, а не только те, которые были признаны прямыми в рамках традиционного лексико-типологического анализа данного семантического поля;
2) абстрактные значения отсекаются на этапах как прямого, так и обратного перевода.

Первое нововведение позволяет отвлечься от уже проведенного исследования и рассматривать данные более объективно, избавившись от необходимости решать сложную проблему разграничения прямых и переносных значений в области физических признаков (тем более, что для семантического поля, не подвергавшегося ранее лексико-типологическому анализу, это сделать было бы крайне затруднительно). Второе изменение гарантирует отсутствие лексем, заведомо не относящихся к рассматриваемому полю.

Эта версия метода обратных переводов менее точно имитирует процесс ручного сбора списка релевантных прилагательных, поскольку не учитывает новые модели метафорических сдвигов, которые, возможно, продемонстрировали бы в дальнейшем типологическую регулярность. Однако полученный результат существенно превосходит предыдущие с точки зрения отсутствия большого количества «лишних» контекстов (см. таблицу 8). К тому же, вполне возможно, что, по крайней мере, часть неучтенных метафор будет компенсирована на этапе анализа сочетаемости набранных прилагательных: новые физические признаки могут развивать переносные значения, не свойственные русской лексеме *острый*, но в целом обычные для слов данного поля в языках мира.

Быстрый	Жаркий	Пронзительный
Высокий	Колючий	Резкий
Горячий	Крутой	Яркий
Душистый	Острый	
Едкий	Отчетливый	

Таблица 8. Список прилагательных, полученный методом обратных переводов (вариант 2: *острый* => переводы физических значений на английский, немецкий и французский => переводы физических значений на русский).

Аналогичную модификацию можно было бы применить и к методу сбора элементов семантического поля путем анализа синонимов. Так, например, за основу можно было бы взять традиционные бумажные словари (например, словарь З.Е. Александровой), извлекать из них прилагательные, синонимичные физическим значениям опорного слова, и смотреть, в свою очередь, синонимы для их физических употреблений. Преимущество этого метода заключалось бы в том, что он, в отличие от метода обратных переводов, не выходил бы за рамки одного языка. С другой стороны, у нас нет уверенности в том, что семантические поля качественных признаков могут состоять только из квазисинонимов. Синонимичными считаются лексемы, которые употребляются в сочетании с одними и теми же словами, но передают при этом разные, хотя и близкие значения. Теоретически признаковые поля, напротив, должны быть составлены из элементов, находящихся между собой в отношении дополнительного распределения: в контексте одних существительных употребляется одно прилагательное, в контексте других – другое, как, например, во французском языке, где инструменты с колющим концом описываются прилагательным *aigu*, а объекты вытянутой формы – лексемой *pointu*. Однако на практике дополнительное распределение редко бывает математически точным: в том же французском языке прилагательное *pointu* может распространяться в том числе и на фреймы, находящиеся в сфере действия лексемы *aigu* (ср. *les instruments pointus et tranchants* – ‘колющие и режущие инструменты’), благодаря чему эти слова в некоторых контекстах оказываются синонимами.

Все дальнейшие эксперименты в рамках настоящего исследования проводятся на материале прилагательных из последнего списка, полученного с помощью модифицированного метода обратных переводов. Вопрос о том, может ли семантическое поле качественного признака состоять только из квазисинонимов, мы оставляем открытым.

2.2. Составление списков коллокаций

Следующий шаг после составления списка прилагательных, относящихся к изучаемому семантическому полю, - анализ контекстов, в которых эти прилагательные употребляются. Как уже было сказано во Введении, мы считаем, что для признаковых слов ключевым является единственный участник вводимой ими ситуации, т.е. носитель свойства. А это означает, что для анализа сочетаемости прилагательного достаточно двухсловных словосочетаний вида «прилагательное + существительное».

Автоматически составить перечень таких словосочетаний для заданной признаковой лексемы на материале русского языка можно множеством способов. Во-первых, благодаря тому, что у определительной конструкции всего две переменных, можно воспользоваться списками биграммов Google[17] или НКРЯ. Биграммы Google разбиты на файлы и отсортированы по алфавиту по первому слову, поэтому извлечение из этих списков всех словосочетаний с искомой признаковой лексемой для русского языка не составляет проблем, в отличие, например, от французского, где прилагательное почти всегда занимает позицию после существительного.

Однако список коллокаций, получаемый таким способом, требует дополнительной обработки. Во-первых, поскольку нас интересует сочетаемость лексем, а не словоформ, все словосочетания необходимо привести к начальному виду. Эта задача требует привлечения морфологического парсера и последующей дизамбигуации, что приводит к появлению достаточно большой порции «мусора» – неправильных разборов. Во-вторых, список биграммов всегда очень обширен и нуждается в фильтрации. Между тем, основания для очистки неочевидны: биграммы Google собираются на основе несбалансированного и никак не контролируемого корпуса текстов, поэтому определить априорно, какие словосочетания следует отсечь, а какие оставить, – непростая задача. Биграммы НКРЯ в этом отношении намного чище и представительнее, однако и они нуждаются в предобработке в виде морфологического анализа.

Морфологического парсинга можно избежать, воспользовавшись ресурсом, в котором необходимый разбор уже проведен. Например, коллекция доступных онлайн корпусов Sketch Engine (www.sketchengine.co.uk, см. Kilgarriff et al. 2004) содержит русский подкорпус объемом в несколько миллиардов словоупотреблений. С помощью нажатия одной-двух кнопок этот ресурс позволяет получить список всех сочетаний искомой лексемы с существительными, приведенных к начальной форме и отсортированных по частотности. Однако, поскольку эта статистическая информация

[17] http://storage.googleapis.com/books/ngrams/books/datasetsv2.html

строится на основе очень объемного корпуса интернет-текстов, результирующий список также требует дополнительной фильтрации. Так же, как и в случае с биграммами Google, отобрать просто определенное количество наиболее частотных словосочетаний нельзя: статистика собирается по несбалансированному корпусу. Более действенный метод – оставлять, например, только такие существительные, которые встретились в сочетании с искомой лексемой в первой тысяче примеров случайной выдачи НКРЯ по запросу «искомое прилагательное + любое имя существительное». В любом случае, этот метод составления списков коллокаций связан с большим количеством дополнительных махинаций, тем более, что бесплатный доступ к коллекции корпусов Sketch Engine ограничен (предоставляется только на 30 дней).

Самый надежный и удобный в наших условиях метод анализа сочетаемости прилагательных – сбор коллокаций по имеющимся в нашем распоряжении корпусам текстов: основному и газетному подкорпусам НКРЯ и корпусу РУВАК. Эти тексты уже снабжены морфологической разметкой, поскольку на этом же материале проводились эксперименты в рамках исследования Кюсева 2014[18], так что достаточно просто посчитать, в сочетании с какими существительными и сколько раз в рамках каждого корпуса встретилась рассматриваемая лексема.

Мы провели соответствующие эксперименты для опорного прилагательного *острый*. Гипотеза о том, что жанр и степень сбалансированности корпусов могут значимо повлиять на результирующий список коллокаций, подтвердилась и на этом материале. Так, набор существительных, сочетающихся со словом *острый* в рамках газетного подкорпуса НКРЯ, оказался явно перекошен в сторону спортивной тематики: одними из самых частотных оказались словосочетания *острая атака*, *острая контратака*, *острая передача*, *острый пас*, в то время как более естественные и ядерные для этого признака употребления (такие, как *острая иголка*) показывают сравнительно низкий уровень частотности (см. таблицу 9).

Существительное, которое встретилось в сочетании с прилагательным *острый*	**Частотность словосочетания (абсолютное значение)**	**Ранг[19]**
Проблема	368	1
Угол	354	2
Вопрос	307	3
Ощущение	226	4
Необходимость	213	5
Дефицит	213	6
Нехватка	211	7

[18] Подробнее о параметрах обработки корпусов см. раздел 2.3 настоящей работы.

[19] Для простоты ранги соответствуют сплошной нумерации коллокаций и не учитывают случаев совпадения уровня частотности.

Момент	174	8
Борьба	101	9
Фаза	94	10
Атака	92	11
Конкуренция	88	12
Боль	82	13
Дискуссия	79	14
Тема	73	15
Ситуация	70	16
Нож	67	17
Потребность	66	18
Отравление	62	19
Предмет	57	20
Контратака	56	21
Форма	55	22
Перец	51	23
Передача	48	24
Реакция	45	25
Конфликт	44	26
Кризис	43	27
Желание	39	28
Пас	37	29
…	…	
Иголка	3	203
Игла	3	205

Таблица 9. Фрагмент списка существительных, с которыми сочетается прилагательное *острый* (по газетному подкорпусу НКРЯ).

