ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»
Факультет гуманитарных наук
Школа лингвистики

На правах рукописи

Рыжова Дарья Александровна

АВТОМАТИЗАЦИЯ ЛЕКСИКО-
ТИПОЛОГИЧЕСКИХ ИССЛЕДОВАНИЙ: МЕТОДЫ
И ИНСТРУМЕНТЫ

Специальность «Теория языка»

Диссертация на соискание ученой степени
кандидата филологических наук НИУ ВШЭ

Научный руководитель
dоктор филол. наук
Е.В. Рахилина

Москва 2018
ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ .......................................................................................................................... 4

ГЛАВА 1. ОБЩИЕ СВЕДЕНИЯ .................................................................................................. 9

§1. Краткий обзор имеющихся методик и подходов к типологическому описанию лексики .... 10
  1. Экспериментальная парадигма Института психолингвистики имени Макса Планка ...... 10
  2. Теория семантических примитивов ............................................................................. 12
  3. Серия подходов, основанных на анализе лексикографических источников .................. 13
  4. Серия подходов, основанных на использовании параллельных корпусов .................... 18
  5. Методологическая ниша .............................................................................................. 20
§2. Теоретическая база: фреймовый подход к лексической типологии ......................... 21

ГЛАВА 2. ВЕРИФИКАЦИЯ ПОНЯТИЯ ФРЕЙМА С ПОМОЩЬЮ МОДЕЛЕЙ
ДИСТРИБУТИВНОЙ СЕМАНТИКИ ....................................................................................... 26

§1. Постановка задачи ........................................................................................................... 26
§2. Модели дистрибутивной семантики ............................................................................. 28
§3. Пилотные эксперименты .............................................................................................. 31
§4. Новая серия экспериментов ......................................................................................... 32
  1. Подготовка типологических и дистрибутивных данных ........................................ 33
    1.1. Подготовка типологического векторного пространства ..................................... 33
    Предобработка Базы данных признаковой лексики .................................................. 34
    Метрика типологической близости ......................................................................... 38
    1.2. Подготовка дистрибутивного векторного пространства .................................... 39
    1.3. Подсчет корреляции ............................................................................................ 45
    1.4. Базовый алгоритм ................................................................................................ 46
  2. Эксперимент 1: признаковые поля 'острый' и 'гладкий' ........................................ 46
  3. Эксперимент 2: глаголы качания .............................................................................. 51
  4. Эксперимент 3: англоязычный обучающий корпус ................................................ 54
  5. Эксперимент 4: визуализация векторных пространств ............................................ 56
  6. Выводы ........................................................................................................................ 61

ГЛАВА 3. АВТОМАТИЧЕСКАЯРАЗРАБОТКА АНКЕТЫ С ПОМОЩЬЮ МОДЕЛЕЙ
ДИСТРИБУТИВНОЙ СЕМАНТИКИ ....................................................................................... 65

§1. Краткий обзор существующих методов составления типологических анкет .................. 65
§2. Постановка задачи ......................................................................................................... 67
§3. Определение круга лексем, относящихся к изучаемому полю ................................... 72
  1. Проблема границ поля .............................................................................................. 72
  2. Проблема метафорических значений ...................................................................... 75
  3. Формализация задачи ................................................................................................ 78
  4. Методы (полу)автоматического составления списка прилагательных ....................... 79
    4.1. Метод анализа синонимов .................................................................................. 79
    4.2. Метод ближайших соседей .............................................................................. 80
    4.3. Определение границ поля по материалам онтологии RuWordNet ....................... 83
    4.4. Метод обратных переводов .............................................................................. 84
  5. Анализ результатов ...................................................................................................... 86
§4. Составление списков коллокаций ............................................................................... 89
  1. Выбор корпуса ........................................................................................................... 89
  2. Установление порога частотности ............................................................................. 96
§5. Разделение коллокаций на группы ........................................................................... 98
  1. Выбор основания для кластеризации и подготовка векторного пространства .......... 98
2. Кластеризация векторного пространства ................................................................. 100
  1. Алгоритмы с автоматическим определением количества кластеров ................. 101
  2. Алгоритмы с заданным числом кластеров .......................................................... 102
§7. УМЕНЬШЕНИЕ ОБЪЕМА АНКЕТЫ ........................................................................ 104
§8. ОЦЕНКА РЕЗУЛЬТАТОВ ...................................................................................... 105
§9. ЭКСПЕРИМЕНТЫ С ДРУГМИ ПОЛЯМИ ................................................................ 108
§10. ВЫВОДЫ ............................................................................................................ 112

ГЛАВА 4. МЕТОДЫ АВТОМАТИЧЕСКОГО СБОРА ДАННЫХ .................................... 115
§1. ПОСТАНОВКА ЗАДАЧИ .......................................................................................... 115
§2. МАТЕРИАЛ ДЛЯ ЭКСПЕРИМЕНТОВ .................................................................. 116
§3. АЛГОРИТМЫ АВТОМАТИЧЕСКОГО ЗАПОЛНЕНИЯ ТИПОЛОГИЧЕСКОЙ АНКЕТЫ ................................................................. 118
  3.1. Перевод анкеты ................................................................................................. 118
    2.1.1. Онлайн-переводчики компаний Yandex и Google ..................................... 119
    2.1.2. Машиночитаемые словари Freedic и Verdict ........................................... 121
    2.1.3. Параллельные корпуса .............................................................................. 122
    2.1.4. Анализ результатов .................................................................................... 123
  2.2. Заполнение анкеты ............................................................................................ 125
§4. ВЫВОДЫ ............................................................................................................. 126

ГЛАВА 5. ПОСТРОЕНИЕ СЕМАНТИЧЕСКОЙ КАРТЫ И АНАЛИЗ ТИПОВ СИСТЕМ ...... 129
§1. СУЩЕСТВУЮЩИЕ МЕТОДЫ СОЗДАНИЯ СЕМАНТИЧЕСКИХ КАРТ .................. 129
   1.1. Графовые семантические карты .................................................................... 129
   1.2. Вероятностные семантические карты ............................................................ 130
§2. ГРАФОВАЯ И ВЕРОЯТНОСТНАЯ МОДЕЛИ В ПРИМЕНЕНИИ К НАШЕМУ МАТЕРИАЛУ ................................................................. 131
§3. ПОСТРОЕНИЕ СЕМАНТИЧЕСКИХ КАРТ С ПОМОЩЬЮ РЕШЕТОК ФОРМАЛЬНЫХ ПОНЯТИЙ ................................................................. 137
   1. Анализ формальных понятий ........................................................................... 137
   2. Решетки формальных понятий как лексические семантические карты ......... 138
   3. Представление метафорических значений ...................................................... 141
§4. ВЫВОДЫ ............................................................................................................. 144

ЗАКЛЮЧЕНИЕ .............................................................................................................. 145

БИБЛИОГРАФИЯ .......................................................................................................... 148

ПРИЛОЖЕНИЕ 1. АНКЕТА ДЛЯ ПОЛЯ ‘ОСТРЫЙ’ .................................................... 161
ПРИЛОЖЕНИЕ 2. АНКЕТА ДЛЯ ПОЛЯ ‘ГЛАДКИЙ’.................................................. 166
ПРИЛОЖЕНИЕ 3. АНКЕТА ДЛЯ ПОЛЯ ГЛАГОЛОВ КАЧАНИЯ .................................. 170
ПРИЛОЖЕНИЕ 4. АВТОМАТИЧЕСКИ СКОНСТРУИРОВАННАЯ АНКЕТА ДЛЯ ПОЛЯ ‘ПРЯМОЙ’ ................................................................. 171
ПРИЛОЖЕНИЕ 5. АВТОМАТИЧЕСКИ СКОНСТРУИРОВАННАЯ АНКЕТА ДЛЯ ПОЛЯ ‘ТОЛСТЫЙ’ ................................................................. 173
Введение

Лексическая типология – сравнительно молодая область лингвистики, основной задачей которой является сопоставительный анализ значений слов в разных языках. Сильнейший импульс для развития лексическая типология получила с выходом знаменитой статьи Berlin & Kay 1969 о типологии цветообозначений, где была предложена четкая методика сопоставления лексических значений, широко применяемая до сих пор. Эта работа вызвала широкий резонанс в лингвистическом сообществе и положила начало активному развитию лексической типологии. В настоящее время интерес к типологическому анализу лексики только продолжает возрастать.

За полвека лексическая типология добилась существенных результатов: разработаны различные методики сбора и анализа материала (см., например, недавний обзор Koptjevskaja-Tamm, Rakhilina, & Vanhove 2016, описан целый ряд семантических полей (глаголы давания (Newman 1998), разделения объектов на части (Majid & Bowerman 2007), движения в воде (Maysak & Pahilina 2007), извлечения объектов (Kopecka & Narasimhan 2012) и многие другие, в том числе не-глагольные). Между тем, некоторые методологические сложности по-прежнему не преодолены. В первую очередь, они связаны с тем, что для анализа лексики необходим обширный и представительный материал, который в большинстве случаев невозможно почерпнуть из лексикографических источников. Это вынуждает исследователей разрабатывать специальные анкеты и собирать команду экспертов по различным языкам, способных провести работу с носителями и проанализировать полученный материал. Трудоемкость всего процесса не позволяет проводить подробный анализ общирных семантических зон в большом количестве языков. Поэтому, в большинстве случаев, приходится серьезно ограничивать либо количество языков в выборке, либо степень подробности их анализа. Автоматизация трудоемкой деятельности по сбору и обработке лексического материала позволила бы получить огромный массив структурированных данных для многих языков мира, подготовленных к лексикографической обработке и непосредственному сопоставлению.

Результаты подробного и обширного сравнительного анализа лексических значений, а также сама алгоритмизация и компьютеризация лексико-типологического исследования представляют несомненную теоретическую ценность: они позволяют не только расширять и уточнять данные, полученные ручным путем, но и уточнять методологические
основания, на которых была построена ручная работа с этими данными. В частности, в данной диссертации мы предполагаем доказать реальность и лингвистическую релевантность такого теоретического понятия, как лексико-типологический фрейм, которое лежит в основе наших исследований. Таким образом, с алгоритмизацией лексическая типология повышает свой статус как научно обоснованная область лингвистических исследований: мы строим не гипотезы, а полноценные модели.

Одновременно привлечение в лексическую типологию больших данных принесло бы и практическую пользу: их можно было бы учитывать при решении задач ручного и машинного перевода, а также при разработке более эффективных методик обучения иностранному языку. Таким образом, актуальность предлагаемой на защиту диссертационной работы, определяется востребованностью методов автоматического анализа лексики как в теоретической, так и в прикладной лингвистике.

Наша работа опирается на фреймовый подход к лексической типологии, разработанный Московской лексико-типологической группой MLexT (Рахилина & Резникова 2013; Rakhilina & Reznikova 2016) и восходящий к традициям Московской семантической школы, см. Апресян 1974. Ключевое для данной парадигмы понятие фрейма обозначает минимальную ситуацию, которая может в каком-либо языке описываться отдельной лексемой. Задача типологического описания некоторого семантического поля в таком случае сводится к определению набора составляющих его фреймов (т.е. типов ситуаций, которые могут покрываться относящимися к нему лексемами) и моделей их лексикализации (т.е. стратегии объединения значений в рамках одного лексического средства – прототипически, слова). Набор фреймов определяется через анализ сочетаемости слов, которая изучается по словарям и корпусам и уточняется в ходе опросов носителей, а принципы объединения фреймов отображаются на семантических картах, подобных тем, что создаются по результатам исследований в грамматической типологии (см. Haspelmath 2003).

Цель нашей работы – обосновать фреймовый подход в качестве методологической основы и теоретической базы лексико-типологических исследований и предложить новые методы автоматического сбора и анализа лексико-типологических данных, которые позволят упростить и ускорить процесс сбора первичных данных и обнаружить новые закономерности в выражении лексических значений.

В соответствии с поставленной целью, работа решает следующие задачи:
(1) формализация базовых понятий и процедуры лексико-типологического исследования, выполняемого в рамках фреймовой парадигмы: выделение основных его этапов и формулировка задач, которые должны быть решены на каждом шаге;
(2) подбор и апробация автоматических методов реализации каждого из этапов;
(3) анализ полученных результатов, определение перспектив применения кванитативных методик в лексической типологии.

Основные методы, на которые мы опираемся при разработке алгоритмов автоматического сбора и анализа лексических данных, − это дистрибутивный анализ (модели дистрибутивной семантики, см. Baroni, Bernardi, & Zamparelli 2014), кластерный анализ (Everitt 2011) и анализ формальных понятий (Ganter & Wille 1999).

Научная новизна исследования обусловлена слабой изученностью лексико-типологической области в целом и узким кругом исследований, посвященных задаче разработки компьютерных методов анализа значений слов. Методы, которые мы используем в диссертации, пока не применялись для решения подобных задач. Мы предлагаем свои собственные алгоритмы их внедрения в процесс типологического анализа лексики.

На защиту выносятся следующие положения:
(1) Фреймовая структура поля имеет кванитативное обоснование и представляет собой пересекающиеся кластеры с ярко выраженными центрами («фокусами»).
(2) Предварительный вариант лексико-типологической анкеты может быть получен на основе одноязычного корпуса текстов с помощью моделей дистрибутивной семантики и кластерного анализа полученного дистрибутивного пространства.
(3) Процесс сбора данных по анкете может быть полностью автоматизирован с помощью параллельных и одноязычных корпусов, машиночитаемых переводных словарей и онлайн-переводчиков.
(4) Решетки формальных понятий могут быть использованы как новый аналог семантических карт. Такие карты независимы от изначальных теоретических предпосылок исследователя и имеют более широкий круг возможностей по сравнению с обычными графовыми и вероятностными моделями. Они позволяют отображать не только относительные расстояния между исходными значениями, но и стратегии объединения прямых значений и системные связи между прямыми и метафорическими употреблениями лексем.
Тем самым, теоретическая значимость работы определяется её вкладом в развитие лексической типологии в целом и фреймового подхода в частности. Результаты, полученные в ходе настоящего исследования, позволяют уточнить наши представления об организации семантического пространства лексических значений и выдвинуть новые гипотезы относительно степени их сопоставимости.

Практическая значимость диссертации заключается в разработке алгоритмов, которые могут позволить оптимизировать процесс лексико-типологического исследования, а значит, ускорить процесс подготовки материала, необходимого для решения задач в области лексикографии (в том числе компьютерной), обучения языку, ручного и машинного перевода.


Эстония, 2017). По теме диссертации опубликовано 9 работ, в том числе 5 в изданиях, рекомендованных ВАК.

**Структура работы.** Помимо Введения, работа включает пять глав, заключение, библиографию из 158 названий и пять приложений. В первом разделе Главы 1 приводится обзор существующих методик сравнительного анализа лексики, в том числе новейших компьютерных. Во втором разделе дается обзор фреймового подхода к лексической типологии, на который мы опираемся в настоящем исследовании. В Главе 2 мы представляем результаты серии экспериментов, направленных на оценку состоятельности ключевого для данной парадигмы понятия – фрейма. Главы 3 – 5 посвящены обсуждению возможных методов автоматизации каждого из этапов исследования: разработки анкеты (Глава 3), заполнения анкеты материалами различных языков (Глава 4) и построения семантической карты (Глава 5). Наконец, в Заключении формулируются основные выводы.
Глава 1. Общие сведения

Лексическая типология – относительно молодая, но бурно развивающаяся область лингвистики. По аналогии с грамматической типологией, изучающей инвентарь грамматических значений и средства их выражения в языках мира, лексическая типология стремится описать набор лексических значений и способы их лексикализации, т.е. их обозначения с помощью слов и других словарных средств естественных человеческих языков. Так, лексико-типологический анализ наименований частей человеческого тела показывает, что в одних языках, например, в английском или японском, нижняя часть руки от пальцев до запястья и верхняя – от запястья до плеча – противопоставляются лексически (ср. англ. arm vs. hand, яп. te vs. ude), а в других языках такого противопоставления на лексическом уровне не наблюдается, или, по крайней мере, оно выражено менее явно (ср. русск. рука, обозначающее всю руку, от пальцев до плеча)\(^1\).

Иногда лексическую типологию противопоставляют семантической и лексико-семантической, говоря, что лексическая типология – это «изучение того, как средствами лексики «обеспечивается» и членится то или иное денотативное поле (область действительности и соответствующая понятийная область), какие свойства и отношения реальных объектов оказываются релевантными для их номинации» (Толстая 2013: 142). Семантическая типология, напротив, занимается изучением семантической структуры слова, т.е. выбирает в качестве отправной точки не действительность и понятийную область, а лингвистическую единицу и её характеристики. Лексико-семантическая типология, как это и следует из ее названия, комбинирует два подхода: «…она либо исходит из денотативного поля, но ограничивает его лексически, либо, наоборот, исходит из лексического поля (гнезда и т. п.) и изучает набор и соотношение обозначаемых лексикой этого гнезда денотативных сфер» (там же). Мы будем использовать термин «лексическая типология» в широком смысле, объединяя под этим ярлыком все три типа задач.

Долгое время лексическая типология оставалась в стороне от магистральных направлений лингвистических исследований. На это есть множество причин. Во-первых, инвентарь лексических смыслов заведомо на несколько порядков объемнее инвентаря грамматических значений, что заметно затрудняет их изучение. Во-вторых, среднестатистическая лексическая единица в разы менее частотна, чем стандартный

---

\(^1\) Подробнее о типологии частей тела см. Brown 2005a, 2005b; Majid, Enfield, & van Staden 2006, Majid 2015.
грамматический показатель, что осложняет процесс сбора материала. И, наконец, до недавнего времени основной причиной отсутствия интереса у лингвистов к лексическим значениями была их эфемерность и отсутствие сколько-нибудь надежных методов их анализа и описания, а значит, и отсутствие базы для их сопоставления. Все эти факторы нередко создавали впечатление не просто невозможности адекватного анализа семантического уровня языка в силу отсутствия необходимых инструментов (ср. взгляд Л. Блумфилда (1926)), но и принципиальной безнадежности, бессмысленности этого предприятия: лексический состав языка казался хаосом, в котором просто нет и быть не может никаких более или менее строгих правил2.

Однако со второй половины XX века лексическая типология начинает привлекать к себе все больший и больший интерес исследователей. Разрабатывается целый ряд подходов к анализу словарных значений, предлагается несколько оснований для сопоставления лексики различных языков. Далее мы представим краткий обзор основных методик (§1), а затем более подробно рассмотрим подход, который мы используем в качестве теоретической базы для данного исследования (§2).

§1. Краткий обзор имеющихся методик и подходов к типологическому описанию лексики

1. Экспериментальная парадигма Института психолингвистики имени Макса Планка

Экспериментальная методика сопоставления лексических значений была впервые предложена в знаменитой и уже упоминавшейся в Введении работе Berlin & Kay 1969, посвященной типологическому анализу цветообозначений. Сегодня этот подход является самым распространенным в области лексико-типологических исследований. Его развивает и активно применяет исследовательская группа Института психолингвистики имени Макса Планка (см. Kopecka & Narasimhan 2012; Majid, 2015; Majid & Bowerman, 2007 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

2 Ср. цитату из Di Sciullo & Williams 1987 (с. 3–4): «The lexicon is incredibly boring by its very nature. It <…> is like a prison – it contains only the lawless and the only thing that its inmates have in common is lawlessness» («Лексикон – вещь невероятно скучная по своей природе. <…> Он как тюрьма: в нем одинрушители законов, и беззаконие – это единственное, что объединяет его обитателей» (перевод наш – ДР)).
на части, – для глаголов разбиения, разрезания и т.п.; пузырьков со смесями различного вкуса – для изучения вкусовых прилагательных и т.д. Анкеты предъявляются испытуемым, которых просят описать стимулы словами своего родного языка. Тем самым, можно сказать, что в рамках этой парадигмы минимальным элементом «значения» слова считается совокупность некоторых физических характеристик объекта действительности или ситуации в целом.

Такая методика 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. Поскольку анкета состоит из экстралингвистических стимулов, ее не нужно переводить и адаптировать для каждого конкретного языка выборки. Тем самым, 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. Кроме того, метод элицитации позволяет получить и отрицательную информацию, т.е. запреты на употребление тех или иных лексем в определенных контекстах.

Однако, несмотря на все преимущества, этот метод имеет свои ограничения. В частности, 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 составить анкету, состоящую из экстралингвистических стимулов, 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") значительно сложнее.

Другое ограничение связано с тем, что в центре внимания этой школы с самого начала оказались особенности когнитивного восприятия цвета, звука, вкуса и других экстралингвистических ситуаций и их параметров людьми, относящимися к разным расам и культурам. Лингвистическое поведение слов, описывающих эти ситуации, становится ключом к пониманию более общих когнитивных закономерностей, но его изучение не является самоцелью. Соответственно, и наборы стимулов, разрабатываемые в Институте имени Макса Планка, часто предполагают 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, исследование цветообозначений нацелено на определение того, к каким физическим характеристикам и их значениям в этой зоне чувствительна когнитивная система человека вообще и представителей разных рас и народностей в частности. Ключом к решению этой задачи служит, в том числе, информация о том, какие физические противопоставления могут быть вербализованы в тех или иных языках, а какие — нет (ср. термины «expressibility» («выразимость»)) и «ineffability» («невыразимость») в работе Levinson & Majid 2014).

Между тем, один и тот же с физической точки зрения цвет может описываться разными лексическими средствами, в зависимости от того, атрибутом какого объекта он является, ср. русск. карие глаза vs. каштановые волосы vs. коричневое платье vs. бурый медведь (подробнее об этом см. Рахилина 2010). Эти лингвистические особенности слов выходят за рамки задач исследовательской группы Института имени Макса Планка. Заметим, однако, что этот подход непрерывно развивается и в настоящее время уже не сводится исключительно к элицитации по анкетам, состоящим из экстралингвистических стимулов (см., например, Majid 2015). Однако сам вектор развития остается тем же: анализ когнитивных механизмов через призму языка.

2. Теория семантических примитивов


Так, например, разницу в семантике польского цветообозначения niebeski и английского blue А. Вежбицка следующим образом отражает в их толкованиях:

X – niebeski:
(a) в некоторые моменты на небе можно видеть солнце
когда люди видят что-то, подобное X-у, они могут подумать о небе в такие моменты
X – blue:

(а) в некоторые моменты на небе можно видеть солнце когда люди видят что-то, подобное X-y, они могут подумать о небе в такие моменты
(б) в некоторых местах можно видеть массу воды не потому, что люди в этих местах что-то делали когда люди далеко от таких мест, они могут видеть эту воду когда они видят что-то, подобное X-y, они могут подумать об этом3

Как можно видеть из приведенных толкований, такой подход частично преодолевает недостатки экспериментальной методики, рассмотренной выше. В частности, А. Вежбицкая, описывая цвет, обращается не к физическим характеристикам спектра (тону, яркости и т.п.), а к прототипу (в данном случае, небу и воде), т.е. к контексту, что кажется более естественным для описания лингвистических единиц.

Сложность, однако, заключается в том, что предложенный А. Вежбицкой и К. Годдардом набор универсальных компонентов настолько мал, что семантическое представление каждого лексического значения оказывается очень громоздким, и сравнивать такие значения между собой оказывается крайне затруднительно (ср. похожую критику в работах Geeraerts 1988, 2014 и др.). Помимо этого, маленький набор универсальных смыслов не позволяет описывать тонкие различия между лексемами и, в особенности, исследовать зоны вариативности. Вызывает вопросы и сам универсальный набор: А. Вежбицка и К. Годдард утверждают, что его элементы должны иметь однословное выражение в любом языке, однако на деле это не так (см. Плунгян & Рахилина 1996, а также Goddard & Wierzbicka 1994). Дополнительным ограничением такого подхода является его трудоемкость: исчерпывающее описание той или иной семантической области на материале представительной языковой выборки в рамках этой парадигмы не представляется возможным.

3 Серия подходов, основанных на анализе лексикографических источников


4 URL: [http://clics.lingpy.org/]
В работе А. Франсуа на материале словарных данных анализируется семантическое поле дыхания. В качестве основной единицы типологического описания выбирается словарное значение (а не словарный вход)⁵. Предполагается, что набор таких значений близок к универсальному, а разные лексемы в разных языках покрывают те или иные его подмножества. Основные принципы лексикализации зоны отражаются на семантической карте.


Методика, предложенная А. Франсуа, не лишена недостатков – в первую очередь, связанных с выбором источника данных. Известно, что разные словари характеризуются разной степенью подробности: единой лексикографической традиции и единого шаблона представления семантики многозначного слова пока не существует. Помимо того, что в одних словарях представлены более дробные списки значений лексем, чем в других, тут есть и более существенные теоретические трудности. Е.В. Рахилина, А.С. Выренкова и Л.О. Наний (Наний 2016, Rakhilina, Vyrenkova, Orekhov to appear) провели подробную экспертизу методики Александра Франсуа (на основании предложенных им

---

⁵ Т.е., в терминах МСШ, лексема (одно из значений многозначного слова), а не вокабула (отдельное слово, основная единица словарного описания), см. определение этих терминов в Активном словаре русского языка (Апресян 2014:11). В рамках данной работы мы пользуемся другим терминологическим аппаратом: слово лексема мы употребляем в качестве синонима слов слово и вокабула, а словарные значения называем значениями или концептами.

⁶ Более подробный обзор методик семантического картирования см. в Главе 5 настоящей диссертации.
семантических карт и их описаний) и возможностей ее дигитализации в системах типа CLICS и выявили проблемные точки такого подхода.

Во-первых, словари не всегда отражают устаревшие значения. Но даже если они помечены, информация об этом может быть проигнорирована, особенно при автоматическом ее извлечении, как в системах типа CLICS. Примером может служить представление как результата системной колексификации в русском языке метафорического значения, реализуемого в словосочетаниях вида дух народа (войска, страны), и лексикализованного значения дух как ‘breathing’ (ср. современное русское вдох / выдох). Употребления второго типа в современном русском языке характеризуются очень узкой и лексикализованной сочетаемостью. Фактически она ограничена только тремя фразами: испустить дух, дух захватывает, перевести дух, которые действительно исторически восходят к глаголу дышать, но уже по данным Малого академического словаря (Евгеньева 1999) являются периферийными и устаревшими и синхронно с другими его производными не ассоциируются. Любопытно, что МАС, который не используется в исследовании François 2008 (в отношении русского материала эта статья опирается на данные словаря Sakhno 2005), разводит эти употребления по разным словарным статьям без взаимных ссылок друг к другу.

Во-вторых, словари не различают прямые и метафорические значения в том смысле, что их статус там в большинстве случаев никак не маркирован. По нашему мнению, это существенно затрудняет сопоставление языков, потому что семантическая карта оказывается разнородной: загроможденной семантическими узлами разного происхождения с разным «семантическим расстоянием» между ними, для которых трудно механически установить непосредственные семантические связи (ср. анализ глагола òl из языка мвотлап в работе А. Франсуа).

Между тем, в лексической типологии, как и в других научных областях, в последние годы набирают все большую популярность компьютерные методы сбора и анализа данных, для многих из которых основным источником информации оказываются именно словари.

бллизкими считаются те концепты, которые во многих языках обозначаются одним и тем же словом (например, ‘день’ и ‘солние’ семантически значительно ближе друг к другу, чем ‘день’ и ‘трава’, ср., в частности, венгерское nap, которое объединяет два первых значения). Частота объединения вычисляется по переводным словарям на основе аккуратно составленной репрезентативной выборки из 81 языка. Выявленные связи между концептами представляются в виде взвешенного графа. Тем самым, в работе проводится автоматизированный типологический анализ сразу нескольких предметных полей на материале значительного количества ареально и генетически разнородных языков. В качестве базы для сопоставления слов из разных языков (т.е. своеобразного аналого лексико-типологической анкеты) используется набор словарных подзначений, составляемый в процессе обработки словарей. Очень близкая методология предложена в работе Sejane & Eger 2013. Эти работы характеризуются теми же недостатками, что и только что рассмотренный подход А. Франсуа.

Несомненное достоинство подобных методов анализа лексики заключается в том, что они могут быть полностью автоматизированы, что и позволяет им учитывать материал обширных и разнообразных языковых выборок. Однако они связаны с рядом методологических проблем другого характера. Часть из них мы обсудили выше в связи с методикой А. Франсуа. Подчеркнем здесь ещё раз, что словарные данные не очень надежны в силу своей плохой сопоставимости: все включаемые в исследование словари оказываются разного объема и разного качества. Вероятно, это не очень существенно для анализа базовой предметной лексики (будучи базовой и частотной, она должна быть представлена в словаре практически любого объема и степени подробности), но для исследования любой другой семантической зоны применить этот метод будет крайне затруднительно.

Проблема сопоставимости словарных данных отчасти преодолена в системе CLICS, в основе которой лежат данные серии «Межконтинентальных словарей» (Intercontinental Dictionary Series, IDS, см. Key & Comrie 2007). В этих словарях есть единая база «концептов» – список (анкета) из 1310 лексических значений, и для каждого из них указывается, каким лексическим средством оно обслуживается в языке, для которого создается словарь. На основе этих данных строятся взвешенные графы (подобные представленным в работе Youn et al. 2016), отражающие стратегии колексификации – объединения разных значений в рамках одного слова. Недостаток этого чрезвычайно ценного ресурса, который ещё только предстоит преодолеть, – это далекий от идеального
список концептов, лежащий в его основе. В большинстве случаев он достаточно хорош для предметной лексики, но глагольные и признаковые зоны в нем представлены довольно бедно. Так, например, концепта ‘качаться / колебаться’ в нем нет совсем, а признаку ‘старый’ соответствует только один концепт, хотя известно, что в языках мира в этой зоне могут противопоставляться по меньшей мере четыре значения: ‘старый о живых существах’ (старый человек) vs. ‘старый о неодушевленных объектах’ (старые тряпки) vs. ‘бывший’ (старый директор) vs. ‘древний, старинный’ (старый город), ср. грузинские лексемы toxuci, dzveli, qop’ili и adrindeli, соответствующие именно этим четырем понятийным фрагментам (см. Rakhilina, Vyrenkova, Orekhov to appear).

Нельзя не отметить, что именно словари как источник обширных лексикографических данных могут служить основой для глобальных лексикотипологических исследований, не фокусирующихся на узкой семантической области, а охватывающих широкий пласт лексики разных языков. Только что рассмотренные нами проекты Youn et al. 2016 и CLICS (List et al. 2014) могут служить примером такого рода типологии лексических систем.

Ещё один пример лексико-типологического исследования на уровне систем, а не их фрагментов представлен в работе Kibrik 2012. Этот подход также основан, в первую очередь, на словарных данных, но его особенность заключается в том, что он фокусируется не на стратегиях колексификации значений, а на степени морфологической сложности слов, их выражающих, которая, как предполагается, отражает и уровень концептуальной сложности соответствующего лексического значения. В работе сопоставляется глагольная лексика трех языков: английского, русского и атабаского языка коюкон. За основу берется список английских глаголов, которые раньше всего усваиваются детьми, предложенный в Tomasello 1992. Предполагается, что такие слова соответствуют базовым понятийным фрагментам, а значит, их можно условно считать ядром глагольной лексики, причем не только английской, но и любой другой лексической системы. Каждому глаголу из этого списка ставится в соответствие его переводной эквивалент в русском и в коюконе и затем для каждого глагола указывается уровень его морфологической сложности – количество деривационных морфем (включая корневую), входящих в его состав. Самым высоким уровнем морфологической сложности в зоне ядерной глагольной лексики обладает коюкон, самым низким – английский, а русский язык представляет собой промежуточный случай. Если разделить список глагольных значений на семантические классы (такие как ‘глаголы манипуляции’, ‘глаголы движения’ и т.п.), то выясняется, что
в одних лексических зонах (в частности, в зонах глаголов движения, положения в пространстве и речи/звучка) доля непроизводных глаголов существенно выше, чем в других семантических областях. Таким образом, этот подход направлен на решение очень фундаментальной задачи: построения типологии лексических систем с точки зрения того, какие значения для данного языка являются более базовыми и когнитивно простыми, а какие представляют собой более сложные концептуальные структуры. Обычно лексические типологии игнорируют эту задачу: ни направление NSM, ни Неймегенская школа ею не занимаются. Однако для методологии MLexT, которую мы используем в качестве теоретической базы в настоящей работе, она постепенно становится релевантна, ввиду того, что в последних исследованиях этого направления начинает привлекаться материал китайских двуслогов (Kholkina to appear), а также славянские приставочные корреляты как участники оппозиций в семантических полях глаголов движения (Levin, Reznikova to appear). Однако ее формализация в рамках парадигмы MLexT требует дополнительных теоретических изысканий, выходящих за рамки поставленных здесь целей.

4. Серия подходов, основанных на использовании параллельных корпусов

Другой способ проведения лексико-типологического исследования на основе автоматической обработки данных значительной языковой выборки – использование параллельных корпусов. Так, например, в работе Eger 2012, которая является развитием метода, предложенного в Sejane & Eger 2013, предлагается составлять с помощью параллельных корпусов наборы переводных эквивалентов для каждой заданной лексемы, т.е. использовать этот ресурс в качестве замены для недостаточно надежных переводных словарей.

Дополнительное преимущество параллельных корпусов по сравнению со словарями заключается в том, что они позволяют не только находить переводные эквиваленты исходных лексем, но и исследовать особенности их дистрибуции, т.е. определять, в каких именно контекстах заданной лексеме соответствует тот или иной эквивалент из другого языка. Так, например, в работе Viberg 2013a – одном из первых исследований лексико-типологического характера, опирающихся на данные параллельного корпуса – проводится сравнительный анализ особенностей употребления глаголов позиции (‘сидеть’, ‘стоять’, ‘лежать’) в роли локативных предикатов в шведском, английском, немецком, французском и финском языках (ср. также близкую работу Viberg 2013b).
Кроме того, сочетаемостные свойства лексем на материале параллельных корпусов значительно проще сопоставлять автоматически, чем более разнородные данные одноязычных корпусов и переводных словарей. Работа Wälchli & Cysouw 2012 является собой пример полностью автоматизированного лексико-типологического анализа базовых глаголов движения. Исследование проводится на базе корпуса переводов Евангелия от Марка на материале 100 генетически и ареально разнородных языков. Эта методология опирается на предположение о том, что семантика лексемы отражается на ее сочетаемости, а значит, может быть изучена по контекстам, в которых эта лексема употребляется. Параллельные корпуса дают возможность проанализировать один и тот же набор контекстов в разных языках и использовать его в качестве единой базы для сравнения материала разных языков (тем самым, набор всех встреченных в корпусе контекстов фактически становится аналогом типологической анкеты 7). На основе собранных типологических данных контексты кластеризуются, т.е. объединяются в группы, соответствующие «минимальным значениям», и отображаются на семантической карте. Заметим, однако, что такая карта будет отличаться от «традиционных» семантических карт, описанных выше. Карта того образца, который мы встречаем в работе Wälchli & Cysouw 2012, строится методом многомерного шкалирования и отражает, во-первых, все пункты анкеты, а не установленные минимальные значения, а во-вторых, является непрерывной, а не дискретной. Подробнее типы семантических карт и различные методы их создания мы обсудим в Главе 5.

Подход, предложенный в статье Wälchli & Cysouw 2012, обладает неоспоримыми преимуществами: он позволяет проводить лексико-типологические исследования на больших выборках; опирается на собственно языковые примеры, не опосредованные обработкой лексикографов; и, наконец, он может быть полностью автоматизирован, как это и демонстрируют сами авторы. Существенный недостаток этой методологии заключается, однако, в том, что на данный момент она применима только для анализа очень частотной лексики: объемы существующих параллельных корпусов на сегодняшний день очень ограничены.

7 Ср. также предложение составлять на базе параллельных корпусов анкеты для исследований в области грамматической типологии в статье (Dahl, 2007).
5. Методологическая ниша

Итак, в современной мировой практике можно выделить несколько подходов к лексической типологии, каждый из которых опирается на один из трех возможных источников данных: полевую работу (психолингвистический подход Института имени Макса Планка), словарную информацию (А. Франсуа, А.А. Кибрик и некоторые другие) и корпусные данные (А. Виберг, Б. Вельхли и М. Сисоу и др.). Особняком стоит подход Анны Вежбицкой и Клиффа Годдарда, которые предлагают анализировать языковое поведение слов, но сам анализ строят преимущественно на основе интроспекции.

Методология, на которую мы опираемся в настоящем исследовании и которую мы рассмотрим подробно в следующем разделе, – фреймовый подход к лексической типологии (Рахилина & Резникова 2013; Rakhilina & Reznikova 2016) – совмещает в себе все три основных подхода и предполагает работу со словарями, с корпусами (в случае, если они доступны) и с носителями-информантами. Это позволяет получить наиболее полное представление о семантике изучаемых лексем. С другой стороны, столь многосторонний анализ оказывается очень трудоемким и заставляет существенно сократить языковую выборку, что и обуславливает необходимость разработки методов автоматизации различных этапов лексико-типологического исследования в этой парадигме.

Как мы только что показали, компьютерные методы сбора и анализа данных в области лексической типологии уже начинают появляться, однако у каждого из них есть довольно существенные недостатки. Принципиальное отличие наших экспериментов от рассмотренных выше разработок заключается в том, что мы располагаем существенным массивом лексико-типологических данных, собранных вручную по методологии, отработанной на разнородном лексическом материале. Эти данные мы можем использовать в качестве “золотого стандарта” и, тем самым, не просто предлагать те или иные методы компьютерной обработки, но и определять, насколько эти методы применимы для поставленных задач, а также подбирать наиболее оптимальные параметры конструируемых систем.

В следующем разделе мы обсудим основные положения фреймового подхода к лексической типологии, выделим этапы исследования в рамках этой парадигмы и сформулируем, какая задача должна быть решена на каждом из шагов.
§2. Теоретическая база: фреймовый подход к лексической типологии

Фреймовый подход к лексической типологии, на который мы опираемся в настоящей работе, разработан Московской лексико-типологической группой (MLexT) под руководством Е.В. Рахилиной. Он восходит к традициям Московской семантической школы (см. Апресян 1974, 2004, 2014) и предполагает сравнительный анализ внутри- и межъязыковых квазисинонимов через призму их сочетаемостных свойств. Для того, чтобы сравнивать слова из различных языков, удобно разбить их семантику на непересекающиеся понятийные фрагменты, т.е. типы ситуаций, в которых эти слова могут употребляться. Каждому типу ситуаций соответствуют разные группы контекстов.