Аналогичная статистика по корпусу РУВАК дает интуитивно более приемлемые результаты, хотя и здесь, как кажется, сказывается недостаточная сбалансированность текстового материала: например, очень частотны и многочисленны названия разного рода заболеваний, некоторые из которых известны далеко не каждому носителю русского языка (см. таблицу 10).

Существительное, которое встретилось в сочетании с прилагательным *острый*	**Частотность словосочетания (абсолютное значение)**	**Ранг**
Лейкоз	729	13
Пиелонефрит	443	32
Инфаркт	435	34
Отравление	432	35
Воспаление	350	40
Бронхит	298	50
Аппендицит	298	51
Панкреатит	278	54
Гастрит	220	64
Инфекция	209	67
Холецистит	138	82
Гепатит	131	87
Ринит	131	88
…		

Таблица 10. Фрагмент списка существительных, с которыми сочетается прилагательное *острый* (по корпусу РУВАК).

Наконец, список коллокаций, полученный по основному подкорпусу НКРЯ, выглядит наиболее представительно (см. таблицу 11). Чтобы не опираться исключительно на интуитивные суждения, мы оценили его более строго. Для каждого существительного, встретившегося в корпусе в сочетании с прилагательным *острый* не менее 10 раз, мы указали фрейм, который это словосочетание иллюстрирует, и затем сверили полученный набор фреймов с собранной вручную анкетой. Оказалось, что в списке коллокаций фигурируют представители всех фреймов, релевантных для русского прилагательного *острый*. Словосочетания, найденные автоматически, не совпадают в точности с примерами анкеты (например, в анкете фрейм ‘режущий (об инструментах с режущим краем)’ иллюстрируется контекстом ‘острый меч’, а в списке коллокаций – *острое лезвие*), но это расхождение не снижает значимости полученного списка. Напротив, для анкеты, составленной вручную, примеры подбирались исследователями интуитивно, в то время как автоматический сбор данных позволяет выбирать иллюстрации на более строгих основаниях: например, по принципу частотности. Тот же контекст ‘острый меч’ на деле не всегда оказывался удачным: меч – это устаревшая реалия, знакомая, к тому, носителям далеко не всех культур, поэтому эта строка анкета часто оставалась незаполненной (например, для коми-зырянского и кабардинского языков).

Существительное, которое встретилось в сочетании с прилагательным *острый*	**Частотность словосочетания (абсолютное значение)**	**Ранг**
Угол	398	1
Боль	290	2
Ощущение	198	3
Нож	197	4
Глаз	196	5
Нос	178	6
Вопрос	166	7
Взгляд	164	8
Чувство	162	9
Форма	156	10
Проблема	152	11
Зуб	150	12
Камень	135	13
Запах	118	14
Конец	101	15
Ум	100	16
Слово	98	17
Необходимость	95	18
Подбородок	91	19
Период	87	20
Край	86	21
Желание	86	22

Характер	79	23
Дефицит	79	24
Момент	78	25
Бородка	74	26
Нехватка	73	27
Жалость	69	28
Потребность	68	29
Колено	66	30
Вершина	64	31
Приступ	63	32
Нужда	62	33
Игла	61	34
Шип	59	35

Таблица 11. Начальный фрагмент списка существительных, с которыми сочетается прилагательное *острый* (по основному подкорпусу НКРЯ). Видно, что уже среди первых 35 примеров есть представители основных физических значений опорной лексемы, ср.: *нос*, *подбородок*, *бородка* – фрейм ‘вытянутая форма’; *нож* – фрейм ‘режущий (об объектах с режущим краем)’; *игла*, *шип* – ‘колющий (об объектах с колющим концом)’ и т.д.

Таким образом, результаты наших наблюдений позволяют использовать в качестве текстового материала для автоматического составления списка коллокаций основной подкорпус НКРЯ. Это решение подкрепляется также данными экспериментов, представленных в работе Кюсева 2014. В рамках исследования М.В. Кюсевой оценивалась корреляция автоматически вычисленной степени сходства слов и словосочетаний в русском языке с типологической близостью иллюстрируемых ими ситуаций (подробнее см. Кюсева 2014, а также введение к настоящей работе). Так, было установлено, что для полного представления сочетаемости словосочетания в целом (например, словосочетания *острый нож* как единой языковой единицы) объема основного подкорпуса НКРЯ недостаточно: уровень корреляции русскоязычных данных с результатами типологических наблюдений растет по мере увеличения объема корпуса. Для существительных, напротив, статистика, собранная на основе НКРЯ, оказывается вполне представительной: значение корреляции, полученное в результате сравнения данных НКРЯ и типологических наблюдений, не меняется в процессе добавления новых текстовых данных (см. таблицу 12).

Корпус	**Языковая единица**	**Коэффициент корреляции Пирсона**
Основной подкорпус НКРЯ	словосочетание	0,308
Основной подкорпус НКРЯ + РУВАК	словосочетание	0,361
Основной подкорпус НКРЯ + РУВАК + газетный подкорпус НКРЯ	словосочетание	0,369
Основной подкорпус НКРЯ	слово	0,317
Основной подкорпус НКРЯ + РУВАК	слово	0,318
Основной подкорпус НКРЯ + РУВАК + газетный подкорпус НКРЯ	Слово	0,317

Таблица 12. Зависимость коэффициента корреляции Пирсона от объёма корпуса

Итак, составление списка коллокаций на материале только основного подкорпуса НКРЯ (без привлечения других текстовых ресурсов) позволяет экономить время без ущерба качеству результатов. И всё же полученный таким способом список нуждается в дальнейшей модификации, а именно в установке порога частотности: необходимо определить, сколько раз словосочетание должно встретиться в корпусе, для того чтобы его можно было включить в исследование.

Такой порог необходим по двум причинам. Во-первых, существительных, сочетающихся с искомым прилагательным, слишком много. Во-вторых, они распределены в соответствии с законом Ципфа: примерно 20% списка занимают частотные сочетания, а остальные 80% – слова, встретившиеся в корпусе в непосредственной близости от интересующей нас лексемы один или два раза, т.е. большая часть существительных иллюстрирует лишь окказиональные употребления изучаемого прилагательного.

Мы провели ряд экспериментов с разными признаковыми словами, чтобы определить точнее, где именно можно установить порог частотности. Мы разметили списки коллокаций для лексем *острый, тяжелый, хороший, шершавый, грубый* (для которых у нас уже разработаны типологические анкеты) и отметили, в какой момент происходит «насыщение», т.е. до какого элемента списка нужно дойти, чтобы набрать хотя бы по две иллюстрации на каждый фрейм, релевантный для данного прилагательного.

Оказалось, что уровень насыщения зависит от двух факторов: степени многозначности и частотности самой признаковой лексемы (см. таблицу 13). С одной стороны, чем больше у лексемы значений, тем больше контекстов требуется для их иллюстрации и тем позднее должен наступать момент насыщения. С другой стороны, чем частотнее лексема, тем выше можно поставить планку: если слово употребляется в целом очень часто, то есть вероятность, что оно будет нередко встречаться и в сочетании некоторого круга других лексем. Так, например, для признака *острый* достаточно уровня частотности, равного 17, в то время как прилагательному *тяжелый*, обладающему более широкой полисемией, требуется уровень не ниже 10. А для адекватного описания лексемы *хороший*, очень частотной и не очень многозначной, достаточно принять во внимание только те существительные, в сочетании с которыми она встретилась около 100 раз.

Однако важно понимать, что параметры, влияющие на уровень насыщаемости лексемы, тоже взаимосвязаны, причём нетривиальным образом. Так, во многих случаях у многозначных слов высокая частотность (ср. лексемы *шершавый, грубый, острый* и *тяжелый*: по мере роста количества значений от слова к слову растёт и уровень частотности), а эффекты на момент насыщения эти конфигурации параметров оказывают

противоположные: многозначность понижает планку, а частотность – повышает. Помимо этого, такая связь параметров не симметрична: встречаются лексемы, частотность которых очень велика, но при этом у них выделяется мало значений (см. *хороший*).

Наконец, основная сложность заключается в том, что мы можем лишь выявить закономерности, определить тенденции: из-за одних факторов порог повышается, из-за других понижается. Вычислить точную формулу зависимости порога частотности коллокаций от частотности и многозначности опорной лексемы пока не удаётся: слишком велик разброс данных, поэтому любое решение связано с риском упустить какой-нибудь нюанс функционирования данной конкретной лексемы.

Исходя из всего вышесказанного, мы установили фиксированное значение частотности, общее для всех слов и не зависящее ни от каких параметров. При этом мы выбрали достаточно низкую планку, позволяющую набрать достаточное количество примеров, в то же время гарантируя отсутствие в итоговом списке заведомо окказиональных употреблений рассматриваемой лексемы. Таким образом, мы отказываемся от эвристик, которые могли бы нам позволить избавиться от «лишних» контекстов (которых, например, для слова *хороший* при таком решении будет множество), в пользу максимальной полноты данных. В качестве порогового мы выбираем значение 10 и считаем, что существительное, встретившееся в корпусе в сочетании с опорной лексемой меньше этого количества раз, не является его достаточно устойчивой коллокацией.