Например, семантику русского прилагательного тонкий можно представить в виде такого набора понятийных фрагментов:

- ‘малый диаметр поперечного среза’; реализуется в контексте названий длинных вытянутых предметов [карандаш, веревка, палка]
- ‘малое расстояние от одной грани объекта до другой’ – о размере плоских предметов («слоев»), таких как книга, ткань или бумага;
- ‘слабая громкость и высокий частотный диапазон’ – качество звука, реализуется в контексте существительных звук, голос;
- ‘учтивый, деликатный, тактичный’ – о качествах человека;
- …

Те же самые элементарные ситуации позволяют устанавливать соответствия между русским прилагательным тонкий и его переводными эквивалентами. Ср., например, перевод этого слова на китайский:

- ‘тонкий’ + название длинного вытянутого предмета => xī (xi gùnzi – ‘тонкая палка’);

Такие ситуации, релевантные для лексикализации рассматриваемой семантической зоны (см. Рахилина & Резникова 2013 (с. 12): «…ситуации, которые чаще всего противопоставляются лексическими средствами в языках мира»), называются фреймами. Эти единицы можно считать эквивалентами традиционных лексических значений (словарных входов), но их основное отличие от значений, обычно выделяемых в словарях, заключается в том, что они типологически ориентированы, и их набор составляется по результатам подробного анализа материала нескольких языков. Кроме того, обычно
фреймы описываются не с помощью толкований, а посредством определения типов контекстов, которые этот фрейм иллюстрируют. Предполагается, что фреймы — это минимальные лексические значения, и любое слово в любом языке, относящееся к рассматриваемой семантической области, должно покрывать ту или иную их комбинацию.

ПРИМЕЧАНИЕ

Термин фрейм был введен Марвином Минском (см. Минский 1979) для обозначения особых единиц памяти, пучков связанных между собой смыслов в задачах моделирования баз знаний в области искусственного интеллекта. Фреймами назывались ситуации-сценарии, предполагающие определенный состав участников и взаимоотношения между ними. Утверждалось, что элементы фрейма тесно связаны между собой: упоминание одного из них актуализирует все остальные (например, слово ресторан вызывает в сознании человека представления об определенной ситуации и особом сценарии поведения: об особом помещении, официантах, меню, еде и напитках, счета, чаевых и т.п.). В лингвистике это понятие активно использовалось Чарльзом Филлмором (см. Fillmore 1982, a также созданный под его руководством ресурс FrameNet 8 ) для моделирования предикатных отношений: анализа наборов участников, релевантных для того или иного фрейма, их семантических ролей и моделей их синтаксического оформления. Понятие фрейма, о котором говорим мы, основано на теории фреймовой семантики Ч. Филлмора, однако в него вносятся определенные модификации: к представлению о семантической роли и синтаксическом оформлении добавляются сведения о семантических типах участников (в простейшем случае, информация о том, к каким таксономическим классам они принадлежат). Например, филлморовский фрейм CUTTING объединяет различные типы разделения объекта на части и включает три ядерных участника (агент (Agent), пациенс (Item) и фрагменты, на которые разделяется пациенс (Pieces)) и несколько периферийных (инструмент, цель, время, место и др.). Этих сведений достаточно для анализа аргументной структуры относящихся к этому фрейму слов, но не для исследования их лексической семантики. У русских глаголов рубить (дрова топором) и резать (бумагу ножницами) аргументные структуры одинаковые, но при этом они обозначают два разных типа действий, которые совершаются над разными типами объектов, а, следовательно, и с помощью разных инструментов – с этим связано и наличие в этой семантической зоне разных лексических средств. Чтобы такого рода лексические противопоставления можно было выявлять и анализировать, группа MLeXТ называет отдельным фреймом каждую когнитивно значимую комбинацию, складывающуюся из типа действия, объекта, инструмента и прочих участников ситуации разделения объекта на части (см. Кашкин 2010).

Тем самым, фреймы в понимании MLeXТ – это более дробные, более специффицированные семантические фрагменты, чем фреймы Ч. Филлмора. Заметим, что и самому Ч. Филлмору такое представление было не чуждо. В статье Fillmore & Atkins 2000, где приводится сравнительный анализ английского глагола crawl 'ползать' и его французского эквивалента ramper, выделяются значения, основанные на различиях в типе движущегося объекта: движение в полном контакте с поверхностью (ср. ползание змей, червяков, улиток) vs. в частичном контакте с поверхностью (характерно для крабов, пауков и т.п.).

8 URL: [https://framenet.icsi.berkeley.edu/]
Не все комбинации фреймов одинаково вероятны: какие-то значения часто объединяются в рамках одного слова, а какие-то, напротив, в большинстве случаев оказываются лексически противопоставлены. Закономерности объединения фреймов в рамках одной лексемы представляются графически в виде лексико-семантических карт. Такие карты строятся по тем же принципам, что и семантические карты в грамматической типологии (см. Haspelmath 2003) и лексические карты А. Франсуа (см. François 2008, а также §1.3), с той разницей, что узлами графа служат фреймы, а не грамматические функции или словарные подзначения.

Этот подход, в отличие от предложенных в работах Youn et al. 2016 и Kibrik 2012 или реализованных в системах типа CLICS, предполагает очень подробный анализ компактных лексических областей, ср. анализ глаголов движения в воде (Майсак & Рахилина 2007), предикатов боли (Брицын и др. 2009), глаголов вращения (Круглякова 2010), глаголов звуков животных (Резникова и др. 2015; Rakhilina, Merle, & Chahine 2017), признаковой лексики (Кашкин 2013; Кюсева 2012; Холкина 2014) и др. В ходе проведенных исследований показано, что этот подход действительно позволяет анализировать и эффективно сопоставлять лексику разных языков, выявляя для каждого поля его фреймовую структуру и сравнивая языки между собой на основании того, как они членят эту понятийную область с помощью лексических средств. Для более глобальных задач в области широкой типологии лексических систем в нем пока накоплено слишком мало данных, однако кажется, что именно списки фреймов могли бы стать надежной основой и для баз, подобных CLICS, и для метода анализа лексических систем в целом, предложенного в работе Kibrik 2012.

Исследование лексических единиц в этой парадигме включает несколько основных этапов:

1. Составление анкеты, т.е. предварительное определение набора фреймов и подбор диагностических контекстов на каждый фрейм.
2. Сбор данных других языков выборки.
3. Составление семантической карты для описания системы каждого языка и ее визуализации.
4. Анализ типов систем, реализованных в разных языках.

Все эти этапы очень трудоемки. Процесс составления анкеты (этап 1) подразумевает определение набора фреймов, релевантных для рассматриваемого поля, и выбор наиболее показательных контекстов-иллюстраций для каждого фрейма. Только для этого, самого
первого шага уже необходимо провести подробный анализ сочетаемости лексем, покрывающих выбранную семантическую область, на материале словарных и корпусных данных 3-5 языков, дополнив их в ходе опросов носителей. Далее, поскольку подготовленная по итогам первого этапа работы анкета состоит из контекстов, в которых могут употребляться изучаемые слова, ее необходимо переводить на каждый из языков, включаемых в выборку, и только после этого заполнять (этап 2), также на основе словарей, корпусов и работы с носителями. Оба этих этапа в рамках фреймового подхода обычно выполняются вручную и требуют долгой, кропотливой и согласованной работы специалистов по всем языкам, включенным в выборку.

Задачи построения семантической карты и последующего анализа типов систем (этапы 3 и 4) также требуют ручной работы, которая осложняется тем, что эти этапы хуже всего формализованы (подробное обсуждение этой проблемы см. в работе Croft & Poole 2008, а также в Главе 5 данной диссертации). Во многих случаях можно построить карты разных конфигураций, которые будут одинаково хорошо соответствовать собранным данным. Кроме того, такие карты обычно строятся вручную, и, если данных много, то вероятность упустить те или иные закономерности очень высока.

Трудоемкость всего процесса вкупе с необходимостью привлечения эксперта для анализа материала каждого нового языка не позволяют проводить исследование на основе достаточно представительных языковых выборок. В свою очередь, небольшие размеры выборок заставляют усомниться в значимости получаемых результатов, в частности, в том, что выделение особых семантических единиц (фреймов), претендующих на статус минимальных лексических значений, действительно лингвистически оправданно. Задача настоящего исследования — предложить возможные алгоритмы автоматизации каждого из этапов работы в рамках фреймовой парадигмы, а также провести процедуру независимой верификации центрального для данного подхода понятия фрейма.

Мы начнем с задачи верификации понятия фрейма (Глава 2), а затем расскажем последовательно о возможных методах автоматизации каждого шага лексико-типологического исследования: составления анкеты (Глава 3), ее заполнения (Глава 4), составления семантической карты и анализа полученных данных (Глава 5). Наша задача — подобрать оптимальные методы и инструменты для каждого шага исследования, поэтому мы тестируем все алгоритмы независимо друг друга. Наше преимущество заключается в том, что в нашем распоряжении есть обширный типологический материал, собранный вручную, поэтому у нас есть возможность подавать каждому алгоритму на вход...
проверенные данные, чтобы результаты текущих экспериментов не были зашумлены ошибками автоматической обработки, накопленными на предыдущих этапах.
Глава 2. Верификация понятия фрейма с помощью моделей дистрибутивной семантики

§1. Постановка задачи

В этой главе мы представим серию экспериментов, направленных на поиски дополнительных обоснований для выделения фреймов. Фреймовая структура поля определяется в терминах семантической близости: ситуации, относящиеся к одному фрейму, наиболее близки семантически, а между ситуациями из разных фреймов расстояния могут быть разные, и именно эти расстояния отражает семантическая карта рассматриваемого поля.

Семантическое расстояние между фреймами определяется на основе типологических данных. Обычно в рамках фреймового подхода учитываются только относительные расстояния: если некоторая лексема L1 может покрывать только фреймы F1 и F2, а лексема L2 – только значения F2 и F3, но при этом нет ни одного слова, которое обозначало бы F1 и F3, не охватывая при этом F2, утверждается, что фреймы F1 и F3 находятся дальше друг от друга, чем F1 и F2 или F2 и F3. Такая конфигурация фреймов иллюстрируется с помощью линейной семантической карты: F1 – F2 – F3.

Мы разработали формулу более точного, численного определения типологической близости между фреймами на основе данных о частоте колексификации минимальных значений, т.е. сведений о том, насколько часто та или иная пара фреймов обозначается одним и тем же словом в языках нашей выборки. Каждый фрейм представляется в виде вектора w, в качестве измерений которого выступают лексемы изучаемого поля из всех языков выборки. В случае, если лексема l может описывать данный фрейм, соответствующее ей измерение w принимает значение 1, а если не может – 0.

Типологическое расстояние между фреймами (мы называем эту метрику TC – typological closeness) определяется с помощью косинусной меры близости между представляющими их векторами (ср. похожую метрику близости в недавней работе Youn et al. 2016). Подробнее процедура вычисления этого значения описана в разделе 4.1 данной главы.

Однако известны и другие методы определения семантических расстояний между лексическими значениями, в частности, представление семантики лексической единицы

9 Сразу оговоримся, что здесь и далее мы используем термин эксперимент в том значении, в каком он обычно используется в компьютерной лингвистике (см., например, Большакова и др. 2011), т.е. для обозначения так называемых вычислительных, а не психо- или нейролингвистических экспериментов.
(слова или словосочетания) с помощью вектора ее сочетаемости (т.н. модели дистрибутивной семантики, см. Baroni et al. 2014). Вектор сочетаемости вычисляется на основе контекстов употребления слова в некотором корпусе текстов. Контексты употребления языковых единиц — это наблюдаемые сущности, более доступные для объективного анализа, чем значения. Понятие контекста можно формализовать и степень сходства между двумя разными контекстами можно так или иначе измерить. Далее, если принять допущение о том, что близкие по значению языковые единицы употребляются в похожих контекстах (см. Harris 1957, Sahlgren 2008 и многие другие работы, в которых высказывается подобная точка зрения), то, имея данные о степени сходства сочетаемостных свойств слов или словосочетаний, можно судить о степени их семантической близости. Именно на эту идею опирается теория дистрибутивной семантики: она позволяет представить любую языковую единицу (будь то морфема, слово или словосочетание) в виде суммы контекстов, в которых она встречается в том или ином корпусе текстов. Такие представления можно сравнивать между собой и использовать результаты этого сравнения (т.н. дистрибутивные расстояния) в качестве ключа к определению степени семантической близости между рассматриваемыми языковыми единицами.

Эта методика уже используется для решения широкого круга задач, в том числе, близких к нашей (например, для семантической дизамбигуации или выбора из ряда квазисинонимов слова, наиболее подходящего для данного контекста). Насколько нам известно, в типологии подобные методики ещё не применялись, однако можно предположить, что в том случае, если фреймы — это действительно некоторые более или менее единые семантические единицы, то словосочетания, иллюстрирующие один и тот же фрейм (например, тонкая палка и тонкий карандаш), должны употребляться в более близких контекстах, чем словосочетания, относящиеся к разным фреймам (ср. тонкая палка и тонкий слой). Кроме того, если мы утверждаем, что фреймы — это элементы так называемого универсального лексического набора (см. Рахилина & Резникова 2013), т.е. любое слово из любого языка, относящееся к рассматриваемому полю, описывает ту или иную их комбинацию, то неважно, на основе какого языка мы будем считать дистрибутивные расстояния: наблюдаемые данные разных языков должны отражать одни и те же семантические, когнитивные структуры.

В этой главе мы опишем серию экспериментов, направленных на проверку этих гипотез, т.е. на выявление степени соответствия типологических расстояний
дистрибутивным. Для этого мы сначала обсудим более подробно теорию моделей дистрибутивной семантики и основные принципы их работы (§2). Затем, в параграфе 3, представим пилотные эксперименты по сопоставлению типологических и дистрибутивных расстояний между фреймами для поля ‘острый’, после чего перейдем к описанию нашей новой серии экспериментов (§4). Параграф 4, основной в данной главе, структурирован следующим образом: сначала мы документируем процедуру подготовки типологического и дистрибутивного пространств (раздел 1), затем последовательно описываем процедуру и результаты экспериментов на материале признаковых полей ‘острый’ и ‘гладкий’ и глагольного поля качания (разделы 2 – 3). В следующем эксперименте (раздел 4) мы снова работаем с полем ‘острый’, но меняем обучающий корпус для дистрибутивных моделей с русскоязычного на англоязычный, чтобы посмотреть, насколько результаты зависят от стартового языка. Наконец, в последнем эксперименте (раздел 5) мы сопоставляем визуализации типологического и дистрибутивного пространств для нескольких признаковых полей и в разделе 6 формулируем основные выводы.

§2. Модели дистрибутивной семантики

Модели дистрибутивной семантики (или DSM – Distributional Semantics Models, см. Baroni et al. 2014) лучше всего подходят для решения наших задач, поскольку они основываются на тех же самых теоретических предпосылках, что и фреймовый подход к лексической типологии (а также Московская семантическая школа и многие другие теории): на гипотезе о том, что представление о значении слова можно получить путем анализа его сочетаемости. 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 языковой единице (чаще всего, слову) 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. Считается, что, 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 получить больше данных для статистики.


Так, например, в модели с двумя измерениями (ось 1: пить, ось 2: есть) и размером окна в одно слово справа или слева от опорного семантические вектора для существительных чай, кофе и мороженое могли бы иметь следующий вид (пример сконструирован нами для простоты изложения, а не основан на реальных данных того или иного корпуса текстов):

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, однако в применении к нашим задачам эта метрика даёт содержательный результат в том случае, если сравниваемые вектора имеют равную длину. Длина вектора отражает уровень частотности моделируемой языковой единицы, поэтому меру евклидова расстояния обычно применяют только к так называемым нормализованным векторам, приведенным к общей длине, чтобы разница в частотности не препятствовала определению степени семантического сходства между языковыми
единицами (т.е., говоря очень грубо, чтобы можно было определить, что не очень частотное слово миска семантически близко к частотному слову тарелка).

Для наиболее точного вычисления степени близости между векторами, помимо нормализации, используются и другие вспомогательные операции. В частности, распространенной проблемой является недостаток статистических данных, который преодолевается разными способами. Прежде всего, с высокой разреженностью векторов (т.е. большим количеством нулевых значений измерений) можно бороться с помощью уменьшения размерности пространства. Эта процедура способствует уменьшению количества времени, требуемого для последующей обработки данных, а в ряде случаев и повышает качество векторных представлений языковых единиц.

Ещё один тип операций полезен для представления информации о сочетаемости единиц, больших, чем слово, в первую очередь, двусловных словосочетаний. Модели дистрибутивной семантики позволяют вычислять вектора сочетаемости и для такого рода опорных элементов: словосочетание может просто считаться единой лингвистической единицей, и значение измерений вычисляется в зависимости от того, в каких контекстах эта единица 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 может быть применен один из методов композиции (см. Mitchell & Lapata 2010): 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. Как и операция по уменьшению размерности пространства, процедура композиции позволяет решить проблему недостатка данных и существенно сэкономить время их обработки: 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 (Agirre & Edmonds 2007; Schütze 1998), кластеризации текстов и нахождения документов по запросам пользователей (см., например, Salton 1991),
извлечения отношений (Lin & Pantel 2001) и т.д. Однако в области типологических исследований такого рода модели, насколько нам известно, ещё никем не применялись.

§3. Пилотные эксперименты

Самый начальный шаг в направлении внедрения алгоритмов дистрибутивной семантики в лексическую типологию был сделан в магистерской работе Кюсева 2014, где был представлен следующий эксперимент. По нескольким корпусам текстов на русском языке были собраны вектора для словосочетаний из анкеты, лежавшей в основе типологического анализа семантического поля качественного признака ‘острый’ (ср. острый нож, острыя стрела, острое воспаление...). К тому времени этот признак уже был изучен ранее на материале более 20 языков (см. нашу совместную статью Kyuseva et al. to appear), поэтому для него существует надежная и 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, посвященным качественному признаку ‘острый’, включены в построенную нами типологически ориентированную базу данных признаковой лексики (Кюсева, Резникова, & Рыжова 2013b). В этой базе данных материал представлен в следующем виде: единицей входа является строка анкеты, и для каждого прилагательного, относящегося к рассматриваемому полу, указано, покрывает оно данную строку анкеты или нет. 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 одну из них, и затем полученное число делится на общее количество прилагательных поля ‘острый’, зарегистрированных в Базе (см. Формулу 1).

\[
Sim_{typ} = \frac{(N_{adj}[+a,+b] + N_{adj}[-a,-b])}{(N_{adj}[-a,+b] + N_{adj}[+a,-b] + N_{adj}[+a,+b] + N_{adj}[-a,-b])},
\]

где a, b – пункты анкеты, Nadj – количество прилагательных, покрывающих или не покрывающих соответствующие контексты.

Формула 1. Типологическая близость двух контекстов (= ситуаций).
Так, если в Базе данных нет ни одной лексемы, которая описывает ‘острый гвоздь’, но не употребляется в сочетании с существительным, обозначающим иглу, и наоборот, нет ни одного слова, которое описывало бы ‘острую иглу’, но не ‘острый гвоздь’, то значение метрики для контекстов ‘острый гвоздь’ и ‘острая игла’ будет равняться единице. А если в Базе есть десять слов, которые 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 соответствующего русскоязычного словосочетания (острый гвоздь, острая игла, острый нож и др.) 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 (на основе типологических сведений vs. русскоязычных данных о сочетаемости), поэтому для каждой пары было получено две разных оценки степени близости. Затем для этих двух мер было подсчитано значение коэффициентов корреляции Пирсона и Спирмена, показывающих, есть ли зависимость между двумя величинами, т.е. есть ли взаимосвязь между значениями показателей дистрибутивного и типологического сходства.

В результате экспериментов было получено достаточно высокое значение коэффициентов корреляции (в частности, 0.726 для коэффициента корреляции Пирсона, при том что максимальное значение коэффициента равняется 1), что свидетельствует о наличии зависимости между двумя метриками близости, а значит, по крайней мере, косвенно поддерживает гипотезу о значимости понятия фрейма для лексических систем естественных языков. Однако эти предварительные эксперименты проводились на очень ограниченном типологическом материале, и полученный результат нуждается в проверке, а сама процедура экспериментов – в дополнительных модификациях.

§4. Новая серия экспериментов

В рамках нашей работы мы провели целую серию новых экспериментов, направленных на анализ сопоставимости двух семантических теорий, дополнительную верификацию центрального для фреймового подхода к лексической типологии понятия фрейма, а также подбор оптимальных для наших задач параметров дистрибутивных моделей.
Прежде всего, мы внесли некоторые изменения в разработанную в ходе предварительных экспериментов методологию, и провели новую серию экспериментов на материале семантического поля ‘острый’. Далее, мы апробировали методику на новых тестовых данных, используя результаты типологических исследований качественного признака ‘гладкий’ (Кашкин 2013; Kashkin & Vinogradova to appear) и поля глаголов качания (Шапиро 2015). Помимо этого, в качестве дополнительного эксперимента мы построили векторные модели на основе не русского, а английского корпуса, сопоставив их с результатами типологического исследования поля ‘острый’. И, наконец, мы провели дополнительную серию экспериментов по построению и сопоставлению визуализаций типологических и дистрибутивных пространств. Далее мы подробно обсудим все технические детали наших экспериментов и представим полученные результаты.

1. Подготовка типологических и дистрибутивных данных

1.1. Подготовка типологического векторного пространства

В рамках этой части нашей работы мы проводили эксперименты с рядом признаковых полей (прежде всего, с полями ‘острый’ и ‘гладкий’) и с глагольным полем качания. Данные о лексикализации зоны качания в нескольких языках были нам предоставлены М. М. Шапиро в формате таблицы, в строках которой расположены пункты анкеты, а в столбцах – проанализированные глагольные лексемы. В ячейках таблицы, на пересечении пункта анкеты и конкретного слова, указано, может ли данное слово употребляться в соответствующем контексте (см. фрагмент в Таблице 1). В основу таблицы положена анкета из 30 строк, иллюстрирующих 11 фреймов, релевантных для лексикализации этой зоны (каждый фрейм представлен несколькими примерами). Эта анкета заполнена материалами семи языков (русского, финского, коми-зырянского, ненецкого, японского, хинди и томо-кана), общее число внесенных в таблицу глагольных лексем – 4110.

---

10 Мы благодарим Марию Шапиро за предоставленные данные и помощь с их оформлением.
Сведения о лексикализации признаковых зон ‘острый’ и ‘гладкий’ были нами взяты из типологически ориентированной базы данных признаковой лексики (Кюсева, Резникова, Рыжова 2013b). В Базе собрана информация о 32 прилагательных из 15 языков для поля ‘острый’ и о 32 лексемах из 9 языков для поля ‘гладкий’.

На основе заполненных типологических анкет было сформировано 5 наборов тестовых данных (см. анкеты в Приложениях 1-3):

1) поле ‘острый’, все пункты анкеты;
2) поле ‘острый’, только те пункты анкеты, которые соответствуют фреймам прямых значений;
3) поле ‘гладкий’, вся анкета;
4) поле ‘гладкий’, только прямые значения;
5) поле глаголов качания, только прямые значения.

Сведениями о типологии метафор в поле глаголов качания мы не располагаем.

Предобработка Базы данных признаковой лексики

База данных признаковой лексики имеет следующую структуру. Единица входа в Базу — это микрофрейм, т. е. релевантная для признакового поля ситуация, или пример реализации того или иного фрейма. Поскольку для семантической дизамбигуации прилагательного в большинстве случаев достаточно определяемого существительного, в качестве ярлыка для микрофрейма, как правило, используется словосочетание “прилагательное + существительное”. Для каждого микрофрейма указывается фрейм, к
которому он относится, семантическое поле, которое содержит этот фрейм, а также таксономический класс слов, которые этот микрофрейм реализуют. Кроме того, для каждого фрейма (а, следовательно, и микрофрейма) указано, к каким значениям он относится: исходным (т.е. прямым, буквальным, физическим) или переносным (как правило, абстрактным метафорическим)\(^{11}\).

Для каждой признаковой лексемы такая анкета-заготовка заполняется полностью: в строках, соответствующих тем контекстам, в которых может употребляться данное слово, оно пишется в начальной форме без дополнительных помет; если слово не употребляется в том или ином контексте, то в нужных строках указывается та же лексема в начальной форме, но со знаком «¬»; если данные о поведении слова в указанном контексте отсутствуют, строка не заполняется. Иными словами, анкета для каждого поля повторяется в Базе столько раз, сколько лексем в нее заносится. В отдельной графе указывается язык, к которому относится лексема. Так, фрагмент Базы данных для лексемы зык (поле ‘острый’, бесленеевский диалект кабардино-черкесского языка) имеет следующий вид:

<table>
<thead>
<tr>
<th>Микро-фрейм</th>
<th>Лексема</th>
<th>Язык</th>
<th>Фрейм</th>
<th>Поле</th>
<th>Такс. класс</th>
<th>Тип значения</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>‘острый нож’</td>
<td>з’an</td>
<td>бесленеевский диаглект кабардино-черкесского языка</td>
<td>‘острый инструмент с режущим краем’</td>
<td>‘острый’</td>
<td>‘физ. свойства’</td>
</tr>
<tr>
<td>2</td>
<td>‘острый меч’</td>
<td>з’an</td>
<td>бесленеевский диаглект кабардино-черкесского языка</td>
<td>‘острый инструмент с режущим краем’</td>
<td>‘острый’</td>
<td>‘физ. свойства’</td>
</tr>
<tr>
<td>3</td>
<td>‘острый коготь’</td>
<td>з’an</td>
<td>бесленеевский диаглект кабардино-черкесского языка</td>
<td>‘острый инструмент с режущим краем’</td>
<td>‘острый’</td>
<td>‘физ. свойства’</td>
</tr>
<tr>
<td>4</td>
<td>‘острая сабля’</td>
<td></td>
<td>бесленеевский диаглект кабардино-черкесского языка</td>
<td>‘острый инструмент с режущим краем’</td>
<td>‘острый’</td>
<td>‘физ. свойства’</td>
</tr>
<tr>
<td>5</td>
<td>‘острое лезвие’</td>
<td></td>
<td>бесленеевский диаглект кабардино-черкесского языка</td>
<td>‘острый инструмент с режущим краем’</td>
<td>‘острый’</td>
<td>‘физ. свойства’</td>
</tr>
<tr>
<td>6</td>
<td>‘острая игла’</td>
<td>-з’an</td>
<td>бесленеевский диаглект кабардино-черкесского языка</td>
<td>‘острый инструмент с колющим концом’</td>
<td>‘острый’</td>
<td>‘физ. свойства’</td>
</tr>
<tr>
<td>7</td>
<td>‘острая стрела’</td>
<td></td>
<td>бесленеевский диаглект кабардино-черкесского языка</td>
<td>‘острый инструмент с колющим концом’</td>
<td>‘острый’</td>
<td>‘физ. свойства’</td>
</tr>
<tr>
<td>8</td>
<td>‘острый гвоздь’</td>
<td>-з’an</td>
<td>бесленеевский диаглект кабардино-черкесского языка</td>
<td>‘острый инструмент с колющим концом’</td>
<td>‘острый’</td>
<td>‘физ. свойства’</td>
</tr>
</tbody>
</table>

Таблица 2. Фрагмент Базы данных для лексемы з’an (поле ‘острый’, бесленеевский диалект кабардино-черкесского языка). Знак «—» перед словом означает, что оно не может быть употреблено в данном контексте.

Таблица формата Базы данных признаковой лексики была нами приведена к формату таблицы глаголов качания. Заметим, что данные в таблице с прилагательными получились очень разреженными. Это связано с особенностями заполнения Базы: далеко не для каждого языка была собрана информация по каждому пункту анкеты. В большинстве случаев эксперты по языкам, собирая данные, исходят из теоретического предположения,
что строки анкеты, относящиеся к одному и тому же фрейму, единообразно обслуживаются доступными лексическими средствами. Поэтому в условиях ограниченных ресурсов (отсутствия корпусов достаточного объема или невозможности длительного общения с информантом) эксперт заполняет только часть строк, относящихся к одному фрейму, по умолчанию предполагая, что введенная информация может быть обобщена и на остальные микрофреймы данного фрейма (ср. Таблицу 2).

Разреженность типологических данных, особенно в условиях того, что все наши выборки языков не очень большие, может привести к тому, что метрика типологической близости будет отражать реальное положение дел недостаточно адекватно. Так, например, пункт ‘острый меч’ в анкете для поля ‘острый’ (строка 2 Таблицы 2) очень часто остается незаполненным, поскольку меч – это очень специфическая реальность, известная не всем культурам. Если объяснить носителю, что это такое, он, вероятнее всего, употребит для описания этого предмета то же самое прилагательное, которое используется в его языке для описания остrego ножа, но обычно, конечно, эксперты стараются не использовать такого рода искусственные контексты в полевой работе и просто оставляют соответствующую строку пустой. С другой стороны, бывают случаи, когда за мечами и другими типами устаревшего оружия закрепляются устаревающие прилагательные (как, например, сербское бридак, почти полностью уступившее место прилагательному оштар), т.е. фактически устаревает все словосочетание целиком. В этом случае, поскольку контекст ‘острый меч’ описывается не тем словом, которое используется в других контекстах, относящихся к тому же фрейму, эксперт заполняет соответствующую строку. Типологически такие случаи очень редки, но, так как фактически информация в Базе есть только о них, именно их отразит метрика типологической близости (см. Формулу 1), и контексты ‘острый нож’ и ‘острый меч’ окажутся, согласно этой метрике, значительно дальше друг от друга, чем они есть на самом деле.

Мы приняли решение преодолеть эту проблему, автоматически дозаполнив пустые клетки в Базе. При этом мы опирались на тот же принцип, что и эксперты по языкам: заполняли пустые слоты в соответствии с данными из непустых строк анкеты, относящихся к тому же фрейму (см. в Таблице 3 вид, который принимает после этой операции фрагмент Базы данных из Таблицы 2). После этого преобразования данные снова были приведены к формату таблицы глаголов качания (см. Таблицу 1).
<table>
<thead>
<tr>
<th>Микро-фрейм</th>
<th>Лексема</th>
<th>Язык</th>
<th>Фрейм</th>
<th>Поле</th>
<th>Такс. класс</th>
<th>Тип значения</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>‘острый нож’</td>
<td>ž’an</td>
<td>бесленеевский диалект кабардино-черкесского языка</td>
<td>‘острый инструмент с режущим краем’</td>
<td>‘острый’</td>
<td>‘физ. свойства’</td>
</tr>
<tr>
<td>2</td>
<td>‘острый меч’</td>
<td>ž’an</td>
<td>бесленеевский диалект кабардино-черкесского языка</td>
<td>‘острый инструмент с режущим краем’</td>
<td>‘острый’</td>
<td>‘физ. свойства’</td>
</tr>
<tr>
<td>3</td>
<td>‘острый коготь’</td>
<td>ž’an</td>
<td>бесленеевский диалект кабардино-черкесского языка</td>
<td>‘острый инструмент с режущим краем’</td>
<td>‘острый’</td>
<td>‘физ. свойства’</td>
</tr>
<tr>
<td>4</td>
<td>‘острая сабля’</td>
<td>ž’an</td>
<td>бесленеевский диалект кабардино-черкесского языка</td>
<td>‘острый инструмент с режущим краем’</td>
<td>‘острый’</td>
<td>‘физ. свойства’</td>
</tr>
<tr>
<td>5</td>
<td>‘острое лезвие’</td>
<td>ž’an</td>
<td>бесленеевский диалект кабардино-черкесского языка</td>
<td>‘острый инструмент с режущим краем’</td>
<td>‘острый’</td>
<td>‘физ. свойства’</td>
</tr>
<tr>
<td>6</td>
<td>‘острая игла’</td>
<td>ž’an</td>
<td>бесленеевский диалект кабардино-черкесского языка</td>
<td>‘острый инструмент с колющим концом’</td>
<td>‘острый’</td>
<td>‘физ. свойства’</td>
</tr>
<tr>
<td>7</td>
<td>‘острая стрела’</td>
<td>ž’an</td>
<td>бесленеевский диалект кабардино-черкесского языка</td>
<td>‘острый инструмент с колющим концом’</td>
<td>‘острый’</td>
<td>‘физ. свойства’</td>
</tr>
<tr>
<td>8</td>
<td>‘острый гвоздь’</td>
<td>ž’an</td>
<td>бесленеевский диалект кабардино-черкесского языка</td>
<td>‘острый инструмент с колющим концом’</td>
<td>‘острый’</td>
<td>‘физ. свойства’</td>
</tr>
</tbody>
</table>

Таблица 3. Фрагмент автоматически дозаполненной Базы данных для лексемы ž’an (поле ‘острый’, бесленеевский диалект кабардино-черкесского языка). Знак ‘-’ перед словом означает, что оно не может быть употреблено в данном контексте. Полужирным выделены слова, добавленные автоматически.

**Метрика типологической близости**

Автоматическое дозаполнение таблицы с типологическими данными позволило нам также видоизменить формулу вычисления степени типологической близости,
максимально приблизив ее к используемой нами формуле определения степени дистрибутивного сходства.

В результате предобработки Базы данных мы получаем таблицу, в которой каждая строка анкеты рассматриваемого поля может быть представлена в виде вектора $w$ (что было невозможно раньше, когда не все клетки типологических таблиц были заполнены). Измерениями этого вектора будут внесенные в Базу лексемы, а значение $i$-ого измерения будет вычисляться следующим образом:

$$w_i = \begin{cases} 1, & \text{если строка } w \text{ покрывается лексемой } l_i; \\ 0, & \text{если строка } w \text{ не покрывается лексемой } l_i \end{cases}$$

Формула 2. Вычисление значения измерений типологического вектора

Так, например, вектор для первой строки (‘гибкие стебли’ – ‘цветок’) из Таблицы 1 будет иметь вид {0, 1, 1, 0}. Поскольку теперь типологическая информация имеет векторное представление, мы можем использовать в качестве метрики типологической близости ($TC$) стандартную меру расстояния между векторами любого типа – косинусное сходство (см. Jurafsky & Martin 2008, а также Формулу 3).

$$\text{similarity}(A, B) = \cos(\theta) = \frac{A \cdot B}{\| A \| \| B \|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

Формула 3. Косинусная мера близости между векторами.

1.2. Подготовка дистрибутивного векторного пространства

С типологическим пространством мы сопоставляли дистрибутивное, которое строилось следующим образом. Для каждого пункта анкеты было подобрано максимально соответствующее ему русское словосочетание вида «прилагательное + существительное»: острый нож для пункта ‘острый нож’, острая стрела – ‘острая стрела’ и т.д. Для каждого полученного словосочетания был построен вектор его сочетаемости.

Дистрибутивную модель можно построить множеством способов, зависящих от значений целого ряда параметров. Мы провели серию экспериментов по сопоставлению значения метрики ТС (которая вычисляется однозначно) с различными значениями метрики дистрибутивной близости, вычисляемых для моделей разной конфигурации.
Некоторые настройки оставались неизменными во всех экспериментах. Так, в качестве измерений для векторов встречаемости мы использовали 10 000 самых частотных лемм самостоятельных частей речи (по основному подкорпусу НКРЯ), а значением каждого измерения считали абсолютную частоту встречаемости слова-измерения на расстоянии ±5 знаменательных слов от опорной единицы. Расстояния между векторами всегда определялись с помощью косинусной меры близости (см. Формулу 3 выше), и именно поэтому мы приняли решение использовать ту же метрику для определения степени близости между типологическими векторами.

Значения всех остальных параметров варьировались. К числу таких параметров относятся: обучающий корпус, дополнительная обработка векторного пространства, тип вектора словосочетания. Рассмотрим подробнее каждый из них.

**Обучающий корпус.**

Для обучения моделей мы использовали три русскоязычных корпуса в разных комбинациях и один дополнительный корпус, составленный из текстов на английском языке:

1) основной подкорпус НКРЯ (общим объемом около 200 млн словоупотреблений);  
2) газетный подкорпус НКРЯ (около 150 млн словоупотреблений);  
3) ruWaC (около 1 млрд словоупотреблений);  
4) ukWaC (около 2 млрд словоупотреблений).

Можно заметить, что это корпуса разных объемов и разных жанров. Основной подкорпус НКРЯ не очень большой, но сбалансированный, газетный подкорпус сопоставимого размера, но содержит только тексты публицистического характера. Эти коллекции текстов были предоставлены нам в 2013 году в неразмеченном виде. Мы аннотировали их с помощью морфологического парсера Mystem2 и провели морфологическую дизамбигуацию с помощью инструмента, представленного в работе (Лакомкин и др. 2013). Тем самым, наши корпуса несколько отличаются по объему и разметке от основного и газетного подкорпусов НКРЯ, поиск по которым доступен на официальном сайте корпуса [https://ruscorpora.ru], в частности, процент ошибок на уровне морфологической аннотации в нашем варианте несомненно выше.

RuWaC значительно больше подкорпусов НКРЯ по объему и состоит в основном из интернет-текстов. UkWaC аналогичен ruWaC-у по жанру, но состоит из текстов не на русском, а на английском языке. Оба корпуса снабжены морфологической и синтаксической разметкой (см. Baroni et al. 2009).
Дополнительная обработка векторов.

Изначально вектор сочетаемости той или иной лингвистической единицы, собранный по нашим параметрам, состоит из 10 000 целых чисел, соответствующих абсолютной частоте встречаемости каждого слова-измерения в контексте слова (или словосочетания), для которого строится вектор. Очевидно, что такие значения измерений отражают не только сочетаемостные предпочтения опорного слова, но и общую частотность самого слова-измерения. Так, например, значение измерения, соответствующего лемме делать, практически для любой опорной единицы будет заведомо выше, чем значение измерения, в качестве которого выступает лемма половник. Чтобы нивелировать этот эффект и учесть частотность как опорного слова, так и слов-измерений, используют несколько техник взвешивания вектора, в большинстве своем основанных на функции взаимной информации (mutual information), см. Martin & Jurafsky 2008.