Лексема	**Количество значений по БТС**[20]	**Общее количество употреблений в корпусе**	**Необходимый уровень частотности по списку коллокаций**
Шершавый	2	Не входит в 10 000 частотных знаменательных слов	19
Грубый	7	12096	18
Острый	10	15929	17
Тяжелый	18	40536	10
Хороший	5[21]	92086	101

Таблица 13. Зависимость порога частотности коллокаций от многозначности и частотности опорной лексемы.

[20] Большой толковый словарь под редакцией С.А. Кузнецова (Кузнецов 1998)

[21] В Большом толковом словаре для прилагательного *хороший* указано больше пяти значений, но многие из них не диагностируются на уровне словосочетания «прилагательное + существительное», ср.: *Ну ты хорош! Хорош ругаться!* Или: *Ты моя хорошая!*

Таким образом, мы получили экспериментальный материал, состоящий в общей сложности из 13 прилагательных и 1818 примеров их употреблений (словосочетаний вида «прилагательное + существительное»). Очевидно, что для типологической анкеты этот объем неприемлем: требуется кластеризация словосочетаний и выбор нескольких представителей из каждого кластера.

2.3. Подготовка векторного пространства

Необходимые для нашего исследования вектора сочетаемости были собраны на основе текстового материала трех корпусов: основного подкорпуса НКРЯ, газетного подкорпуса НКРЯ и корпуса РУВАК. Обе части НКРЯ были нам доступны только в виде простых наборов неразмеченных текстов, а РУВАК – в размеченном виде, в частности, с морфологической разметкой, которая была для нас особенно существенна. Потребность в морфологической аннотации была связана с двумя причинами: во-первых, нас интересовала сочетаемость лемм, а не словоформ; во-вторых, в качестве измерений мы предполагали использовать лексемы только знаменательных частей речи. Поэтому подкорпуса НКРЯ мы разметили с помощью морфологического парсера MyStem и самостоятельно дизамбигуировали, воспользовавшись системой снятия морфологической омонимии, сконструированной нами в рамках научно-исследовательского семинара факультета филологии НИУ ВШЭ «Проектирование лингвистических ресурсов и систем» (см. Лакомкин и др. 2013).

Далее на базе основного подкорпуса НКРЯ был собран частотный словарь из 10 000 лексем знаменательных частей речи (глаголов, существительных, прилагательных и наречий), которые и использовались в дальнейшем в качестве набора измерений векторных пространств. Значением по каждой из 10 000 осей было число раз, когда слово-измерение встретилось в корпусе на расстоянии не более ±5 знаменательных слов от опорной единицы.

Так как основная наша задача – кластеризация контекстов употребления качественных прилагательных, т.е. оценка степени сходства между различными словосочетаниями вида «прилагательное + существительное», в качестве опорных единиц для векторов мы использовали два вида сущностей: целые словосочетания и отдельные слова (для последующей композиции).

Как уже было сказано выше, для полного представления сочетаемости словосочетаний нужны очень объемные корпуса текстов. В работе Кюсева 2014 продемонстрировано, что даже имеющихся у нас трёх корпусов недостаточно. Так, из приводившейся выше таблицы 12 видно, что уровень корреляции данных русского языка и типологических наблюдений растёт по мере увеличения объема корпуса, однако всё равно остаётся на сравнительно низком уровне. При этом если рассматривать не словосочетания в чистом виде, а соответствующие им композиции двух векторов (прилагательного и существительного), то значение корреляции резко подскакивает до

0,7. Это наблюдение даёт нам основание считать, что метод композиции не искажает информацию, а, напротив, позволяет компенсировать недостаток данных.

Безусловно, этот результат требует проверки на материале других качественных признаков, однако он правдоподобен и теоретически объясним. Так, известно, что значение прилагательного напрямую зависит от семантики существительного, с которым оно сочетается (см., например, Апресян 1974, Рахилина 2008). Из этого следует, что чем больше похожи друг на друга существительные, т.е. чем ближе они друг к другу в векторном пространстве, тем больше вероятности, что они будут описываться одним и тем же прилагательным изучаемого признакового поля в русском языке (или их переводы – одним и тем же признаковым словом соответствующего языка). Именно это и демонстрирует вторая часть таблицы 12, в которой приведены значения уровня корреляции дистрибуции русских существительных с одной стороны и закономерностей употребления прилагательных со значением ‘острый’ в контексте существительных с аналогичной семантикой в языках мира – с другой. Таким образом, многое о словосочетании «прилагательное + существительное» можно сказать, исходя из дистрибуции одного существительного. Многое, но не всё. Композиция, т.е. добавление к вектору существительного вектора прилагательного, позволяет модифицировать значения измерений, усилив одни оси и ослабив другие, в зависимости от особенностей сочетаемости признаковой лексемы. Этот процесс можно сравнить с механизмом формирования гласных звуков: поток воздуха, поступающий из легких и уже получивший некоторую периодичность от голосовых связок (= вектор существительного) попадает в ротовую и/или носовую полость, где, в зависимости от конфигурации речевого тракта (= вектора прилагательного), одни его частоты усиливаются, а другие ослабляются. Отдельная задача – подобрать такую передаточную функцию (= такую формулу зависимости результирующих значений измерений от значений соответствующих осей векторов прилагательного и существительного), которая будет максимально репрезентативно отражать свойства целого словосочетания как единой языковой единицы[22]. В работе Кюсева 2014 показано, что простая сумма векторов дает хорошие результаты, поэтому в рамках нашего исследования мы будем использовать такой же вариант композиции, однако вопрос подбора оптимальной формулы требует дальнейшего изучения.

Следует отметить, что метод композиции, как и ожидалось, позволяет существенно сокращать время расчетов. Во-первых, в отличие от словосочетаний в чистом виде, для каждого из которых нужно собирать отдельный вектор, для композиции достаточно один

[22] Подробнее о методе композиции см. Mitchell&Lapata 2010

раз собрать вектора для всех необходимых прилагательных (которых всего 13, см. раздел 2.1) и для нескольких сотен частотных существительных, впоследствии лишь «добирая» вектора менее частотных слов по мере необходимости. Так, например, для готовых словосочетаний *яркий свет, резкий свет, острое слово, резкое слово, яркий ответ, резкий ответ, яркий человек, резкий человек, острый взгляд, резкий взгляд* понадобится 10 итераций вычисления векторов, а для построения этих же словосочетаний методом композиции – 8, причем эта разница будет резко возрастать по мере увеличения количества словосочетаний. Во-вторых, поскольку для адекватного представления сочетаемости отдельных слов, в отличие от целых словосочетаний, достаточно одного основного подкорпуса НКРЯ, статистику для композиции можно считать только по нему, без привлечения РУВАКа и газетного подкорпуса, что сокращает количество измерений ещё по крайней мере втрое.

Таким образом, в рамках нашего исследования мы проводим две серии экспериментов, основанные на разных типах пространств. Одно пространство состоит из векторов словосочетаний, вычисляемых на основе трех корпусов, а другое – из аналогичных словосочетаний, построенных на основе суммы векторов прилагательных и существительных, собираемых по основному подкорпусу НКРЯ. В обоих случаях в качестве измерений изначально выступают 10 000 частотных знаменательных слов, затем размерность пространств уменьшается до 300 (подробнее о механизме уменьшения размерности см. раздел 1.2). Значения измерений вычисляются на основе окна в ±5 лексем знаменательных частей речи. Все модификации пространств и операции над векторами мы проводим с помощью тулкита DISSECT (**DIS**tributional **SE**mantics **C**omposition **T**oolkit, http://clic.cimec.unitn.it/composes/toolkit/), являющегося частью проекта COMPOSES (**COMP**ositional **O**perations in **SE**mantic **S**pace, http://clic.cimec.unitn.it/composes/).

2.4. Кластеризация и составление анкеты

Все методы кластеризации можно разделить на два типа:

1.Алгоритмы, требующие указания числа кластеров, на которые следует разбить все данные;

2.Алгоритмы, определяющие количество кластеров автоматически.