Мы использовали четыре модели взвешивания:

1) Positive Point-wise Mutual Information (PPMI)

\[ pmi(r,c) = \log \frac{P(r,c)}{P(r)P(c)} \]

\[ pppmi(r,c) = pmi(r,c), \text{если } pmi(r,c) \geq 0 \]

\[ pppmi(r,c) = 0, \text{если } pmi(r,c) < 0 \]

Формула 4. Схема взвешивания PPMI

2) Positive Local Mutual Information (PLMI)

\[ plmi(r,c) = pppmi(r,c) \times \text{count}(r,c) \]

Формула 5. Схема взвешивания PLMI

3) Exponential Point-wise Mutual Information (EPMI)

\[ epmi(r,c) = \frac{P(r,c)}{P(r)P(c)} \]

Формула 6. Схема взвешивания EPMI
4) Positive Log Weighting

\[ plo{g}(r, c) = \log \text{count}(r, c), \text{если } \log \text{count}(r, c) \geq 0 \]

\[ plo{g}(r, c) = 0, \text{если } \log \text{count}(r, c) < 0 \]

Формула 7. Схема взвешивания PLOG

Второе преобразование векторного пространства, которое мы использовали, – уменьшение его размерности. Пространство, которое мы получаем на первом шаге, в результате сбора векторов сочетаемости, представляет собой матрицу размером N x 10 000, где N – количество опорных единиц, для которых строятся вектора, а 10 000 – количество слов-измерений. В таких матрицах очень много нулейых значений. Чтобы понизить их долю, используют технику разложения матриц, подразумевающую представление исходной матрицы в виде произведения матриц меньшей размерности. В нашем исследовании мы используем сингулярное разложение (singular-value decomposition, SVD), с помощью которого сокращаем размерность дистрибутивного пространства до 300 измерений. Предполагается, что это может, во-первых, сократить время обработки данных в ходе последующих манипуляций с ними, а во-вторых, повысить качество модели за счет объединения похожих измерений.

Тип вектора словосочетания.

Поскольку в наших экспериментах типологическим расстояниям между пунктами анкеты должны соответствовать дистрибутивные расстояния между словосочетаниями, а не отдельными словами, возникает необходимость подбора оптимального метода подготовки векторного представления для словосочетания. Есть два основных способа построения вектора для единицы более длинной, чем слово. С одной стороны, можно рассматривать словосочетание как единое целое и вычислять значения измерений по контекстам, в которых оно встречается (такой вектор называется наблюдаемым – observed). В этом случае исследователь неминуемо сталкивается с проблемой нехватки данных: словосочетания значительно менее частотные, чем слова, поэтому для качественного представления их дистрибутивных свойств нужны корпуса очень больших размеров. С другой стороны, вектор словосочетания можно строить методом компонирования векторов его элементов, т.е. сначала собирать отдельные вектора для прилагательного и для существительного, а затем их объединять. Существует несколько стандартных моделей вычисления результирующего векторного представления словосочетания на

1. Аддитивная (additive) и аддитивная взвешенная (weighted additive).

Эта схема компонозиции подразумевает сложение векторов прилагательного и существительного (т.е. попарное суммирование значений по каждому из измерений) с присвоением слагаемым некоторых весов. Если применяется простейшая модель без взвешивания, значение коэффициентов приравнивается к 1. Для взвешенной модели значения весовых коэффициентов вычисляются на основе обучающего корпуса – набора векторов соответствующих наблюдаемых словосочетаний. В процессе обучения минимизируется евклидова норма разности между векторами, сгенерированными моделью, и векторами из обучающего корпуса.

\[ \mathbf{p} = \alpha \mathbf{u} + \beta \mathbf{v}, \]

где \( \mathbf{p} \) – вектор словосочетания, \( \mathbf{u} \) - вектор прилагательного, \( \mathbf{v} \) - вектор существительного, \( \alpha \) и \( \beta \) – весовые коэффициенты.

Формула 8. Аддитивная модель композиции.

2. Мультипликативная (multiplicative).

Эта модель аналогична аддитивной, только вместо суммирования значений по каждому измерению применяется их попарное перемножение.

\[ \mathbf{p} = \mathbf{u} \odot \mathbf{v}, \]

где \( \mathbf{p} \) – вектор словосочетания, \( \mathbf{u} \) - вектор прилагательного, \( \mathbf{v} \) - вектор существительного. Символ \( \odot \) используется для обозначения операции попарного перемножения значений измерений двух векторов.

Формула 9. Мультипликативная модель композиции.

3. Расширительная (dilation).

В этой модели один из векторов (в нашем случае, вектор существительного) раскладывается на два компонента: параллельный второму вектору (вектору прилагательного, см. Формулу 10а) и ортогональный ему (Формула 10б). Затем параллельному компоненту дается больший вес с помощью коэффициента \( \lambda \), а ортогональный компонент остается неизменным (Формула 10в). Наконец, каждое слагаемое из уравнения 10в домножается на скалярный квадрат вектора прилагательного (Формула 10г). Итоговый модифицированный вектор существительного используется в
качестве векторного представления всего словосочетания. Значение коэффициента $\lambda$ подбирается так же, как и коэффициенты для аддитивной взвешенной модели (см. выше). В расширительной модели без тренировки для коэффициента $\lambda$ выбирается значение 2.

$$\hat{x} = \frac{\mathbf{u} \cdot \mathbf{v}}{\mathbf{u} \cdot \mathbf{u}},$$

где $\mathbf{u}$ - вектор прилагательного, $\mathbf{v}$ - вектор существительного.

Формула 10а. Компонент вектора $\hat{\mathbf{v}}$, параллельный вектору $\mathbf{u}$.

$$\hat{y} = \mathbf{v} - \hat{x} = \mathbf{v} - \frac{\mathbf{u} \cdot \mathbf{v}}{\mathbf{u} \cdot \mathbf{u}} \mathbf{u}$$

Формула 10б. Компонент вектора $\mathbf{v}$, ортогональный вектору $\mathbf{u}$.

$$\mathbf{v}' = \lambda \hat{x} + \hat{y} = \lambda \frac{\mathbf{u} \cdot \mathbf{v}}{\mathbf{u} \cdot \mathbf{u}} \mathbf{u} + \mathbf{v} - \frac{\mathbf{u} \cdot \mathbf{v}}{\mathbf{u} \cdot \mathbf{u}} \mathbf{u} = (\lambda - 1) \frac{\mathbf{u} \cdot \mathbf{v}}{\mathbf{u} \cdot \mathbf{u}} \mathbf{u} + \mathbf{v},$$

где $\lambda$ - скалярная величина, вычисляемая на основе обучающего корпуса (векторов наблюдаемых словосочетаний).

Формула 10в. Модифицированный вектор существительного.

$$\hat{p} = (\lambda - 1)(\mathbf{u} \cdot \mathbf{v}) \mathbf{u} + (\mathbf{u} \cdot \mathbf{u})\mathbf{v},$$

Формула 10г. Вектор словосочетания: расширительная модель композиции.

4. Лексическая функция\(^{12}\) (lexical function) и практическая лексическая функция (practical lexical function, PLF).

В моделях лексической функции и практической лексической функции существительное представляется в виде вектора, а прилагательное – в виде матрицы, т.е. прилагательное выступает в роли функции, которая может так или иначе влиять на сочетаемость существительного. Матрица для признакового слова рассчитывается на основе наблюдаемых векторов словосочетаний с участием данного прилагательного. Модель практической лексической функции представляет собой сумму результата лексической функции и вектора прилагательного (см. Формулы 11 и 12, а также Paperno, Pham, & Baroni 2014).

\(^{12}\) Название этой модели неудачно совпадает с широко известным в лингвистике термином лексическая функция, введённым А.К. Жолковским и И.А. Мельчуком (см. Мельчук, Жолковский 1984) для описания нетривиальной сочетаемости слов, но не имеет с ним практически ничего общего.
\[ \vec{p} = U\vec{v}, \]
где \( \vec{p} \) – это вектор словосочетания, \( U \) – матричное представление прилагательного (т.е. собственно «лексическая функция»), \( \vec{v} \) – вектор существительного.
Формула 11. Лексическая функция.

\[ \vec{p} = U\vec{v} + \vec{u}, \]
где \( \vec{p} \) – это вектор словосочетания, \( U \) – матричное представление прилагательного, \( \vec{v} \) – вектор существительного, \( \vec{u} \) – вектор прилагательного.
Формула 12. Практическая лексическая функция.


1.3. Подсчет корреляции

Чтобы определить степень сопоставимости двух семантических пространств (типологического и дистрибутивного), мы берем декартово произведение анкеты самой на себя (т.е. множество всех возможных пар строк из этой анкеты) и вычисляем для каждой пары значение обеих метрик близости (см. Таблицу 4), а затем вычисляем значение коэффициента корреляции для метрик типологического и дистрибутивного сходства. Поскольку мы ожидаем, что связь между двумя метриками должна быть линейной, мы используем коэффициент корреляции Пирсона (см. Формулу 13).

Формула 13. Коэффициент корреляции Пирсона.

\[ r_{xy} = \frac{\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{m}(x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{m}(y_i - \bar{y})^2}} = \frac{cov(x, y)}{\sqrt{s_x^2 s_y^2}} \]
где \( x^m = (x_1, ..., x_m), y^m = (y_1, ..., y_m) \) – две выборки, \( \bar{x}, \bar{y} \) – выборочные средние \( x^m \) и \( y^m \), \( s_x^2, s_y^2 \) – выборочные дисперсии, \( r_{xy} \in [-1, 1] \).
Формула 13. Коэффициент корреляции Пирсона.
<table>
<thead>
<tr>
<th>Пары пунктов анкеты</th>
<th>Типологическая близость (выборка $x^m$)</th>
<th>Дистрибутивная близость (выборка $y^m$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘острый нож’ vs. ‘острая коса’</td>
<td>0.961</td>
<td>0.923</td>
</tr>
<tr>
<td>‘острый нож’ vs. ‘острый нос’</td>
<td>0.484</td>
<td>0.910</td>
</tr>
<tr>
<td>‘острый нож’ vs. ‘острое зрение’</td>
<td>0.615</td>
<td>0.907</td>
</tr>
</tbody>
</table>

Таблица 4. Иллюстрация пары выборок, для которых подсчитывался коэффициент корреляции Пирсона (поля ‘острый’). Данные о дистрибутивной близости приведены для модели, построенной с помощью практической лексической функции.

1.4. Базовый алгоритм

В качестве отправной точки для оценки результатов мы использовали самую простую дистрибутивную модель, в которой вектора сочетаемости вычислялись не для словосочетаний, а только для существительных, без учета прилагательных (ср. нож вместо острый нож, стрела вместо острая стрела и т.д.). Это дистрибутивное пространство сопоставлялось с неизменным типологическим.

Предполагалось, однако, что уровень корреляции между такими пространствами будет достаточно высоким, поскольку, как правило, семантика существительного в большой мере определяет значение прилагательного, а значит, и связь соответствующего контекста с тем или иным фреймом (см. Рахилина 2010, а также попытки использовать это явление для решения задачи автоматической семантической дизамбигуации НКРЯ, описанные в Шеманаева и др. 2007).

2. Эксперимент 1: признаковые поля ‘острый’ и ‘гладкий’

В первой серии экспериментов мы использовали в качестве тестовых данные о лексикализации признаков ‘острый’ и ‘гладкий’ в нескольких языках и статистику сочетаемости соответствующих русских словосочетаний вида «прилагательное + существительное».

Для лексикализации прямых значений поля ‘острый’ типологически релевантны два основных противопоставления:

---

13 Основные результаты этого эксперимента отражены в публикации Ryzhova et al. 2016.
1) острота, воспринимаемая тактильно vs. острота, воспринимаемая зрительно (острые инструменты vs. объекты вытянутой формы);

2) острый край vs. острый кончик (режущие инструменты vs. колющие инструменты и объекты вытянутой формы). 

Так, например, французское прилагательное tranchant описывает только режущие инструменты (ножи, лезвия и т.п.), в то время как китайская лексема jianrui сочетается только с наименованиями колющими инструментов (таких, как ‘стрел’, ‘копье’, ‘шило’), а японское слово togatta описывает только объекты острой формы (ср. ‘носок ботинка’), подробнее см. Kyuseva et al. to appear. Между тем, в русском языке на лексическом уровне эти фреймы не противопоставлены: прилагательное острий может описывать и режущие (острий нож), и колющие (острая игла) инструменты, а также объекты вытянутой формы (ботинки с острыми носами). И если последний тип употреблений обычно выносится в отдельное значение в рамках словарных статей для слова острий, то значения функциональной остроты колющих и режущих инструментов в русской лексикографической традиции практически никогда не противопоставляются.

В поле отсутствия неровностей на поверхности объекта (‘гладкий’) тоже выделяется два базовых противопоставления, дающих три основных фрейма:

1) тактильное vs. зрительное восприятие неровностей (ср. гладкая/скользкая доска vs. ровная стена);

2) «функциональное» отсутствие неровностей, приятное или удобное (может оцениваться как зрительно, так и тактильно) vs. отсутствие неровностей, приводящее к потере функциональности (ср. гладкая кожа / ровный пол vs. скользкий пол).

Для некоторых лексических систем релевантна только первая оппозиция. В таких языках выделяется лексически ситуация описания поверхностей, у которых отсутствие неровностей оценивается зрительно, а гладкие и скользкие поверхности (т.е. такие, чья неровность определяется тактильно) описываются одной и той же лексемой. Такой тип лексикализации засвидетельствован, например, в эрзянском и ненецком языках. В других языках, напротив, действует только второе противопоставление, и тогда для описания гладких и ровных поверхностей выбирается одно и то же слово, а для «нефункциональных» скользких – другое. Именно такая система представлена, например, в марийском (и – с некоторыми оговорками – в китайском), подробнее о типологии

14 Кажущаяся логически возможной комбинация ‘линия’ + ‘форма’ не лексикализована ни в одном языке выборки.

(1) У них было свое дерево, платан, и перед тем как идти спать, они проводили пальцами по его гладкой коже — воздух в темноте свежел, а она оставалась теплая. [Михаил Шишкин. Венерин волос (2004) // «Знамя», 2005]

(2) Все пригорочки, холмики и впадинки будут срезаны и слажены, вместо них появится ровная поверхность газона с травой регламентированного зеленого цвета и заданной высоты. [неизвестный. Слухи про неё верны // «Русская жизнь», 2012]

(3) Она все-таки упала, не удержавшись на скользком полу, и больно ударилась коленкой. [Наталья Александрова. Последний ученик да Винчи (2010)]

Таким образом, два выбранных нами поля сопоставимы по уровню сложности фреймовых структур, лежащих в основании концептуализации этих зон в языках мира. С другой стороны, в русском языке стратегии лексикализации этих полей существенно различаются: в поле ‘острый’ одно доминантное прилагательное покрывает все три фрейма, в то время как в поле ‘гладкий’ функционируют три лексемы, эксплицитно противопоставляющие три основных значения в этой признаковой области.

Для каждого из полей мы провели полный цикл экспериментов, варьируя все параметры, упомянутые в разделе 1 настоящей главы. Для обоих полей максимальный коэффициент корреляции получился очень высоким (0.766 для поля ‘острый’ и 0.946 для поля ‘гладкий’). Обращает на себя внимание тот факт, что максимальное значение коэффициента корреляции при базовом алгоритме оказывается неожиданно низким: 0.268 (‘острый’) и 0.244 (‘гладкий’).

Особенно важно отметить, что лучшие результаты для двух признаковых зон были получены на одних и тех же настройках: в качестве обучающего корпуса использовался основной подкорpus НКРЯ, вектора взвешивались по схеме PPMI, размерность итогового векторного пространства сокращалась до 300, а вектора словосочетаний складывались из векторных представлений составляющих их слов с помощью модели композиции PLF (практическая лексическая функция). Заметим также, что лучшие результаты получены на материале только прямых значений рассматриваемых признаков. Учет метафорических
фреймов существенно снижает показатели: коэффициент корреляции Пирсона для поля ‘острый’ в этом случае (при прочих равных условиях) равняется 0.462, для зоны ‘гладкий’ – 0.604, что, по-видимому, означает, что прямые значения обладают более четкой, а главное, предсказуемой фреймовой структурой, чем переносные, которые, хотя и являются мотивированными, охватывают материал конкретного языка менее равномерно. Наиболее показательные результаты приведены в Таблице 5.

Таблица 5. Коэффициенты корреляции Пирсона для дистрибутивных моделей разных конфигураций и разных тестовых данных (поля ‘острый’ и ‘гладкий’). Все результаты статистически значимы (p-value < 0.01 во всех случаях). Полужирным выделены лучшие результаты.

<table>
<thead>
<tr>
<th>корп.</th>
<th>обработка вектора</th>
<th>модель композиции</th>
<th>‘острый’</th>
<th>‘острый’: прямые</th>
<th>‘гладкий’</th>
<th>‘гладкий’: прямые</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>НКРЯ none</td>
<td>noun only</td>
<td>0.092</td>
<td>0.12</td>
<td>0.121</td>
<td>0.196</td>
</tr>
<tr>
<td>2</td>
<td>НКРЯ ppmi</td>
<td>noun only</td>
<td>0.139</td>
<td>0.167</td>
<td>0.237</td>
<td>0.21</td>
</tr>
<tr>
<td>3</td>
<td>НКРЯ ppmi,SVD</td>
<td>noun only</td>
<td>0.167</td>
<td>0.268</td>
<td>0.274</td>
<td>0.244</td>
</tr>
<tr>
<td>4</td>
<td>НКРЯ ppmi</td>
<td>none</td>
<td>0.097</td>
<td>0.194</td>
<td>0.134</td>
<td>0.154</td>
</tr>
<tr>
<td>5</td>
<td>НКРЯ ppmi</td>
<td>additive</td>
<td>0.36</td>
<td>0.654</td>
<td>0.589</td>
<td>0.74</td>
</tr>
<tr>
<td>6</td>
<td>НКРЯ ppmi</td>
<td>Multiplicative</td>
<td>0.253</td>
<td>0.421</td>
<td>0.585</td>
<td>0.7</td>
</tr>
<tr>
<td>7</td>
<td>НКРЯ ppmi</td>
<td>Dilation w/Training</td>
<td>0.19</td>
<td>0.222</td>
<td>0.379</td>
<td>0.443</td>
</tr>
<tr>
<td>8</td>
<td>НКРЯ ppmi</td>
<td>Dilation w/Training</td>
<td>0.207</td>
<td>0.35</td>
<td>0.249</td>
<td>0.313</td>
</tr>
<tr>
<td>9</td>
<td>НКРЯ ppmi,SVD</td>
<td>LexFunc</td>
<td>0.116</td>
<td>0.345</td>
<td>0.443</td>
<td>0.703</td>
</tr>
<tr>
<td>10</td>
<td>НКРЯ ppmi,SVD</td>
<td>PrLexFunc</td>
<td>0.39</td>
<td>0.766</td>
<td>0.449</td>
<td><strong>0.946</strong></td>
</tr>
<tr>
<td>11</td>
<td>НКРЯ none</td>
<td>WeightedAdd</td>
<td>0.443</td>
<td>0.754</td>
<td>0.589</td>
<td>0.849</td>
</tr>
<tr>
<td>12</td>
<td>НКРЯ plog</td>
<td>WeightedAdd</td>
<td>0.387</td>
<td>0.76</td>
<td>0.477</td>
<td>0.765</td>
</tr>
<tr>
<td>13</td>
<td>НКРЯ epmi</td>
<td>WeightedAdd</td>
<td><strong>0.462</strong></td>
<td>0.763</td>
<td>0.59</td>
<td>0.865</td>
</tr>
<tr>
<td>14</td>
<td>НКРЯ ppmi</td>
<td>WeightedAdd</td>
<td>0.42</td>
<td>0.764</td>
<td><strong>0.604</strong></td>
<td>0.905</td>
</tr>
<tr>
<td>15</td>
<td>НКРЯ plmi</td>
<td>WeightedAdd</td>
<td>0.443</td>
<td>0.762</td>
<td>0.603</td>
<td>0.791</td>
</tr>
<tr>
<td>16</td>
<td>все ppmi</td>
<td>WeightedAdd</td>
<td>0.418</td>
<td>0.764</td>
<td>0.564</td>
<td>0.899</td>
</tr>
<tr>
<td>17</td>
<td>все plmi</td>
<td>WeightedAdd</td>
<td>0.438</td>
<td>0.763</td>
<td>0.549</td>
<td>0.712</td>
</tr>
<tr>
<td>18</td>
<td>НКРЯ ppmi,SVD</td>
<td>Additive</td>
<td>0.269</td>
<td>0.443</td>
<td>0.404</td>
<td>0.566</td>
</tr>
<tr>
<td>19</td>
<td>НКРЯ ppmi,SVD</td>
<td>Dilation w/Training</td>
<td>0.388</td>
<td>0.766</td>
<td>0.448</td>
<td>0.936</td>
</tr>
<tr>
<td>20</td>
<td>НКРЯ ppmi,SVD</td>
<td>WeightedAdd</td>
<td>0.388</td>
<td>0.717</td>
<td>0.421</td>
<td>0.682</td>
</tr>
<tr>
<td>21</td>
<td>НКРЯ ppmi,SVD</td>
<td>Dilation</td>
<td>0.231</td>
<td>0.519</td>
<td>0.374</td>
<td>0.512</td>
</tr>
<tr>
<td>22</td>
<td>НКРЯ ppmi,SVD</td>
<td>Multiplicative</td>
<td>0.062</td>
<td>0.41</td>
<td>0.194</td>
<td>0.228</td>
</tr>
</tbody>
</table>

Таким образом, вопреки распространенному мнению (см., например, Bullinaria & Levy 2012), увеличение объема обучающего корпуса не всегда ведет к повышению качества модели: в нашем случае отдельно взятый основной подкорпус НКРЯ дает более высокий (или, по меньшей мере, сопоставимый) результат, чем объединенный корпус общим объемом около 1,35 млрд словоупотреблений, включающий основной и газетный подкорпусы НКРЯ и корпус ruWaC (ср. аналогичное наблюдение в работе Kutuzov & Kuzmenko 2015, где утверждается, что основной подкорпус НКРЯ достаточно представителен для анализа лексической семантики русских слов).
Для создания качественного векторного представления отдельных лемм основного подкорpusa НКРЯ достаточно. Что же касается профиля сочетаемости двусловных сочетаний, то для решения этой задачи даже объединенный обучающий корпус оказывается мал: применение любой модели композиции существенно улучшает результат по сравнению с использованием наблюдаемых векторов словосочетаний. При этом наши результаты не позволяют с уверенностью говорить о том, какая именно модель композиции лучше всего подходит для нашей задачи. Практически столь же высокий, что и для модели PLF, коэффициент корреляции между двумя наборами данных наблюдается при применении взвешенной аддитивной модели без уменьшения размерности дистрибутивного пространства (см. строки 11-15), а также расширительной (dilation) модели с тренировкой и сокращением размерности (см. строка 19). И аддитивная, и расширительная схемы при этом экономнее, чем модель практической лексической функции, для обучения которой требуется больше данных (и, соответственно, времени на их обработку).

Не удается однозначно определить и лучшую схему предварительного взвешивания векторного пространства: все модели, представляющие собой различные вариации метрики взаимной информации, работают примерно одинаково – и существенно лучше, чем простая логарифмическая метрика plog (ср. строки 13, 14, 15 vs. 12).

Однако эта серия экспериментов позволяет сделать несколько важных предварительных выводов. Во-первых, при определенных настройках значение коэффициента корреляции оказывается очень высоким. Содержательно это говорит о том, что типологическое и дистрибутивное расстояние взаимосвязаны, т.е. если два русских словосочетания, иллюстрирующие определенные значения прилагательных из поля 'острый' или 'гладкий', употребляются в похожих контекстах / ситуациях, то велика вероятность, что для их перевода на некоторый другой язык будет использована одна и та же признаковая лексема. Иными словами, чем выше дистрибутивное сходство между словосочетаниями, тем больше вероятность, что они относятся к одному и тому же фрейму. В свою очередь, это означает, что фреймовая структура, которую мы обычно выявляем в ходе подробного анализа данных нескольких языков, так или иначе находит отражение в сочетаемостных характеристиках слов и словосочетаний одного конкретного языка.

Во-вторых, типологические данные, собранные вручную, вероятно, можно считать золотым стандартом и использовать их для оценки качества дистрибутивных моделей. В-третьих, оптимальные параметры дистрибутивной модели, вероятно, можно будет
учитывать в дальнейшем при использовании этой техники для решения задач автоматического анализа лексики.

3. Эксперимент 2: глаголы качания

Эксперименты с признаковыми полями ‘острый’ и ‘гладкий’ показали очень высокий коэффициент корреляции между типологическим и дистрибутивным пространствами. Однако можно предположить, что успех связан с тем, что поля качественных признаков устроены довольно просто: каждое из них сводится всего к двум базовым противопоставлениям, дающим три основных фрейма в прямых значениях, при этом в поле ‘гладкий’ эти фреймы ещё и эксплицитно противопоставлены в русском языке на лексическом уровне. В качестве дополнительной проверки значимости полученных результатов мы провели дополнительный эксперимент на материале глагольного поля качания, характеризующегося значительно более сложной фреймовой структурой.

Материал по глаголам качания, которым мы располагаем, показывает, что эта зона в среднем значительно богаче лексически, чем признаковые поля ‘острый’ и ‘гладкий’: во всех языках выборки выделяется от 4 до 15 слов с семантикой качания / колебания. Эти лексемы кодируют несколько противопоставлений. Наиболее существенную роль в лексикализации таких ситуаций играет тип качающегося объекта, определяющий характер его движения и, соответственно, выбор лексического средства для его описания. Объект может быть вертикально или горизонтально ориентированным (ср. занавеска vs. поверхность воды или поросшего высокой травой луга), прикрепленным сверху или снизу (маятник vs. дерево), жестким или гибким (дерево vs. цветок). Кроме того, могут быть противопоставлены лексически движение, связанное с нормальным функционированием объекта (ср. качание маятника, качелей, кресла-качалки), и движение, вызванное нарушением целостности объекта (ср. шатающийся стул, болтающаяся ножка стола), – что, опять же, определяется типом объекта (его исходным предназначением).

В русском языке выделяется 6 глаголов качания (качаться, колыхаться, колебаться, развеваться, болтаться, шататься), маркирующих только часть из всех возможных противопоставлений. Доминантный глагол качаться покрывает все фреймы поля, кроме ситуаций потери функциональности (нарушения целостности объекта) и колыхания

15 Заметим, что ситуации каузированного движения (ср. ‘раскачивать качели’) были полностью исключены из рассмотрения авторами типологического исследования глаголов качания, см. Шапиро 2015. Соответственно, и наши эксперименты затрагивают только некаузативный фрагмент данного семантического поля.
гибких / мягких объектов. Зону потери функциональности обслуживают глаголы болтаться и шататься, а зону колебательного движения гибких объектов – лексемы колебаться и колыхаться. Наконец, последний глагол – развеваться – имеет очень узкую сферу употребления и описывает преимущественно колебание мягких объектов (прототипически – флага) на ветру.

Поскольку все противопоставления в этой зоне основаны на типе движущегося объекта, в качестве диагностических контекстов в анкете фигурируют словосочетания вида «существительное + глагол» (ср. маятник качается, зуб шатается), содержащие два лексических элемента, как и контексты для прилагательных, что позволяет говорить о сопоставимых условиях признакомого и глагольного экспериментов.

На материале глаголов качания мы провели редуцированную серию экспериментов: протестировали только такие дистрибутивные модели, которые требуют небольших объемов тренировочных данных и наименьших временных затрат. Тем самым, в качестве обучающего корпуса мы использовали только НКРЯ, вектора словосочетаний строили только с использованием моделей композиции, а из возможных моделей композиции выбрали только варианты, не предполагающие никакого обучения: аддитивную, мультипликативную и расширительную (dilation) схемы. Из возможных типов операций над векторами в этой серии экспериментов мы по-прежнему варьировали параметр уменьшения размерности (либо без уменьшения, либо с уменьшением до 300 измерений по технике SVD) и параметр взвешивания векторов (сведя его к двум значениям: без взвешивания vs. со взвешиванием по схеме PPMI). Напомним также, что мы располагаем типологическими данными только для прямых значений этой семантической зоны. Результаты этой серии экспериментов представлены в Таблице 6.
Из Таблицы 6 видно, что максимальный уровень корреляции между двумя наборами данных по-прежнему очень высок (0.7). Это значение дает сопоставление типологического пространства с дистрибутивным, построенным с использованием схемы взвешивания PPMI и аддитивной модели композиции. Подчеркнем, что этот результат соответствует полученому на материале признаковых полей при тех же настройках: для поля ‘острый’ коэффициент корреляции Пирсона при таких же экспериментальных условиях равнялся 0.654, для поля ‘гладкий’ – 0.74 (см. строку 5 Таблицы 5). Это подтверждает предварительные выводы, сделанные нами по результатам предыдущей серии экспериментов, а также позволяет предположить, что более сложные модели композиции могли бы позволить повысить коэффициент корреляции между двумя семантическими представлениями глаголов качания (как это произошло в случае с признаковыми полями). Важно, однако, особо отметить, что приемлемого результата можно добиться и на максимально простых («экономных») настройках, не требующих ни корпуса очень большого объема, ни дополнительного обучения моделей композиции.

Интересно при этом, что значение коэффициента корреляции при базовом алгоритме (с векторами сочетаемости существительных вместо словосочетаний) в случае глаголов качания оказывается очень высоким (0.539 при оптимальных настройках) и существенно превышает соответствующие значения для признаковых полей (0.167 – ‘острый’ и 0.21 –

<table>
<thead>
<tr>
<th>Корпус</th>
<th>Операции над векторами</th>
<th>Модель композиции</th>
<th>Коэффициент корреляции Пирсона</th>
</tr>
</thead>
<tbody>
<tr>
<td>НКРЯ</td>
<td>none</td>
<td>noun only</td>
<td>0.394</td>
</tr>
<tr>
<td>НКРЯ</td>
<td>ppmi</td>
<td>noun only</td>
<td>0.539</td>
</tr>
<tr>
<td>НКРЯ</td>
<td>ppmi, SVD</td>
<td>noun only</td>
<td>0.464</td>
</tr>
<tr>
<td>НКРЯ</td>
<td>none</td>
<td>additive</td>
<td>0.434</td>
</tr>
<tr>
<td>НКРЯ</td>
<td>none</td>
<td>multiplicative</td>
<td>0.443</td>
</tr>
<tr>
<td>НКРЯ</td>
<td>none</td>
<td>dilation</td>
<td>0.57</td>
</tr>
<tr>
<td>НКРЯ</td>
<td>ppmi</td>
<td>additive</td>
<td>0.7</td>
</tr>
<tr>
<td>НКРЯ</td>
<td>ppmi</td>
<td>multiplicative</td>
<td>0.648</td>
</tr>
<tr>
<td>НКРЯ</td>
<td>ppmi</td>
<td>dilation</td>
<td>0.691</td>
</tr>
<tr>
<td>НКРЯ</td>
<td>ppmi, SVD</td>
<td>additive</td>
<td>0.643</td>
</tr>
<tr>
<td>НКРЯ</td>
<td>ppmi, SVD</td>
<td>multiplicative</td>
<td>0.267</td>
</tr>
<tr>
<td>НКРЯ</td>
<td>ppmi, SVD</td>
<td>dilation</td>
<td>0.647</td>
</tr>
</tbody>
</table>

Таблица 6. Коэффициенты корреляции Пирсона для дистрибутивных моделей разных конфигураций и типологических данных для поля глаголов качания. Полужирным выделен лучший результат.
‘гладкий’, см. строку 2 Таблицы 5). Возможно, это различие могло бы позволить сделать какие-то содержательные выводы об особенностях структуры признаковых и глагольных семантических полей, но для этого необходимы дополнительные эксперименты с данными других семантических зон, выходящие за рамки этого исследования.

4. Эксперимент 3: англоязычный обучающий корпус

Итак, эксперименты с двумя признаковыми и одним глагольным полем демонстрируют высокие коэффициенты корреляции между типологическим и дистрибутивным пространствами (при использовании одних и тех параметров дистрибутивных моделей). Однако нельзя не принять во внимание тот факт, что, согласно методологии Московской лексико-типологической группы, типологические анкеты составляются изначально на русском материале. Впоследствии они уточняются и дополняются в ходе анализа других языков, но при этом опасность влияния стартового языка на итоговый результат типологического исследования теоретически все равно не устраняется полностью (или, по крайней мере, его отсутствие пока никак не доказано).

Если стартовый язык действительно существенно влияет на представления исследователей об организации той или иной семантической зоны, то результаты, полученные нами в предыдущих экспериментах, неудивительны и не очень показательны. Мы продемонстрировали, что данные русского языка действительно соответствуют результатам типологического анализа рассматриваемых семантических зон, что, возможно, как раз говорит в пользу слишком сильного влияния первого языка на последующий типологический анализ, а не в пользу принципиальной сопоставимости типологического и дистрибутивного пространств.

Исходя из этих соображений, мы провели дополнительную серию экспериментов на материале поля ‘острый’, но в этот раз сопоставили типологическое пространство с дистрибутивным, построенным по данным англоязычного обучающего корпуса. Для этого каждой строке типологической анкеты были поставлены в соответствие английские словосочетания вида «прилагательное + существительное» (sharp knife ‘острый нож’, sharp arrow ‘острая стрела’, sharp nose ‘острый нос’, см. полный список в Приложении 1), а векторные представления для них были построены с помощью корпуса текстов на английском языке ukWaC. Вектора словосочетаний строились с помощью нескольких моделей композиции: аддитивной, мультипликативной, расширительной, аддитивной
взвешенной, лексической функции и практической лексической функции. Исходные вектора существительных взвешивались по схеме PPMI и подвергались операции сокращения размерности семантического пространства до 300 измерений (SVD).

Таблица 7. Коэффициенты корреляции Пирсона для дистрибутивных моделей разных конфигураций, обученных на англоязычном корпусе, и типологических данных для поля 'острый'. Полужирным выделены лучшие результаты.

<table>
<thead>
<tr>
<th>Корпус</th>
<th>Операции над векторами</th>
<th>Модель композиции</th>
<th>Коэффициент корреляции Пирсона</th>
</tr>
</thead>
<tbody>
<tr>
<td>ukWaC</td>
<td>ppmi, SVD</td>
<td>additive</td>
<td>острый 0.311  острый: прямые значения 0.728</td>
</tr>
<tr>
<td>ukWaC</td>
<td>ppmi, SVD</td>
<td>multiplicative</td>
<td>0.139  0.521</td>
</tr>
<tr>
<td>ukWaC</td>
<td>ppmi, SVD</td>
<td>dilation</td>
<td>0.165  0.326</td>
</tr>
<tr>
<td>ukWaC</td>
<td>ppmi, SVD</td>
<td>weighted additive</td>
<td>0.296  0.624</td>
</tr>
<tr>
<td>ukWaC</td>
<td>ppmi, SVD</td>
<td>LF</td>
<td>0.239  0.516</td>
</tr>
<tr>
<td>ukWaC</td>
<td>ppmi, SVD</td>
<td>PLF</td>
<td>0.281  0.668</td>
</tr>
</tbody>
</table>

Таблица 7 показывает, что максимальный коэффициент корреляции по-прежнему высок: 0.728. При этом наиболее удачные параметры дистрибутивной модели снова те же, что и в предыдущих экспериментах: самые высокие результаты показывают аддитивная и аддитивная взвешенная модели и модель практической лексической функции. При этом стабильнее всего работает самая простая аддитивная модель, которая, к тому же, и оптимальна с точки зрения необходимых временных и текстовых ресурсов.

Это результат позволяет сделать три теоретических вывода:

1) по-видимому, действительно можно считать сопоставимыми моноязычные дистрибутивные и типологические пространства;

2) понятие фрейма можно считать достаточно обоснованным;

3) методику типологического анализа лексики, предложенную Москвой лексико-типологической группой, можно считать достаточно корректной: результаты нашей последней серии экспериментов говорят в пользу того, что итог такого исследования не зависит напрямую от первого языка выборки.

Оговоримся, что все эти выводы носят очень общий характер и требуют дополнительной тщательной проверки. Однако наши эксперименты позволяют их по крайней мере сформулировать в качестве рабочих гипотез.
5. Эксперимент 4: визуализация векторных пространств

Наконец, соответствие между типологическим и дистрибутивным пространствами хорошо иллюстрируется их визуализациями. Оба пространства для всех наших семантических зон исходно многомерны (напомним, что в типологическом пространстве в качестве измерений выступают слова всех языков выборки, относящихся к рассматриваемому полю, а в дистрибутивном – 10 000 частотных знаменательных слов русского языка, которые потенциально могут встречаться в текстах рядом со словосочетаниями, для которых мы строим дистрибутивное представление). Для каждого поля мы отобразили оба пространства на плоскость с помощью техники многомерного шкалирования (MDS, см. Cox & Cox 2001), т.е. сократили количество измерений до двух и визуализировали полученную картину, обозначив одним цветом точки, относящиеся к одному и тому же фрейму. Такие отображения иллюстрируют относительные расстояния между объектами в многомерном пространстве (они строятся так, чтобы попарные расстояния между всеми точками на двумерной карте максимально точно отражали расстояния между соответствующими объектами в исходном пространстве). Значения измерений итогового двумерного пространства очевидной смысловой нагрузки не несут, хотя иногда и поддаются интерпретации.


Рисунки 1-3 демонстрируют интересный эффект. Визуализация типологического пространства (Рис. 1) наглядно отображает фреймовую структуру поля: объекты, относящиеся к одному фрейму, как и ожидалось, оказываются близко друг к другу на карте. Визуализация дистрибутивного пространства, напротив, отражает только те противопоставления, которые лексикализованы в данном языке. Так, например, карта на

---

16 Заметим, что метод многомерного шкалирования успешно применяется в типологии как раз для автоматического построения семантических карт, см. Croft & Poole 2008; Georgakopoulos & Polis 2018; Wälchli & Cysouw 2012 и др. Мы будем говорить об этом подробнее в Главе 5.
Рис. 2 построена на основе данных русского языка, и по ней четко выделяются контексты для прилагательного колючий, а контексты, обслуживаемые прилагательным острый, представляют собой неделимый континуум. На Рис. 3 представлена визуализация дистрибутивного пространства поля ‘острый’, построенного на основе франкоязычного корпуса (для наглядности – только те фреймы, которые не разделяются на материале русского языка). Французские данные позволяют противопоставить фрейм ‘острый инструмент с режущим краем’ фреймам ‘острый инструмент с колющим концом’ и ‘объект вытянутой формы’, поскольку первый описывается прилагательным tranchant, а два других – pointu, т.е. именно это противопоставление лексикализовано во французском.

Рисунок 1. Визуализация типологического пространства поля ‘острый’.
Рисунок 2. Визуализация дистрибутивного пространства поля ‘острый’, построенного на основе русскоязычного корпуса.