Для решения нашей задачи удобнее воспользоваться методом, который вычислял бы число кластеров автоматически: мы предполагаем, что исследователь изначально не знает, сколько фреймов будет в его анкете. Исходя из этих соображений, мы провели[23] ряд пробных кластеризаций семантических векторов по трём алгоритмам, не требующим указания числа кластеров: Affinity Propagation (см., например, Frey and Dueck 2007), Mean-shift (см. Comaniciu and Meer 2002) и DBScan (см. Ester et al. 1996). Результаты работы всех трех алгоритмов были примерно одинаковыми: они выделяли достаточно большое количество кластеров (около 150 на 1818 словосочетаний), однако среди сформированных групп была одна очень большая и много очень маленьких (в основном единичных). Изменения параметров кластеризации позволяли варьировать число кластеров, однако более дробное деление получалось за счет отщепления от доминанты новых единичных элементов.

Такой тип разбиения данных в нашем случае мало информативен (мы стремимся получить сравнимые по размеру группы, чтобы затем выбрать из них наиболее ярких представителей для анкеты), поэтому мы приняли решение провести серию экспериментов с алгоритмами, требующими изначального указания числа кластеров. Количество кластеров мы определили следующим образом: посчитали сумму числа значений всех прилагательных нашей выборки (по Малому академическому словарю) и умножили её на два, получив тем самым число 112 (см. формулу на схеме 1). Удваивание суммы значений делает общее количество кластеров более независимым от одного конкретного словаря: во-первых, наш опыт показывает, что фреймы часто оказываются более дробными, чем словарные значения; а во-вторых, надежнее получить заведомо большее количество кластеров и выбросить лишнее при последующей обработке.

[9 (*острый*) + 6 (*резкий*) + 2 (*быстрый*) + 1 (*душистый*) + 6 (*горячий*) + 5 (*яркий*) + 2 (*едкий*) + 4 (*колючий*) + 5 (*крутой*) + 2 (*отчетливый*) + 3 (*пронзительный*) + 7 (*высокий*) + 4 (*жаркий*)] * 2 = **112**

Схема 1. Вычисление количества кластеров на основе словарных данных

[23] С помощью тулкита scikit-learn (http://scikit-learn.org/stable/index.html).

Для кластеризации с помощью алгоритмов без автоматического определения количества кластеров мы использовали пакет программ Cluto[24]. Этот тулкит предлагает несколько методов кластерного анализа. Их названия, наборы параметров и основные принципы работы представлены в таблице 14.

Алгоритм	**Параметры**	**Принцип работы**
rb (repeated bisections)	• Критерий кластеризации; • критерий выбора кластера, который будет разбиваться на части следующим	Всё пространство объектов делится на две части, потом одна из частей делится ещё на две, ещё одна из частей делится ещё на две и т.д. до тех пор, пока не будет получено нужное число кластеров.
Rbr	• Функция оптимизации; • критерий выбора кластера, который будет разбиваться на части следующим	Берет результат кластеризации по методу rb и оптимизирует его.
Direct	Критерий кластеризации	Пространство объектов сразу разделяется на нужное количество классов
Agglo	Функция оптимизации	В основе кластеризации – процесс оптимизации некоторой функции
Graph	• Мера близости между объектами (косинус угла, Евклидово расстояние, коэффициент корреляции, коэффициент Жаккара); • количество ближайших соседей; • критерий выбора кластера, который будет разбиваться на части следующим	Из пространства объектов строится граф, который затем разбивается на нужное число фрагментов
Bagglo	Функция оптимизации	Конкатенация алгоритмов rb и agglo: сначала применяется метод rb, затем увеличивается размерность пространства и необходимый результат достигается с помощью алгоритма agglo

Таблица 14. Используемые нами алгоритмы кластеризации с заданным числом кластеров.

В ходе настоящего исследования мы провели кластеризацию наших данных, используя все методы, перечисленные в таблице 14, со всеми допустимыми комбинациями параметров. Помимо этого, любой из алгоритмов (кроме agglo и bagglo) можно модифицировать с помощью критериев agglofrom и agglocrfun. Метод agglofrom указывает, на какое число кластеров (большее, чем требуется в конечном итоге) алгоритм должен разбить исходное пространство объектов; с помощью параметра agglocrfun задается функция, путем оптимизации которой кластеры будут объединяться до тех пор, пока не будет получено нужное число элементов. Иными словами, сначала происходит

[24] http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview

избыточная кластеризация на основе некоторого метода, а затем «лишние» кластеры упраздняются с помощью метода agglo. Мы проводили эксперименты с применением в том числе и данных параметров, указывая в качестве значения agglofrom 168 – количество значений рассматриваемых нами прилагательных по МАС, умноженное на три.

В результате мы получили 168 различных конфигураций параметров кластеризации и провели две серии экспериментов: применили все 168 алгоритмов к каждому из наших двух векторных пространств (пространству словосочетаний и пространству композиций).

Следующий шаг – собственно составление анкеты. Для того, чтобы человек мог пользоваться автоматически составленной анкетой-опросником, необходимо прежде всего сократить количество входящих в неё словосочетаний. Мы делаем это двумя путями: во-первых, отбрасываем слишком маленькие кластеры, состоящие из одного-двух элементов, а во-вторых, сокращаем размеры всех остальных кластеров до трех словосочетаний.

Маленькие кластеры мы исключаем из итоговой анкеты на том основании, что их изоляция не даёт пользователю представления о структуре семантического поля. Мы исходим из предположения, что контексты употребления любого прилагательного можно разделить на классы (потенциальные типологически релевантные фреймы), причем эти классы будут достаточно устойчивыми паттернами, включающими в себя не один и не два примера. В таком случае, в изоляции могут оказаться либо недостаточно представительные контексты, либо примеры фразеологических сочетаний (которые, на наш взгляд, крайне редки), либо окказиональные употребления рассматриваемого прилагательного, либо вообще невозможные сочетания, явившиеся результатом ошибки дизамбигуатора или морфологического парсера.

Сокращение всех кластеров до размера 3, напротив, кажется нам продуктивным. С одной стороны, три примера не перегрузят анкету. С другой стороны, вполне вероятно, что и из этих трех в дальнейшем останется только два. Три элемента – это то минимальное количество, которое позволяет исследователю увидеть систему, определить основание объединения сочетаний в один кластер и, возможно, удалить один из его элементов уже вручную. Так, например, по кластеру [*острая болезнь*, *острый кризис*, *острая паранойя*] понятно, что конституирующим в нём может считаться значение высокой степени проявления заболевания, и словосочетание *острый кризис*, в таком случае, можно из него устранить как неоднозначное (*острый кризис* может обозначать как пик болезни, так и ситуацию, требующую немедленного разрешения, ср. *острая нехватка*, *острый недостаток*). Как бы то ни было, мы считаем, что основная задача составленной автоматически анкеты – указать исследователю на некоторые возможные закономерности организации поля, выделить основные релевантные для него противопоставления. Для

достижения этой цели необходимо иметь прежде всего результаты более или менее устойчивых объединений, а не маленькие кластеры, являющие собой разные менее надежные примеры употребления изучаемых прилагательных и не демонстрирующие никакой системы.

Из кластеров, размеры которых превышали порог в три элемента, мы извлекали представителей двумя способами. Первый способ простой – методом случайного выбора. Второй вычислительно и содержательно сложнее: он подразумевает определение центра кластера и выбор трех наиболее близких к нему векторов. Для реализации второго метода для каждого кластера вычислялся эталонный «центральный вектор», значения измерений которого представляли собой среднее арифметическое значений всех элементов кластера. Затем определялись три вектора, которые ближе всего (по косинусной мере близости) к эталону. Таким образом, мы получили четыре набора анкет, по 168 вариантов в каждом:

1. анкеты на основе кластеризации пространства словосочетаний, случайный выбор представителей кластеров;

2. анкеты на основе кластеризации пространства словосочетаний, представители кластеров – три элемента, находящихся ближе всего к центру кластера;

3. анкеты на основе кластеризации пространства композиций, случайный выбор представителей кластеров;

4. анкеты на основе кластеризации пространства композиций, представители кластеров – три элемента, находящихся ближе всего к центру кластера.

2.5. Оценка результатов

В результате кластеризации и её последующей фильтрации мы получили 672 варианта анкет, из которых необходимо было выбрать оптимальные варианты, чтобы определить, какой набор параметров лучше всего подходит для решения наших задач.

Для автоматической оценки результатов мы разметили все входные словосочетания, т.е. все коллокации частотностью не ниже 10 употреблений в НКРЯ (см. раздел 2.2) для каждого из отобранных нами 13 прилагательных. Разметка заключалась в том, что для каждого словосочетания мы указали, какой фрейм семантического поля ‘острый’ оно иллюстрирует. В случае, если коллокация относилась к типу значений, не предусмотренному эталонной анкетой (например, *жаркий день* или *душистое сено*), мы приписывали ей нулевой номер фрейма.