Рисунок 3. Визуализация дистрибутивного пространства поля ‘острый’, построенного на материале франкоязычного корпуса (без учета фрейма ‘объект с колючей поверхностью’).
Важно, однако, что, если не отображать на плоскость все точки дистрибутивного пространства, а выделить ядро каждого фрейма и только эти ядерные элементы помещать на карту, то картина меняется. Мы вычислили средние арифметические значения по каждому измерению для каждого «кластера», определив тем самым центр каждого фрейма, и эти новые разреженные пространства снова отобразили на плоскость. По Рисунку 5 видно, что такая методика позволяет получить прямой аналог традиционной дискретной семантической карты (Рис. 4) на материале одного-единственного языка.

Рисунок 4. Семантическая карта поля ‘острый’, составленная вручную на основе типологических данных.

Рисунок 5. Семантическая карта поля ‘острый’, составленная автоматически путем отображения на плоскость векторного пространства, состоящего из центральных представителей каждого фреймового кластера.

Наш эксперименты по автоматическому составлению лексико-типологических анкет, также основанные на построении пространств векторов сочетаемости и их последующей кластеризации (см. Главу 3), показывают, что словосочетания, в которых прилагательные разные, а существительные одинаковые (ср. гладкий пол, ровный пол и скользкий пол), часто оказываются в одном кластере, несмотря на то, что иллюстрируют разные фреймы. Однако проекция векторного пространства для поля ‘мокрый’, построенного на материале корпуса текстов на русском языке, вопреки ожиданиям, демонстрирует, что словосочетания, представляющие один фрейм, оказываются ближе друг к другу, чем словосочетания из разных фреймов, включающие одинаковые существительные (см. Рис. 6). Примечательно, что единственная область, где два фрейма не удалось различить даже с помощью анализа широкой сочетаемости, - это зона ‘мокрый после контакта с водой’, внутри которой различаются разные степени проявления признака: ‘мокрый’ и ‘влажный’ (= уже не такой мокрый, но ещё не совсем сухой, или, наоборот, уже не сухой, но ещё не мокрый).

6. Выводы

Полученные результаты позволяют сделать три основных вывода:

(1) Поскольку во всех четырех экспериментах между типологическим и дистрибутивным пространствами есть заметная корреляция, тщательно собранные вручную типологические данные могут использоваться для оценки качества дистрибутивных моделей. Такая метрика имеет ряд преимуществ по сравнению с уже существующими (такими, как сопоставление дистрибутивных расстояний со спонтанными суждениями носителей языка или с длиной пути от одного слова к другому по дереву того или иного тезауруса): в частности, она значительно более объективная и универсальная, т.е. не зависит от языка обучающей выборки для дистрибутивной модели. Основной ее недостаток связан, в первую очередь, с тем, что надежных типологических данных пока собрано очень мало, однако мы предполагаем, что разработка алгоритмов автоматического сбора материала позволит в ближайшем будущем разрешить эту проблему.
(2) Полученные результаты служат дополнительным подтверждением тому, что за понятием фрейма стоит некоторая лингвистически значимая семантическая реальность. В пользу этого свидетельствует тот факт, что применение методологии внешней по отношению к фреймовому подходу семантической теории позволяет получить сопоставимые результаты: русские словосочетания, иллюстрирующие один фрейм, употребляются в более близких контекстах, чем словосочетания из разных фреймов. Однако все же считать фрейм точкой в семантическом пространстве — это некоторое упрощение, проистекающее из необходимости ручной обработки данных. По-видимому, фреймовая структура семантического поля ближе к континуальной, хотя в этом континууме значений отчетливо выделяются фокусные центры (ср. Кибрик 2013) — фреймы, — которые в большинстве случаев и определяют принципы лексикализации данного поля.

(3) Методология дистрибутивной семантики позволяет определять основные контуры фреймовой структуры поля на материале одного языка, а эксперимент на материале англоязычного корпуса позволяет предположить, что выбор того или иного языка в качестве отправной точки лексико-типологического исследования не влияет на итоговый результат.

Третий вывод особенно важен, поскольку он согласуется со многими интуитивными предположениями разработчиков фреймового подхода к лексической типологии. Очевидно, что уже один язык дает очень важную с типологической точки зрения информацию, если некоторое противопоставление в нем лексикализовано (ср. поле ‘гладкий’ выше, где для всех основных фреймов в русском языке есть отдельные лексемы гладкий, ровный и скользкий). Но и в тех случаях, когда в стартовом языке все поле покрывается одним доминантом словом, разные косвенные свидетельства нередко позволяют предугадать, на какие классы (т.е. фреймы) разбиваются его употребления. Во-первых, таким свидетельством могут служить антонимы: русское слово старый объединяет фреймы ‘старый (об одушевленных существах) – пожилой’ (старый человек) и ‘старый (о неодушевленных объектах) – износиовшийся, уже негодный к употреблению’ (старая тряпка, старый башмак), однако этим фреймам соответствуют разные антонимы (молодой vs. новый), см. подробнее Rakhilina, Vyrenkova & Orekhov to appear. Другим таким свидетельством могут быть периферийные лексемы, которые дублируют основную, но только в некоторых типах контекстов. Так, например, в русском языке поле ‘острый’ покрывается одним словом острый, но есть также и периферийное низкочастотное слово
остроконечный, выступающее синонимом к слову острый только в контекстах, соответствующих фрейму вытянутой формы.

Наши же эксперименты показывают, что есть ещё один способ выявлять (с некоторой долей условности, но зато систематически, и не только в тех случаях, когда в языке есть «удачные» антонимы или синонимы) типологически релевантные противопоставления там, где они не видны невооруженным глазом. Она заключается в анализе широкого контекста употребления рассматриваемых лексических единиц. Интересно, что сама по себе эта методология отнюдь не нова для фреймового подхода: напомним, что в рамках этой парадигмы именно сочетаемость слов является основным ключом к выделению значений. Новшество, однако, в том, что обычно дистрибутивные свойства слов исследуются вручную и, как следствие, во внимание принимаются только минимальные «диагностические» контексты. Для прилагательных это определяемые существительные, для одноместных глаголов – существительные, заполняющие их единственную валентность, и т.д. Вычислительные методы дистрибутивной семантики позволяют принять во внимание более широкий контекст. И оказывается, что информация об этом широком контексте оказывается полезной там, где узкий контекст уже не дает достаточно надежных сведений. Наша последняя серия экспериментов с визуализациями хорошо иллюстрирует это явление: лексикализованные в языке противопоставления заметны сразу, а нелексикализованные противопоставления становятся видны, если перейти на следующий уровень обобщения.

значит, нуждаются в дополнительной проверке, но и то, что во всех экспериментах мы получили сопоставимые результаты, вряд ли можно назвать случайностью.

Таким образом, полученные нами результаты дополнительно подтверждают гипотезу о том, что фрейм – это не плод фантазии исследователя, а естественное объединение похожих ситуаций, воспроизводимое от языка к языку и проявляющееся в том, что в ареально и генетически разных языках ситуация, входящие в один фрейм, описываются единообразно. 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: 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. Таким экспериментам и посвящена следующая глава.
Глава 3. Автоматическая разработка анкеты с помощью моделей дистрибутивной семантики

§1. Краткий обзор существующих методов составления типологических анкет

В последние годы в лингвистике, как и во многих других науках, происходит бурное развитие компьютерных методов сбора, хранения и анализа данных. В частности, активно развивается корпусная лингвистика, и для многих языков (не только крупных европейских) уже доступны объемные корпуса текстов. Однако по-прежнему далеко не каждый язык, включаемый в типологическую выборку, снабжен достаточным количеством готовых электронных ресурсов, к тому же сопоставимых по качеству и объему с имеющимися ресурсами для других языков выборки. В таких условиях основным инструментом сбора и анализа материала оказывается типологическая анкета.

Несмотря на то, что от качества анкеты, как правило, зависит результат всего типологического исследования, четкой методологии составления лингвистических опросников, насколько нам известно, по-прежнему не существует ни в грамматической, ни в лексической типологии. Исключение составляют психолингвистические исследования лаборатории Института имени Макса Планка в Неймегене, где в качестве анкеты используется набор экстралингвистических стимулов, подобранных по определенным параметрам. В этом случае анкета представляет собой перебор всех возможных комбинаций заданных параметров, т.е. строится по четким и понятным принципам. Самый известный пример исследования по такого рода опросникам — типология цветообозначений (см. Berlin & Kay 1969; Kay et al. 2007), где в качестве основной анкеты используется цветовая система Манселла, в рамках которой каждый цвет определяется тремя числами: значениями тона, яркости и насыщенности.

В подходах, опирающихся на лингвистическое поведение языковых единиц, а не на соотношение языкового знака с его денотатом (к их числу относится и фреймовый подход к лексической типологии, на который мы опираемся в нашем исследовании), наиболее распространенная методология составления анкет заключается в подробном анализе корпусных данных для нескольких языков с богатой письменной традицией и большим количеством доступных ресурсов. На основе этого материала выявляются параметры

---

варьирования изучаемых языковых единиц и строится анкета, учитывающая все эти параметры (см., например, Рахилина & Резникова 2013).

Среди этой группы методик наиболее строгой оказывается недавно предложенная методология автоматического составления анкет на основе параллельных корпусов, см. Dahl 2007; Wälchli & Cysouw 2012. В рамках этого подхода пунктом анкеты считается каждое вхождение в корпус анализируемого слова или грамматического показателя “стартового” языка. Такая анкета сразу заполняется сведениями о том, что соответствует этим единицам в других языках в тех же самых контекстах.

Ещё один метод автоматической подготовки типологических анкет разрабатывается в рамках принятого нами фреймового подхода к лексической типологии, см. Кюсева, Резникова, & Рыжова 2013а; Орехов & Резникова 2015, а также дипломную работу Абдурашитова 2017. Эта методика опирается на данные о сочетаемости рассматриваемых лексем (пока эксперименты проводились только на материале признаковой лексики), которые можно почерпнуть из коллекций биграмм корпорации Google18. Выбирается несколько прилагательных, относящихся к изучаемому полю в русском, немецком и английском языках. Для каждого прилагательного составляется список существительных, в сочетании с которыми оно встречается достаточно часто (т.е. которые формируют вместе с этим признаковым словом достаточно частотную биграмму вида «прилагательное + существительное»). На следующем этапе все существительные автоматически переводятся на английский язык и объединяются в общий список, который затем кластеризуется на основе данных о сочетаемости соответствующих существительных с рассматриваемыми признаковыми словами.

При всех явных достоинствах каждого из подходов, у них есть очевидные недостатки: психолингвистическая методика не позволяет изучать языковое поведение слов с достаточной степенью подробности; анкеты, составленные вручную на основе корпусов текстов, всегда могут оказаться неполными и во многом зависят от точки зрения исследователя; материала параллельных корпусов пока недостаточно для сколько-нибудь подробного анализа большинства семантических полей; анкеты, составленные автоматически на основе коллекции биграмм, пока получаются слишком громоздкими (один из вариантов анкеты для поля ‘твердый’ содержит 448 кластеров) и в таком виде не могут использоваться в работе с информантами; кроме того, чистота таких кластеризаций пока не очень удовлетворительна.

18 URL: [http://storage.googleapis.com/books/ngrams/books/datasetv2.html]
Метод составления анкеты, который предлагаем мы, хоть и обладает некоторыми своими недостатками, преодолевает ряд перечисленных выше, поэтому может служить по крайней мере дополнением к упомянутым методикам. Во-первых, он основан на анализе лингвистического поведения языковых единиц (корпусных данных); во-вторых, он полностью автоматический; в-третьих, он основывается на данных не параллельных, а одноязычных корпусов, что позволяет использовать более сбалансированные и более объемные коллекции текстов, а следовательно, набирать достаточно материала для анализа более широкого круга семантических полей; в-четвертых, он опирается на данные только одного языка и позволяет получить в результате анкету приемлемого размера.

В рамках данного исследования мы разработали пилотную версию такого алгоритма на материале признаковых слов и близких к ним со структурной точки зрения одноактантных глаголов. Методика отлаживалась преимущественно на материале поля ‘острый’ (см. также наше пилотное исследование Рыжова 2014), после чего оптимальная версия алгоритма была опробована на дополнительном материале нескольких признаковых и глагольных полей. В следующем параграфе (§2) мы подробно опишем традиционную методику подготовки типологической анкеты в рамках фреймового подхода к лексической типологии, сформулируем требования к анкете, которые необходимо соблюсти при попытке автоматизации этого процесса, и поставим формальные задачи, которые должен решать разрабатываемый алгоритм. Затем, в параграфах 3-8, мы представим все вариации алгоритма, которые мы тестировали на материале поля ‘острый’. В параграфе 9 мы демонстрируем возможности применения оптимальной конфигурации параметров алгоритма к анализу нескольких признаковых и глагольных полей. В параграфе 10 мы подводим итоги этого этапа работы.

§2. Постановка задачи

Фреймовый подход к лексической типологии основан на анализе правил сочетаемости исследуемых лексем. Этот общий принцип определяет и формат анкет, которые представляют собой наборы контекстов, а не, например, экстралингвистических стимулов.

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 (предложений), иллюстрирующих круг возможных употреблений 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 многих других языках.

С другой стороны, анкета должна быть не очень большой: в противном случае экспертам по языкам будет очень сложно с ней работать. Поэтому важно не просто собрать список всех возможных словосочетаний, в которых могут фигурировать лексемы изучаемого семантического поля, но и расклассифицировать их, выделив набор различных типов ситуаций, в которых могут употребляться рассматриваемые слова. При этом важно учесть все потенциально возможные лексические противопоставления, но включить в анкету всего по несколько (3-5) примеров реализации каждого из предполагаемых типов ситуаций.

Далее, примеры для анкеты должны быть достаточно показательными. 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 ситуации описания размера поперечного среза длинного вытянутого объекта правильнее использовать формулировку ‘толстая палка’, а не ‘толстый лом’.

КОММЕНТАРИЙ

В рамках фреймового подхода к лексической типологии термин анкета может использоваться для обозначения двух разных типов инструментов. С одной стороны, анкетой может называться набор стимулов, предъявляемых носителям изучаемых языков. В этом случае стимул практически никогда не сводится к словосочетанию из двух элементов: контекст расширяется до предложения или короткого фрагмента текста, чтобы информант имел возможность максимально полно и точно представить себе ситуацию, которую имеет в виду исследователь. Например, ситуации ‘толстая палка’ мог бы соответствовать такой пункт анкеты: Я выбрал слишком толстую палку, ее неудобно держать в руке.

С другой стороны, анкетой называют набор коротких словосочетаний (минимальных диагностических контекстов), который выдается экспертам по языкам, включенным в выборку, и используется для сопоставления собранных данных. Такая анкета служит ориентиром для исследователя: частотность употребления именно
таких единиц оценивается по корпусам; на основе именно таких словосочетаний исследователь создает расширенную версию опросника (т.е. анкету первого типа), учитывающую специфику обследуемого языка.

В рамках настоящей работы мы будем говорить исключительно об анкетах второго типа.

На следующем этапе анкета «тестируется» на материале другого языка, причем желательно, чтобы этот язык был также обеспечен обширными и доступными лексикографическими ресурсами и чтобы у исследователя была возможность общения с его носителями. Всё это необходимо для того, чтобы новый язык можно было изучить столь же подробно, опираясь не только на анкету, но и на другие источники информации. Привлечение нового материала всегда влечет за собой уточнение и расширение анкеты: обнаружаются 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 языку после анализа каждого нового идиома и последующее уточнение анкеты называется челюновым 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 только в деталях, касающихся преимущественно метафорических употреблений. 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 1, а также её фрагмент в Таблице 8).
Задача данного этапа нашего исследования – разработать алгоритм автоматического (или полуавтоматического) составления лексико-типологической анкеты на основе данных одного языка – русского. Для того, чтобы составить полноценную анкету-опросник, удовлетворяющую всем перечисленным требованиям, необходимо решить несколько подзадач. А именно:

1) определить, какие слова в русском языке могут иметь отношение к изучаемому полю;
2) для каждого такого слова составить список коллокаций (например, в случае с прилагательными, – существительных, в сочетании с которыми они могут употребляться);
3) разделить список контекстов на минимальные семантически однородные группы;
4) в случае необходимости сократить анкету до приемлемого размера;
5) оценить результат.

В последующих разделах мы рассмотрим каждую из этих задач и подробно обсудим возможные методы их решения. Все базовые эксперименты по подбору методов проводились преимущественно на материале поля ‘острый’, а затем настройки, признанные оптимальными, дополнительно тестировались на базе других признаковых и некоторых глагольных полей. Именно так мы и построим наше дальнейшее изложение: сначала опишем все опробованные нами методики, затем сделаем промежуточный вывод об эффективности каждой из них, а после этого приведем результаты дополнительных экспериментов с другими семантическими полями.
§3. Определение круга лексем, относящихся к изучаемому полю

Задача составления списка прилагательных, относящихся к тому или иному семантическому полю, при своей кажущейся простоте связана с рядом нетривиальных моментов, которые так или иначе преодолеваются в процессе ручной работы, но могут серьезно осложнить задачу автоматической разработки анкеты-опросника. Прежде всего, задача составления списка слов (в нашем случае, прилагательных), на основе которых будет строиться анкета, естественным образом связана с задачей определения границ изучаемого поля. Однако, по-видимому, семантические поля (по крайней мере, качественных признаков) никогда не бывают полноценно замкнутыми, поэтому задача определения их границ становится очень нетривиальной. Мы обсудим ее в разделе 1. Дополнительную сложность создают метафорические расширения изучаемых полей, о которых мы поговорим в разделе 2. В разделе 3 мы сформулируем требования, которые будем предъявлять к составляемым автоматически спискам слов, относящихся к полю ‘острый’, с учетом этих двух проблем. Наконец, в разделе 4 мы рассмотрим сами алгоритмы (полу)автоматического составления таких списков и оценим, насколько результаты их работы соответствуют нашим требованиям.

1. Проблема границ поля

Традиционно семантическим полем считается «…множество слов, объединенных общностью содержания, или, говоря более конкретно, имеющих общую нетривиальную часть в толковании» (Кронгауз 2005: 130). В рамках нашего подхода под семантическим полем подразумевается набор связанных между собой фреймов. Алгоритм составления списка фреймов мы уже обсуждали 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.

Схематически эта процедура уточнения первоначальных данных с помощью материала других языков представлена в Таблице 9. 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 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.

<table>
<thead>
<tr>
<th>Step 1</th>
<th>Step 2</th>
<th>Step 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1, lex 1</td>
<td>L2, lex 2</td>
<td>L2, lex 3</td>
</tr>
<tr>
<td>Sem 1</td>
<td>Sem 1</td>
<td>Sem 1</td>
</tr>
<tr>
<td>Sem 2</td>
<td>Sem 2.2</td>
<td>Sem 2.2</td>
</tr>
<tr>
<td>Sem 3</td>
<td>Sem 3</td>
<td>Sem 6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L1, lex 4</td>
<td>L1, lex 5</td>
<td>L1, lex 6</td>
</tr>
<tr>
<td>Sem 4</td>
<td>Sem 5</td>
<td>Sem 6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sem 7</td>
<td>Sem 8</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sem 9</td>
<td></td>
</tr>
</tbody>
</table>

Таблица 9. Процедура уточнения 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 §2 мы отметили, что процесс составления списка фреймов конечен. Однако на самом деле это верно лишь отчасти. Обычно 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 pointu (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, что, конечно, не позволяет
утверждать, что их нет ни в одном языке мира, но позволяет предположить, что основные, наиболее когнитивно значимые закономерности уже обнаружены.