Далее оценка проводилась по трем параметрам:

1. Общее количество представленных фреймов
2. Чистота кластеризации
3. Количество «лишних» словосочетаний

Первый параметр соответствует метрике полноты и подразумевает проверку того, все ли ожидаемые фреймы проиллюстрированы в данном варианте анкеты. Чистота кластеризации примерно соответствует точности и оценивает степень однородности кластеров: если в итоговую группу элементов попали словосочетания, относящиеся к трем разным фреймам, то метрика чистоты кластеризации показывала уровень 1/3; если хотя бы у двух словосочетаний совпадала фреймовая принадлежность – 2/3, у всех трех – 1. Числа, полученные в результате оценки каждого кластера, суммировались и делились на общее количество групп. Наконец, третий параметр показывал долю словосочетаний с ненулевым значением фрейма, попавших в итоговую анкету. Таким образом, максимально возможное значение каждой из трех метрик равнялось единице.

Для установления оптимального баланса между полнотой и точностью, мы посчитали F-меру по формуле:

$F = 2PR / (P+R)$, где P – точность, R - полнота

Максимальное значение F-меры (0,72) было получено при конфигурации параметров, приведенной в таблице 15. Метрика «лишних» словосочетаний показывает при этом значение 0,5, не очень высокое, но одно из лучших во всей выборке.

Параметр	Значение
Пространство	Композиция
Метод кластеризации	Rb
Дополнительные модификации	Agglofrom = 168, agglocrfun = i1
Метод отбора элементов кластера	Выбор центроидов

Таблица 15. Оптимальный набор параметров

Внимательный просмотр анкеты, получившей самую высокую оценку (см. Приложение 2), показал, что в ней действительно представлены в большинстве случаев осмысленные группировки, и даже непосредственно основное для нас прилагательное *острый* раскластеризовано довольно неплохо. Однако для того, чтобы проверить, какие параметры будут лучше всего кластеризовать именно словосочетания с опорным прилагательным, мы оценили все анкеты ещё раз, но по другим критериям: на этот раз размечены были исключительно коллокации прилагательных *острый* и *резкий* (которые мы считаем центральными для поля ‘острый’ в русском языке), и по этой разметке были посчитаны полнота и точность, т.е. количество представленных фреймов и чистота кластеров. В данном случае важнее всего для нас была чистота, поэтому мы посчитали F-меру по формуле, позволяющей дать больший вес точности:

F = 1.09*PR / (0.09*P+R), где P – точность, R - полнота

Полученный результат сложно оценить однозначно. Лучше значение F-меры – 0,943, однако просмотр первых двух вариантов анкет показывает, что второй результат (0,94) для нас значительно интереснее: при этой конфигурации достигается наиболее точное разбиение на фреймы прямых значений слова *острый*. Так, выделяется фрейм режущих инструментов (*нож, ножик, лезвие*), фрейм формы (*локоть, локоток, коленка*). Именно это достижение для нас особенно важно: наши эксперименты показывают, что метафоры отстоят дальше от прямых значений, чем различные прямые употребления – друг от друга: даже алгоритмы, автоматически определяющие число кластеров, от которых мы сразу отказались, выделяют некоторые метафоры. Но вычленение прямых значений – очень редкий и важный результат, который достигается при помощи нескольких разных модификаций метода graph (с расширением agglofrom), применяемого на материале пространства композиций с последующим извлечением центральных элементов кластеров.

Заключение

В рамках настоящего исследования мы разработали алгоритм, позволяющий автоматически составлять первичный вариант анкеты для лексико-типологического исследования. Наша процедура включает следующие этапы:

1. Определение списка прилагательных с помощью метода обратных переводов;
2. Сбор коллокаций для каждого прилагательного по основному подкорпусу НКРЯ (с включением в исследование только тех сочетаний, которые встретились в корпусе не менее 10 раз);
3. Подготовка векторного пространства (методом композиции);
4. Кластеризация пространства (методами rb и graph с расширением agglofrom; с предварительным вычислением нужного количества кластеров);
5. Выбор трех центральных представителей каждого кластера.

В ходе работы мы сделали ряд теоретических наблюдений. Так, было экспериментально доказано, что семантические поля предметных имен и относительных прилагательных строятся на отношение когипонимии, а поля качественных признаков – на отношении дистрибуции.

Наши опыты также показали, что для адекватного представления сочетаемости слова важны как объем, так и сбалансированность исходного корпуса текстов. Так, основного подкорпуса НКРЯ достаточно для того, чтобы определить закономерности сочетаемости отдельных лексем. Для моделирования дистрибуции целых словосочетаний его объемов не хватает. С другой стороны, мы видели, что статистика по большим корпусам текстов (таких, как РУВАК) может быть несколько искаженной из-за несбалансированности корпуса. Возможно ли вообще собрать такой корпус текстов, который будет и достаточно представителен (т.е. построен не только и не столько на интернет-текстах), и в то же время очень велик?

На данном этапе, пока ответ на этот вопрос не найден, оптимальным решением является применение моделей композиции, которые позволяют экономить время вычислений и дает более чем приемлемые результаты.

Составленные нами автоматически анкеты, конечно, нельзя рассматривать как готовый инструмент для дальнейших типологических исследований: даже самые лучшие варианты не обладают стопроцентной полнотой и точностью. Безусловно, требуется дальнейшее вмешательство исследователя, который сможет оценить, какие строки из полученного опросника ему заведомо не нужны, какие нуждаются в уточнении. Конечно, при переходе к сбору данных других языков он будет открывать новые фреймы. Помимо

этого, не стоит забывать и о том, что результат, которого мы добились, требует проверки на материале других признаковых полей.

Однако первый (и, возможно, самый важный) шаг уже сделан: мы получили подтверждение для целого ряда наших теоретических установок и положили начало переводу по крайней мере некоторых этапов лексико-типологического исследования на машинную основу.

Так, наша работа лишний раз показывает, что сочетаемость действительно может сказать очень многое о семантике лексемы. Эта идея, безусловно, не нова: на неё опираются Московская Семантическая Школа (см. Апресян 1974), теория Грамматики Конструкций (Goldberg 1995), Московская лексико-типологическая группа MLexT (см. Рахилина 2013) и многие другие, однако с появлением теории Дистрибутивных моделей в семантике эта идея перешла в статус математически доказуемых.

Ещё более важный результат, который мы получили – доказательство того, что анализ сочетаемости позволяет многое узнать не только об одной конкретной лексеме одного конкретного языка, но и об устройстве семантического поля в принципе. Ранее мы утверждали, что для обнаружения основных закономерностей организации поля достаточно изучения нескольких языков, в том числе близкородственных (см. Рахилина, Плунгян 2007, Кюсева, Рыжова 2012). Настоящая работа демонстрирует, что очень многие выводы можно делать уже на материале одного языка. В принципе, мы опирались на эту идею раньше, составляя первичную анкету на русском материале, но если раньше нам были доступны только более или менее поверхностные критерии выделения фреймов (такие, как описание разных типов ситуаций разными лексемами, разные наборы метафор и т.п.) и теоретические рассуждения, то теперь в нашем распоряжении имеется точный инструмент, способный находить более тонкие различия, более глубинные закономерности, скрывающиеся в законах дистрибуции лексем и словосочетаний. И, кстати, насколько нам известно, никогда ранее не применявшийся для решения типологически ориентированных задач. Таким образом, объединение двух парадигм – фреймового подхода к лексической семантике и теории дистрибутивных моделей – позволяет нам сделать новый шаг в развитии каждой из них.

Безусловно, настоящая работа – только начало процесса автоматизации различных этапов лексико-типологического исследования. Наша модель нуждается в дополнительной обработке и апробации: требуются морфологический разметчик и дизамбигуатор более высокого качества; необходима проверка результатов на материале других семантических полей (в том числе, не только признаковых). Однако первый шаг на пути частичной замены ручного труда на машинный уже сделан. В случае, если все дальнейшие

эксперименты будут давать столь же удачные результаты, разработанные алгоритмы можно будет оформить в виде компьютерного инструмента – помощника лексического типолога.

Библиография

Александрова З.Е. 2001. Словарь синонимов русского языка: Практический справочник: Ок. 11 000 синоним. рядов. 11-е изд., перераб. и доп. М.: «Русский язык».

Апресян Ю. Д. 1974. Лексическая семантика: синонимические средства языка. М.: Наука

Апресян Ю.Д. 2004. (под общим рук.) Новый объяснительный словарь русских синонимов. 2-е изд., испр. и доп. Москва – Вена: Языки славянской культуры; Венский славистический альманах.

Большой толковый словарь русского языка. Гл. ред. С. А. Кузнецов. Первое издание: СПб.: Норинт, 1998.

Брицын, В.М.; Рахилина, Е.В.; Резникова, Т.И.; Яворская, Г.М. (ред.) 2009. Концепт БОЛЬ в типологическом освещении. Киев: Видавничий Дім Дмитра Бураго

Кашкин Е.В. 2013. Языковая категоризация фактуры поверхностей (типологическое исследование наименований качественных признаков в уральских языках). Дисс.канд. филол. наук. М.: МГУ

Кронгауз М.А. 2005. Семантика: Учебник для студентов лингвистических факультетов высших учебных заведений, 2-е изд. М.: Издательский центр «Академия».

Круглякова В.А. 2010. Семантика глаголов вращения в типологической перспективе. Дисс.канд. филол. наук. М.: РГГУ

Кюсева М.В. 2012. Лексическая типология семантических сдвигов названий качественных признаков ‘острый’ и ‘тупой’. Дипломная работа. М.: МГУ

Кюсева М.В. 2014. Верификация фреймового подхода к лексической типологии с помощью векторных моделей. Выпускная квалификационная работа, магистратура НИУ ВШЭ.

Кюсева М. В., Резникова Т.И., Рыжова Д. А. 2013. Совершенствование одноязычных, двуязычных и мультиязычных словарей: автоматизация процесса сбора материала // В кн.: Доклады всероссийской научной конференции АИСТ’2013 / Отв. ред.: Е. Л. Черняк; науч. ред.: Д. И. Игнатов, М. Ю. Хачай, О. Баринова. М. : Национальный открытый университет «ИНТУИТ». С. 225-232.

Кюсева М. В., Рыжова Д. А. 2012. Прогностика в лексической семантике: анализ данных близкородственных языков (на материале русской и сербской лексики) // В кн.: Проблемы лингвистической прогностики: Сборник научных трудов Вып. 5. Воронеж: Воронежский государственный университет. С. 58-71.

Кюсева М.В., Рыжова Д.А., Холкина Л.С. 2012. Прилагательные *тяжелый* и *легкий* в типологической перспективе // Компьютерная лингвистика и интеллектуальные

технологии: по материалам ежегодной Международной конференции «Диалог» (Бекасово, 30 мая – 3 июня 2012 г.) В 2-х томах / Отв. ред. А.Е.Кибрик. Т.1: Основная программа конференции. Вып. 11. М.: РГГУ, с. 247-255

Кюсева М. В., Рыжова Д. А., Холкина Л.С. 2013. Лексическая типология: к проблеме определения границ семантического поля (на примере признаков 'толстый' и 'тонкий') // В кн.: Tipologia lexica. Гранада : Jizo Ediciones, С. 255-262.

Лакомкин Е.Д., Пузыревский И.В., Рыжова Д.А. 2013. Анализ статистических алгоритмов снятия морфологической омонимии в русском языке // В кн.: Доклады всероссийской научной конференции АИСТ’2013 / Отв. ред.: Е. Л. Черняк; науч. ред.: Д. И. Игнатов, М. Ю. Хачай, О. Баринова. М. : Национальный открытый университет «ИНТУИТ». С. 184-196.

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

Павлова Е.К. 2014. ‘Мягкий’ и ‘твердый’: к построению лексической типологии. Дипломная работа. М.: МГУ

Рахилина Е. В. 2000/2008. Когнитивный анализ предметных имен: семантика и сочетаемость. М.: Русские словари; М.: Азбуковник, изд.2, испр. и доп.

Рахилина Е. В. О Московской лексико-типологической группе (MLexT) // В кн.: Tipologia lexica. Гранада : Jizo Ediciones, 2013. С. 9-16.

Рахилина Е.В; Резникова Т.И. 2013. Фреймовый подход к лексической типологии // Вопросы языкознания №2, с. 3-31.

Рахилина, Е.В.; Плунгян, В.А. 2007. О лексико-семантической типологии // Т.А. Майсак, Е.В. Рахилина (ред.), Глаголы движения в воде: лексическая типология. М.: Индрик, 9–26.

Резникова Т.И., Кюсева М.В., Рыжова Д.А. 2013. Типологическая база данных адъективной лексики// Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Международной конференции «Диалог» (Бекасово, 29 мая – 2 июня 2013 г.) В 2-х томах /под ред. В.П.Селегей. Т.1: Основная программа конференции. Вып. 12 (19). М.: РГГУ, с. 407-419

Словарь русского языка в 4-х томах («МАС», Малый академический словарь). 1999. М., Русский язык. Т. 1–4

Тагабилева М.Г., Холкина Л.С. 2010. Качественные признаки ‘пустой’ и ‘полный’ в типологическом освещении // Acta Linguistica Petropolitana. Труды Института лингвистических исследований, т. VI, ч.3, Санкт-Петербург: Наука

Фрей А. Потребность в экспрессивности // Грамматика ошибок. М.: УРСС Эдиториал, 2006. С. 216-273.

Фрумкина, Р.М. 1984. Цвет, смысл, сходство. М.: Наука

Холкина Л.С. 2014. Категория качества в китайской лексике. Опыт типологического описания. Дисс. канд. филол. наук. М.: МГУ

Шеманаева О.Ю., Кустова Г.И., Ляшевская О.Н., Рахилина Е.В. 2007. Семантические фильтры для разрешения многозначности в Национальном корпусе русского языка: прилагательные // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2007» (Бекасово, 30 мая - 3 июня 2007 г.) / Под ред. Л.Л. Иомдина, Н.И. Лауфер, А.С. Нариньяни, В.П. Селегея. - М.: Изд-во РГГУ.

Agirre, E. and P.G. Edmonds. 2006. Word Sense Disambiguation: Algorithms and Applications. Springer, 2006.

Baroni, M.; R. Bernardi; R. Zamparelli. 2013. Frege in Space: A Program for Compositional Distributional Semantics // Linguistic Issues in Language Technologies, Vol.9. CSLI Publications.

Comaniciu, D. and P. Meer. 2002. Mean Shift: A robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence. pp. 603-619.

Corbett, G.G. and I.R.L. Davies. 1995. Linguistic and behavioural measures for ranking basic colour terms // Studies in language 19.2, 301-357

Ester, M., H. P. Kriegel, J. Sander, and X. Xu. 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, OR, AAAI Press, pp. 226–231.

François, A. 2008. Semantic maps and the typology of colexification: Intertwining polysemous networks across languages. In: Vanhove, M. (ed.) From Polysemy to Semantic change: Towards a Typology of Lexical Semantic Associations. Studies in Language Companion Series, 106. Amsterdam, Philadelphia: Benjamins. Pp. 163-216

Frey, B. J. and D. Dueck. 2007. Clustering by Passing Messages Between Data Points, Science Feb.

Goldberg A. 1995. Constructions. A Construction Grammar Approach to Argument Structure. Chicago, London: The University of Chicago Press.

Kilgarriff, A., P. Rychly, P. Smrz, D. Tugwell. 2004. The Sketch Engine. Proc EURALEX 2004, Lorient, France; Pp 105-116.

Levinson, S. C. 2000. Yélî Dnye and the theory of basic color terms. Journal of Linguistic Anthropology, 10(1), 3-55.

Lin, D. and P. Pantel. 2001. DIRT – Discovery of Inference Rules from Text. In Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 323–328

Majid, A. and M. Bowerman. (Eds.) 2007. Cutting and breaking events: A crosslinguistic perspective [Special Issue]. Cognitive Linguistics, 18(2)

Majid, A. and S.C. Levinson. 2008. Language does provide support for basic tastes [Commentary on A study of the science of taste: On the origins and influence of the core ideas by Robert P. Erickson]. Behavioral and Brain Sciences, 31. P.86-87.

Mitchell, J.; M. Lapata. 2010. Composition in distributional models of semantics. *Cognitive science* 34(8): 1388-1429

Padó, S.; Lapata, M. 2007. Dependency-based construction of semantic space models. *Computational linguistics* 33(2): 161-199

Salton, G. 1971. The SMART retrieval system: Experiments in automatic document processing. Prentice-Hall, Upper Saddle River, NJ

Schütze, H. 1998. Automatic word sense discrimination // *Computational linguistics* 24(1): 97-123

Wierzbicka, A. 1990. Semantics of color terms: cultural and cognitive aspects // Cognitive linguistics 1.1.

Приложение 1. Типологическая анкета для признакового поля ‘острый’

Фрейм	**Иллюстрация**
острый инструмент с режущим краем	острый нож
	острый меч
	острый коготь
	острая сабля
	острое лезвие
	острая коса
	острые ножницы
	острая пила
острый инструмент с колющим концом	острая иголка
	острая стрела
	острый коготь
	острый кол
	острая спица
	острая палка
	острое копье
	острый зуб
	острый ноготь
	острый рог
предмет, суживающийся к концу	острый нос
	острый носок (ботинка)
	острая гора
	острый локоть
	острый клюв
	острый колпак
	острый подбородок
	острый каблук
	острая мачта
колющаяся поверхность/колющийся предмет	острый шип
	колючее одеяло
	колючая щетина
	колючая шерсть
	колючий куст
	колючий свитер
полоса, резко меняющая направление	крутой поворот
	крутой изгиб
	крутая излучина реки
отвесный, обрывистый, крутой склон	крутой подъем
	крутой склон
	крутой берег
	крутая гора
четкая линия/изображение с четкими линиями	четкая линия
	четкая фотография
	резкий контраст
	четкая картинка
	сильные очки

умный, хорошо соображающий (о человеке)	проницательный человек
	острый ум
	острый, проницательный взгляд
	проницательный наблюдатель
	умная девочка
	умная голова
остроумный, язвительный, меткий (о человеке)	острый журналист
	острое слово
	острый язык
	острое высказывание
	острый критик
	острое замечание
	острая критика
	острая статья
	острый анекдот
грубый, жестокий, резкий (о человеке)	резкий человек
	резкий взгляд
	резкое слово
	резкая реакция
	резкая брань
	грубый ответ
	грубая бесцеремонность
	грубый цинизм
строгий (о человеке)	строгий закон
	строгий правитель
	строгий распорядок
	строгое наказание
	строгая диета
непоседливый (о человеке)	непоседливый ребенок
узкоспециализированное исследование	узкоспециализированное исследование
	узкоспециализированная работа
блюдо с большим количеством специй и пряностей	острый соус
	острый вкус
	острый перец
	острое лечо
газированный напиток	газированный напиток
	газированная вода
неприятно действующий на органы чувств: звук	резкий звук
	острый визг
	резкий голос
	резкий лай
	резкий храп

неприятно действующий на органы чувств:вкус	резкий вкус
неприятно действующий на органы чувств:свет	резкий свет
	резкая вспышка
	острый блеск
неприятно действующий на органы чувств:запах	резкая вонь
	резкий запах
высокий звук	высокий звук
	высокая нота
яркий цвет	цвет
	синева
	оттенок
сильный запах	сильный запах
	сильный аромат
яркий свет	яркий свет
	яркое солнце
хорошо функционирующий орган чувств	острый слух
	острое зрение
	хорошие глаза
	хорошие уши
	острый нюх
	острый нос
отчетливое ощущение	острое ощущение
	острое осознание
сильная эмоция	острое желание
	острая ревность
	острая зависть
	острая обида
	острое стремление
погодное явление, проявляющееся в высокой степени	острый мороз
	резкий ветер
	сильный холод
	сильная жара
требующий немедленного разрешения	острая проблема
	острая нехватка
	острый вопрос
интенсивное физиологическое ощущение	острая боль
	острая жажда
	острый недосып
	острая дрожь
	острая слабость
сильная болезнь	острое воспаление
	острый грипп
	острое заболевание
	острое расстройство
	острый аппендицит
напряженный (о взаимодействиях)	острый конфликт
	острый момент
	острое сопротивление
	острый спор

	острая перепалка
интенсивный по степени проявления (о качествах человека)	острое любопытство
	острая наблюдательность
	острая впечатлительность
	острая гордость
резкое изменение	резкий подъем
	резкий спад
внезапное,быстрое движение	резкий взмах
	резкий вздох
	резкое движение
	резкий бросок
	резкий скачок
быстрый (о скорости)	быстрый темп
	быстрый бег

Приложение 2. Анкета, сконструированная автоматически (метод rb)

№ кластера	**Элемент 1**	**Элемент 2**	**Элемент 3**
1	яркий_халат	яркий_платье	яркий_сарафан
2	колючий_глаз	колючий_глазок	колючий_взгляд
3	пронзительный_взгляд	пронзительный_взор	пронзительный_глаз
4	резкий_выкрик	резкий_крик	резкий_звук
5	быстрый_речка	быстрый_река	быстрый_ручей
6	яркий_светило	яркий_солнце	яркий_сияние
7	жаркий_схватка	жаркий_сражение	жаркий_битва
8	яркий_свет	яркий_светоч	яркий_огонек
9	крутой_плечо	крутой_бок	крутой_шея
10	резкий_увеличение	резкий_ухудшение	резкий_понижение
11	едкий_барит	едкий_натр	едкий_щелочь
12	крутой_обрыв	крутой_скат	крутой_скала
13	острый_приправа	острый_соус	острый_блюдо
14	резкий_критика	резкий_неприятие	резкий_отповедь
15	быстрый_игра	резкий_команда	быстрый_шахматы
16	яркий_образ	яркий_характер	яркий_форма
17	жаркий_глаз	жаркий_дыхание	жаркий_губа
18	острый_запах	острый_аромат	острый_холодок
19	острый_критика	острый_вопрос	острый_полемика
20	высокий_награда	высокий_звание	высокий_заслуга
21	жаркий_день	жаркий_пора	жаркий_время
22	высокий_доход	высокий_издержка	высокий_расход
23	колючий_ветер	колючий_пыль	резкий_ветер
24	высокий_текучесть	высокий_жесткость	высокий_теплопроводность
25	быстрый_переход	быстрый_способ	быстрый_действие
26	резкий_тень	резкий_глаз	резкий_взмах
27	высокий_ряд	высокий_ставок	высокий_сторона
28	высокий_сапожок	высокий_голенище	высокий_сапог
29	яркий_пример	яркий_том	яркий_вид
30	высокий_порядок	высокий_право	высокий_признание
31	высокий_представитель	высокий_совет	высокий_глава
32	горячий_пыль	горячий_след	горячий_воздух
33	острый_зрение	острый_момент	острый_случай
34	горячий_время	горячий_пора	горячий_вид
35	высокий_арка	высокий_башня	высокий_купол
36	колючий_забор	колючий_проволока	колючий_стерня
37	горячий_бульон	горячий_борщ	горячий_суп
38	высокий_чина	высокий_планка	высокий_ноль
39	высокий_внимание	высокий_малый	высокий_род
40	быстрый_говорок	быстрый_взмах	быстрый_шажок
41	высокий_частокол	высокий_стена	высокий_помост
42	высокий_достоинство	высокий_духовность	высокий_призвание

43	высокий_жизнь	высокий_человек	высокий_душа
44	высокий_берег	высокий_льдина	высокий_штиль
45	высокий_подставка	высокий_конторка	высокий_стол
46	высокий_шея	высокий_локоть	высокий_грудь
47	высокий_женщина	высокий_молодая	высокий_девушка
48	отчетливый_голос	отчетливый_звук	отчетливый_стук
49	быстрый_конь	быстрый_рысь	быстрый_лошадь
50	высокий_береза	высокий_сосна	высокий_дуб
51	яркий_цвета	яркий_расцветка	яркий_красная
52	жаркий_дискуссия	жаркий_дебаты	жаркий_спор
53	горячий_благодарность	горячий_признательность	горячий_сочувствие
54	яркий_картинка	яркий_плакат	яркий_этикетка
55	острый_стадия	острый_фаза	острый_инфекция
56	острый_гребень	острый_зубец	острый_вершина
57	горячий_бой	горячий_схватка	горячий_война
58	отчетливый_образ	отчетливый_форма	отчетливый_представление
59	крутой_мера	крутой_место	крутой_перемена
60	горячий_дебаты	горячий_тема	горячий_дискуссия
61	высокий_гор	высокий_увал	высокий_сопка
62	горячий_душа	горячий_душ	горячий_человек
63	резкий_слово	резкий_фраза	резкий_речь
64	высокий_урожай	высокий_урожайность	высокий_сорт
65	резкий_скачок	резкий_ослабление	резкий_перепад
66	высокий_активность	высокий_токсичность	высокий_проба
67	высокий_заросль	высокий_кустарник	высокий_папоротник
68	крутой_ступенька	крутой_лестница	крутой_лесенка
69	крутой_тесто	крутой_каша	крутой_яйцо
70	быстрый_время	быстрый_другой	быстрый_человек
71	высокий_коридор	высокий_зало	высокий_порог
72	высокий_мера	высокий_отношение	высокий_положение
73	острый_ум	острый_слово	острый_мысль
74	высокий_особа	высокий_протекция	высокий_персона
75	высокий_образ	высокий_смысл	высокий_понимание
76	высокий_творчество	высокий_литература	высокий_произведение
77	высокий_должность	высокий_начальство	высокий_чиновник
78	высокий_небо	высокий_облако	высокий_солнце
79	резкий_ход	резкий_движение	резкий_толчок
80	острый_меч	острый_копье	острый_сабля
81	крутой_подъем	крутой_вираж	крутой_спуск
82	едкий_ирония	едкий_сарказм	едкий_насмешка
83	острый_болезнь	острый_паранойя	острый_помешательство
84	душистый_цветок	душистый_роза	душистый_цвета
85	крутой_улочка	крутой_дорожка	крутой_переулок
86	колючий_куст	колючий_трава	колючий_ветка
87	горячий_коф	горячий_пунш	горячий_чай
88	острый_носок	острый_каблук	острый_носка

89	душистый_запах	душистый_воздух	душистый_аромат
90	острый_серп	острый_спора	острый_лопатка
91	высокий_плата	высокий_плат	высокий_зарплата
92	острый_свет	острый_блеск	острый_луч
93	быстрый_увеличение	быстрый_снижение	быстрый_рост
94	горячий_смесь	горячий_спирт	горячий_жидкость
95	острый_чувство	острый_горечь	острый_жалость
96	жаркий_небо	жаркий_солнце	жаркий_тьма
97	горячий_пирога	горячий_спора	горячий_денечек
98	горячий_ладонь	горячий_рука	горячий_пот
99	острый_нос	острый_кадык	острый_палец
100	пронзительный_крик	пронзительный_визг	пронзительный_свист
101	душистый_сена	душистый_горошек	душистый_хлеб
102	колючий_подбородок	колючий_усы	колючий_щека
103	высокий_табурет	высокий_стул	высокий_табуретка
104	высокий_зависимость	высокий_значение	высокий_характеристика
105	высокий_уровень	высокий_доля	высокий_рост
106	яркий_впечатление	яркий_мысль	яркий_жизнь
107	яркий_глаз	яркий_красное	яркий_иза
108	высокий_качество	высокий_результативность	высокий_оперативность
109	быстрый_нейтрон	быстрый_превращение	быстрый_реакция
110	резкий_характер	резкий_образ	резкий_разница
111	высокий_фальцет	высокий_тенорок	высокий_голос
112	острый_камень	острый_кромка	острый_край

Приложение 3. Анкета, сконструированная автоматически (метод graph)

№ кластера	Элемент 1	Элемент 2	Элемент 3
1	высокий_образ	высокий_смысл	высокий_форма
2	острый_ситуация	острый_проблема	острый_кризис
3	горячий_кран	горячий_цинкование	горячий_труба
4	отчетливый_образ	отчетливый_форма	отчетливый_представление
5	высокий_откос	высокий_бугор	высокий_пригорок
6	высокий_глаз	высокий_бровь	высокий_лоб
7	острый_недостаток	острый_интерес	острый_состояние
8	пронзительный_крик	пронзительный_свист	пронзительный_писк
9	едкий_сарказм	едкий_ирония	едкий_горечь
10	высокий_шея	высокий_рука	высокий_локоть
11	яркий_том	яркий_образ	яркий_вид
12	жаркий_день	жаркий_часы	жаркий_вечер
13	высокий_вопрос	высокий_положение	высокий_мера
14	острый_лезвие	острый_ножик	острый_нож
15	острый_ощущение	острый_сознание	острый_переживание
16	острый_словцо	острый_разговор	острый_минута
17	горячий_луч	горячий_свет	горячий_солнце
18	высокий_особа	высокий_протекция	высокий_честь
19	острый_нос	острый_глаз	острый_голова
20	быстрый_вод	быстрый_поворот	быстрый_езда
21	колючий_куст	колючий_изгородь	колючий_трава
22	резкий_слово	резкий_речь	резкий_фраза
23	высокий_сана	высокий_ставок	высокий_скало
24	яркий_женщина	яркий_воспоминание	яркий_тон
25	высокий_глава	высокий_назначение	высокий_представитель
26	быстрый_струя	быстрый_вода	быстрый_волна
27	высокий_текучесть	высокий_жесткость	высокий_зависимость
28	быстрый_разговор	быстрый_шепоток	быстрый_шаг
29	яркий_видение	яркий_сознание	яркий_воплощение
30	горячий_пунш	горячий_спирт	горячий_чай
31	острый_аромат	острый_запах	острый_вкус
32	горячий_след	горячий_ветер	горячий_земля
33	быстрый_сокращение	быстрый_увеличение	быстрый_рост
34	быстрый_превращение	быстрый_обогащение	быстрый_реакция
35	крутой_мера	крутой_перемена	крутой_переход
36	горячий_пора	горячий_человек	горячий_друг
37	горячий_бульон	горячий_лепешка	горячий_суп
38	острый_стрела	острый_пик	острый_камень
39	яркий_печать	яркий_иллюстрация	яркий_подтверждение
40	крутой_место	крутой_ступень	крутой_ступенька
41	высокий_льдина	высокий_штиль	высокий_ординар
42	быстрый_глаз	быстрый_шажок	быстрый_рука
43	жаркий_схватка	жаркий_бой	жаркий_перестрелка

44	высокий_помост	высокий_башня	высокий_стена
45	высокий_литература	высокий_произведение	высокий_искусство
46	высокий_качество	высокий_достижение	высокий_задача
47	острый_блеск	острый_искра	острый_верхушка
48	высокий_небо	высокий_облако	высокий_солнце
49	резкий_диссонанс	резкий_разница	резкий_неприятие
50	жаркий_воздух	жаркий_лет	жаркий_небо
51	горячий_время	горячий_вид	горячий_речь
52	горячий_калач	горячий_блин	горячий_коф
53	душистый_аромат	душистый_сена	душистый_запах
54	высокий_душ	высокий_счастие	высокий_душа
55	яркий_форма	яркий_особенность	яркий_представление
56	высокий_сапожок	высокий_голенище	высокий_сапог
57	высокий_благородство	высокий_честность	высокий_доблесть
58	острый_горечь	острый_жалость	острый_стыд
59	высокий_плата	высокий_плат	высокий_заработок
60	высокий_молодая	высокий_женщина	высокий_девушка
61	резкий_звук	резкий_выкрик	резкий_крик
62	яркий_цвета	яркий_цвет	яркий_расцветка
63	быстрый_переход	быстрый_способ	быстрый_действие
64	яркий_сарафан	яркий_халат	яркий_платье
65	яркий_картинка	яркий_плакат	яркий_витрина
66	резкий_нападки	резкий_выпад	резкий_замечание
67	высокий_белый	высокий_белая	высокий_воротничок
68	резкий_характер	резкий_отношение	резкий_форма
69	высокий_человек	высокий_правда	высокий_жизнь
70	резкий_тень	резкий_очертание	резкий_свет
71	высокий_духовность	высокий_гражданственность	высокий_идеал
72	высокий_достоинство	высокий_репутация	высокий_дарование
73	горячий_дебаты	горячий_призыв	горячий_участие
74	горячий_благодарность	горячий_одобрение	горячий_признательность
75	высокий_звание	высокий_чин	высокий_ранг
76	горячий_уголь	горячий_металл	горячий_газ
77	резкий_толчок	резкий_взмах	резкий_шаг
78	высокий_подсвечник	высокий_свеча	высокий_фонарь
79	высокий_конторка	высокий_стульчик	высокий_табурет
80	яркий_светило	яркий_сияние	яркий_блик
81	высокий_заросль	высокий_папоротник	высокий_кустарник
82	острый_локоть	острый_коленка	острый_локоток
83	крутой_обрыв	крутой_скат	крутой_скала
84	резкий_скачок	резкий_движение	резкий_ход
85	высокий_сосна	высокий_ель	высокий_дуб
86	горячий_пирога	горячий_денечек	горячий_спора
87	быстрый_ход	быстрый_выход	быстрый_движение
88	острый_мысль	острый_конец	острый_человек
89	высокий_гор	высокий_увал	высокий_долина

90	горячий_объятие	горячий_шепот	горячий_поцелуй
91	горячий_любовь	горячий_чувство	горячий_натура
92	горячий_ладонь	горячий_рука	горячий_грудь
93	яркий_фонарь	яркий_огонек	яркий_пламень
94	высокий_дух	высокий_жребий	высокий_подвиг
95	острый_разногласие	острый_противоречие	острый_полемика
96	высокий_стремление	высокий_сознательность	высокий_сознание
97	крутой_плечо	крутой_шея	крутой_парень
98	высокий_лестница	высокий_ступенька	высокий_этаж
99	высокий_степень	высокий_предел	высокий_точка
100	быстрый_время	быстрый_другой	быстрый_дело
101	высокий_ряд	высокий_сторона	высокий_место
102	острый_зуб	острый_палец	острый_ноготь
103	резкий_ухудшение	резкий_понижение	резкий_возрастание
104	быстрый_конь	быстрый_лошадь	быстрый_рысь
105	высокий_порядок	высокий_право	высокий_признание
106	высокий_фальцет	высокий_тенорок	высокий_голос
107	горячий_струйка	горячий_капля	горячий_воздух
108	жаркий_глаз	жаркий_лицо	жаркий_рука
109	яркий_пятно	яркий_красный	яркий_фон
110	острый_болезнь	острый_заболевание	острый_паранойя
111	высокий_доля	высокий_рентабельность	высокий_процент
112	крутой_горка	крутой_тесто	крутой_тачка