Иначе обстоит дело с появлением совсем новых значений, ранее в анкету не входивших. Процесс добавления новых контекстов употребления слов, видимо, потенциально бесконечен, поскольку семантические поля (по крайней мере, признаковые) 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 (см. Плунгян & Рахилина 2007, Кюсеева и др. 2013, Козлов и др. 2016).


Таким образом, представленная нами процедура определения набора фреймов может продолжаться бесконечно, поскольку между разными семантическими полями есть переходные фрагменты, позволяющие переходить из одной зоны в другую, захватывая всё новые и новые контексты. Поэтому для типологического исследования границы поля обычно определяются не только собственно устройством языка, но и решением исследователя, который сам выбирает набор фреймов, которым он хочет ограничиться. Это означает, что у задачи составления списка прилагательных, относящихся к тому или
иному семантическому полю, заведомо нет единственно правильного ответа: результат зависит отчасти от субъективного решения, принимаемого исследователем.

2. Проблема метафорических значений

Важная черта нашего алгоритма составления списка фреймов, которую мы до сих пор не упоминали, заключается в следующем: 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 острий прилагательное oштар 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. Подобная методика отбора слов применяется также в работе François 2008.

КОММЕНТАРИЙ

Такое ограничение заметно сокращает область поиска новых фреймов. Однако введение его не означает, что метафоры вообще исключаются из анкеты: если у слова, описывающего хотя бы один из фреймов прямых значений, есть и метафорические употребления, то все они учитываться в опроснике. Для нас эта информация очень важна: все наши исследования конкретных полей (ср. Кашкин 2013; Кюсева 2012; Майсак & Рахилина 2007; Павлова 2014 и др.) показывают, что метафоры мотивированы исходными значениями слов, и поэтому модели таких сдвигов воспроизводятся от языка к языку столь же регулярно, что и принципы лексикализации прямых значений. С другой стороны, между прямыми и метафорическими 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. Мы предполагаем, что число таких фреймов может оказаться конечным, однако охватить их все в рамках одного исследования крайне затруднительно. Ситуация с метафорами принципиально иная: если анализ новых языков при ограничении области исследования очень быстро перестаёт приносить новые фреймы прямых значений (5 языков обычно уже достаточно), то новые метафоры, пусть единичные при наличии большого пласта повторяющихся моделей, 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 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

19 См. аналогичное рассуждение в статье François 2008.
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 экспериментов, описанных в Главе 2, мы преодолевали это затруднение 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 бесленевском 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. Всё это в очередной раз свидетельствует, что единственного правильного варианта анкеты, как и единственно верного списка прилагательных, которые должны быть включены в исследование, не существует.

20 В большинстве случаев такой широкий набор возможностей возникает в зоне качеств человека, а также в области интенсификаторов и оценочных прилагательных. Вследствие антропоцентричности языка 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, а также нашу работу об оценочной лексике Рыжова 2016.
3. Формализация задачи

Анкета, созданная вручную и использовавшаяся при проведении экспериментов из Главы 2 (она же представлена в Приложении 1), включает 15 прилагательных: острый, резкий, крутой, колючий, четкий, быстрый, яркий, умный, проницательный, сильный, непоседливый, хороший, высокий, грубый, газированный. Очевидно, что многие слова попали сюда из-за «метафорических лакун» и непосредственно с семантическим полем ‘острый’ не связаны (ср. непоседливый, газированный, сильный, высокий, яркий...). Поэтому в экспериментах по определению круга слов, относящихся к данному полю, мы будем использовать в качестве золотого стандарта (набора прилагательных, который мы условно считаем оптимальным) модифицированный вариант этого списка.

Если считать, что лексема острый во всех методах используется в качестве стартовой («сида»), то обязательными, помимо нее, мы будем считать слова резкий, крутой, колючий и четкий / отчетливый, поскольку они покрывают фреймы физических значений, которые могут относиться к полю ‘острый’ (и не обслуживаются русским прилагательным острый), и, по-видимому, не имеют ядра, выходящего за пределы этой зоны, т.е. все основные контексты их употребления могут в том или ином языке покрываться словом ‘острый’. Далее, мы будем считать допустимыми прилагательные, деривационно связанные со словом острый и близкими к нему семантически глагольными лексемами точить, проникать, пронзать / пронизывать, т.е. наличие в итоговом списке лексем заточенный, остроконечный, проницательный, остроумный и проч. мы не будем рассматривать как ошибку системы. Но и необходимыми эти слова мы считать не будем, поскольку, согласно методологии MLexT, основной интерес для исследователя представляют прежде всего непроизводные лексемы или лексемы, производность которых уже совсем не ощущается носителями. Тот же статус, и что словообразовательным дериватам, мы приписываем слову прыгун: оно обслуживает фрейм, который может относиться к полю ‘острый’ и не имеет при этом внешнего ядра (т.е. других значений, которые не могли бы описываться словами из поля ‘острый’), но и обязательным мы его не считаем, потому что этот же фрейм в русском языке свободно покрывается доминантным словом острый. Прилагательные, которые относятся только к фреймам метафорических значений и имеют семантическое ядро за пределами этой зоны (такие как высокий, тонкий, яркий и др.), будут считаться ошибками в работе алгоритма.
4. Методы (полу)автоматического составления списка прилагательных

В этом разделе мы рассмотрим последовательно четыре метода составления списка релевантных прилагательных и выделим их достоинства и недостатки:

1) метод анализа синонимов (на материале одного языка);
2) метод ближайших соседей в пространстве векторов сочетаемости (на материале одного языка);
3) метод, опирающийся на онтологию RuWordNet (полуавтоматический);
4) метод обратных переводов (полуавтоматический, с привлечением словарей трех языков).

4.1. Метод анализа синонимов

Наиболее очевидный метод решения задачи отбора прилагательных, относящихся к одному и тому же полю, – обращение к словарям синонимов.

В Новом объяснительном словаре синонимов под редакцией 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. В словаре под редакцией З. Е. Александровой (Александрова 2001) есть нужный нам раздел, и в нём приведены следующие синонимы интересующей нас лексемы:

<table>
<thead>
<tr>
<th>Заостренный</th>
<th>Острорезкий</th>
<th>Остроумный</th>
<th>Напряженный</th>
</tr>
</thead>
<tbody>
<tr>
<td>Наточенный</td>
<td>Прямой</td>
<td>Мучительный</td>
<td>РЕЗКИЙ</td>
</tr>
<tr>
<td>Вострый (прост.)</td>
<td>Проницательный</td>
<td>Актуальный</td>
<td></td>
</tr>
</tbody>
</table>

Таблица 10. Синонимы прилагательного острый по словарю З.Е. Александровой (2001)

В Таблице 10 малыми прописными буквами написано слово, которое входит в установленный нами золотой стандарт, курсивом выделены необязательные, но допустимые лексемы, а полужиром набраны прилагательные, которые мы бы хотели исключить из результирующего списка. Таким образом, словарь синонимов З. Е. Александровой (2001) демонстрирует достаточно высокую точность (0.73 – из 11 слов 8 считаются допустимыми), но низкую полноту (0.25 – из четырех обязательных слов присутствует только одно).

Помимо традиционных бумажных словарей синонимов, мы использовали в качестве источника лексикографических данных электронные базы 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 для нашей задачи неприемлемо. Мы провели несколько этапов фильтрации: сначала в каждом из списков мы оставили только те слова, которые входят в словарь из 10 000 наиболее частотных знаменательных лексем по основному подкорпусу НКРЯ\(^{21}\), а затем объединили два списка, оставив только те прилагательные, которые встречаются в обоих. Полученный в результате набор представлен в Таблице 11. В этой таблице используется та же аннотация, что и в Табл. 10: малыми прописными буквами набраны необходимые слова, курсивом – допустимые, полужирным шрифтом выделены лишние прилагательные.

Количество необходимых слов, относящихся к фреймам физических значений, по-прежнему мало (всего одно из четырех, т.е. полнота по-прежнему равна 0.25). Синонимов метафорических значений, напротив, остаётся очень много даже после всех уровней фильтрации, что приводит к очень низкому уровню точности: 0.16 (из 19 слов всего три относятся к нужным или допустимым).

<table>
<thead>
<tr>
<th>Актуальный</th>
<th>Критический</th>
<th>Нежный</th>
<th>Сильный</th>
</tr>
</thead>
<tbody>
<tr>
<td>Драматический</td>
<td>Лютый</td>
<td>Остроумный</td>
<td>Современный</td>
</tr>
<tr>
<td>Жгучий</td>
<td>Мучительный</td>
<td>Пронзительный</td>
<td>Тонкий</td>
</tr>
<tr>
<td>Жестокий</td>
<td>Напряженный</td>
<td>РЕЗКИЙ</td>
<td>Чуткий</td>
</tr>
<tr>
<td>Крайний</td>
<td>Насущный</td>
<td>Своевременный</td>
<td></td>
</tr>
</tbody>
</table>

Таблица 11. Список синонимов прилагательного острый, полученный в результате обработки данных, предоставленных ресурсами synonymizer.ru и synonymonline.ru.

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, слишком много. Иными словами, этот метод неудовлетворителен ни по одному из показателей: ни по полноте, ни по точности.

4.2. Метод ближайших соседей

Другой метод определения круга семантически близких слов, в настоящее время особенно популярный, – это выделение лексем, находящихся на минимальном расстоянии

\(^{21}\) Словарь составлен нами совместно с М.В. Кюсевой для предыдущей серии экспериментов (см. Главу 2).
друг от друга в некотором векторном пространстве, см. (Baroni et al. 2014; Dubossarsky et al. 2016; Шеянова 2016). Этот подход опирается на идею о том, что близкие по смыслу слова употребляются в похожих контекстах (см. Sahlgren 2008 о так называемой дистрибутивной гипотезе, а также Главу 2 данной диссертации), а значит, предположительно, слова из одного семантического поля могут оказаться рядом друг с другом (т.е. ближайшими соседями) в пространстве векторов сочетаемости.

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 имеющемуся в нашем распоряжении основному подкорпусу НКРЯ; этот же текстовый материал использовался в качестве обучающего корпуса. Вектор каждого слова состоял из значений 10 000 измерений, причём в качестве набора измерений также использовались самые частотные знаменательные слова по основному подкорпусу 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: 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 неприлагательные.

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 появились в первом варианте списка, но зато добавляет множество лексем, не имеющих непосредственного отношения к полю ‘острый’ (см. Таблицу 12, аннотация та же, что и в предыдущем разделе).

<table>
<thead>
<tr>
<th>Широкое окно</th>
<th>Узкое окно</th>
</tr>
</thead>
<tbody>
<tr>
<td>Тупой</td>
<td>Тупой</td>
</tr>
<tr>
<td>Тяжелый</td>
<td>Жгучий</td>
</tr>
<tr>
<td>Тонкий</td>
<td>Неприятный</td>
</tr>
<tr>
<td>------------</td>
<td>------------</td>
</tr>
<tr>
<td>Легкий</td>
<td>Тяжелый</td>
</tr>
<tr>
<td>Жуткий</td>
<td>Мучительный</td>
</tr>
<tr>
<td>Блестящий</td>
<td>Господний</td>
</tr>
<tr>
<td>Глубокий</td>
<td>Любопытный</td>
</tr>
<tr>
<td>Жгучий</td>
<td>Выразительный</td>
</tr>
<tr>
<td>Печальный</td>
<td>Сильный</td>
</tr>
<tr>
<td>Похожий</td>
<td>Жуткий</td>
</tr>
<tr>
<td>Злобный</td>
<td>Темный</td>
</tr>
<tr>
<td>Жесткий</td>
<td>Жесткий</td>
</tr>
<tr>
<td>Непривычный</td>
<td>Серьезный</td>
</tr>
<tr>
<td>Видимый</td>
<td>Приятный</td>
</tr>
<tr>
<td>Страшный</td>
<td>Тонкий</td>
</tr>
<tr>
<td>Маленький</td>
<td>Грустный</td>
</tr>
<tr>
<td>Сильный</td>
<td>Внимательный</td>
</tr>
<tr>
<td>Странный</td>
<td>Сложный</td>
</tr>
<tr>
<td>Беспокойный</td>
<td>Беспокойный</td>
</tr>
<tr>
<td>Уродливый</td>
<td>Злобный</td>
</tr>
<tr>
<td>Мрачный</td>
<td>Нежный</td>
</tr>
<tr>
<td>Сухой</td>
<td>Нестерпимый</td>
</tr>
<tr>
<td>Отвратительный</td>
<td>Разнообразный</td>
</tr>
<tr>
<td>Наглый</td>
<td>Печальный</td>
</tr>
<tr>
<td>Мягкий</td>
<td>Похожий</td>
</tr>
<tr>
<td></td>
<td>Странный</td>
</tr>
<tr>
<td></td>
<td>Жадный</td>
</tr>
<tr>
<td></td>
<td>Привлекательный</td>
</tr>
<tr>
<td></td>
<td>Схожий</td>
</tr>
<tr>
<td></td>
<td>Интересный</td>
</tr>
<tr>
<td></td>
<td>Осторожный</td>
</tr>
<tr>
<td></td>
<td>Непривычный</td>
</tr>
<tr>
<td></td>
<td>Стоящий</td>
</tr>
<tr>
<td></td>
<td>Уродливый</td>
</tr>
<tr>
<td></td>
<td>Длинный</td>
</tr>
<tr>
<td></td>
<td>Страшный</td>
</tr>
<tr>
<td></td>
<td>Тревожный</td>
</tr>
</tbody>
</table>
Таблица 12. Ближайшие соседи прилагательного *острый* в широком и узком окне (из 50 соседей отобраны имена прилагательные).

Таким образом, на материале зоны ‘острый’ этот метод демонстрирует очень низкие показатели полноты и точности: полнота равна 0 (и для узкого, и для широкого окна), точность – 0.04 при широком окне, 0.026 – при узком.

4.3. Определение границ поля по материалам онтологии RuWordNet


Словари такого типа имеют древесную структуру. Листьями такого дерева оказываются слова рассматриваемого языка, которые иерархически группируются в семантические поля. При этом, если у слова несколько значений, оно попадает сразу в несколько полей, т.е. оказывается сразу на нескольких ветвях семантического дерева.

Теоретически такого рода ресурсы идеально подходят для решения нашей задачи. Однако на практике и их использование сопряжено с рядом затруднений.

С одной стороны, выделение у слова нескольких значений (т.е. включение его в несколько разных синсетов – семантических полей) может позволить избежать включения в итоговый список прилагательных таких слов, которые синонимичны признаку *острый* только в его метафорических употреблениях. С другой стороны, неочевидно, как отбирать нужные синсеты автоматически. Теоретически эта проблема могла бы решаться путем обращения к гиперонимам, которые для каждого синсета свои, но, по крайней мере, в случае со словом *острый*, гиперонимы в онтологии RuWordNet оказываются не очень информативны. Прилагательное *острый* попадает в четыре синсета со следующими гиперонимами: качественный vs. насмешливый vs. сильный, вкусовой vs. хороший, положительный. Внимательное изучение всех четырех синсетов показывает, что в основном интересующие нас прилагательные относятся к синсету 1 с гиперонимом *качественный*, но неочевидно, с помощью какого алгоритма можно было бы выбрать этот синсет автоматически.
Если считать, что синсет(ы) можно выбирать вручную, то дальше можно автоматически извлекать из нужных синсетов синонимы и гипонимы (и антонимы, в случае необходимости). В таком случае для поля ‘острый’ мы получим набор прилагательных, представленный в Таблице 13 (аннотация та же, что и в предыдущих разделах).

<table>
<thead>
<tr>
<th>вострый</th>
<th>отточенный</th>
</tr>
</thead>
<tbody>
<tr>
<td>точенный</td>
<td>заостренный</td>
</tr>
<tr>
<td>острийный</td>
<td>остроконечный</td>
</tr>
</tbody>
</table>

Таблица 13. Набор прилагательных, относящихся к полю ‘острый’, полученный путем анализа онтологии RuWordNet.

Из Таблицы 13 видно, что этот метод позволяет добиться высокого уровня точности (1: из шести прилагательных ни одного недопустимого), но дает минимальное значение полноты (0: из обязательных лексем нет ни одной).

4.4. Метод обратных переводов

Наконец, последний метод, который мы условно назвали методом обратных переводов, подразумевает выход за рамки русского языка и имитирует, хотя и в сильно упрощенном виде, ручную работу лексического типолога. Этот механизм основан на словарном материале, но источником данных являются не тезаурусы и 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 физических значений на английский, немецкий и французский языки, которые затем переводились обратно на русский. В качестве источника для перевода использовалась коллекция электронных словарей Яндекс (https://slovari.yandex.ru)\(^22\). Этот ресурс группирует переводные эквиваленты по значениям лексемы, для которой ищется перевод (в отличие, например, от словаря «Мультитран» - www.multitran.ru), что позволяет определить, какие переводы следует учесть, а какие исключить. Однако в этом случае, как и в ситуации с использованием онтологии RuWordNet, в алгоритм приходится включать элемент ручной работы.

Такая процедура наиболее точно воспроизводит челночный метод сбора прилагательных: в других языках находятся слова, способные покрывать фреймы ядерных (прямых) 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 14 (аннотация та же).

Из таблицы видно, что метод обратных переводов позволяет сформировать широкий набор прилагательных, в который попадает больше обозначений физических свойств объектов, чем в списки предыдущих версий, однако лексем, описывающих абстрактные характеристики объектов, по-прежнему очень много (значение полноты – 1, точности – 0.19). Особенно это касается прилагательных, обозначающих качества человека (см. сноску 18), которые мы для наглядности выделили в отдельный столбик таблицы.

<table>
<thead>
<tr>
<th>Физические признаки</th>
<th>Абстрактные признаки (не качества человека)</th>
<th>Качества человека</th>
</tr>
</thead>
<tbody>
<tr>
<td>ОСТРЫЙ</td>
<td>Пронзительный</td>
<td>Остроумный</td>
</tr>
<tr>
<td>КОЛЮЧИЙ</td>
<td>Быстрый</td>
<td>Умный</td>
</tr>
<tr>
<td>КРУТОЙ</td>
<td>Точный</td>
<td>Наблюдательный</td>
</tr>
<tr>
<td>ЕДКИЙ</td>
<td>Яркий</td>
<td>Внимательный</td>
</tr>
<tr>
<td>ОТЧЕТЛИВЫЙ</td>
<td>Обидный</td>
<td>Решительный</td>
</tr>
<tr>
<td>ВЫСОКИЙ</td>
<td>Крепкий</td>
<td>Суровый</td>
</tr>
<tr>
<td>РЕЗКИЙ</td>
<td>Неприятный</td>
<td>Жестокий</td>
</tr>
<tr>
<td>СИЛЬНЫЙ</td>
<td>Хитрый</td>
<td>Коварный</td>
</tr>
<tr>
<td>ОПРЕДЕЛЕННЫЙ</td>
<td>Смелый</td>
<td>Энергичный</td>
</tr>
<tr>
<td>НАПРАВЛЕННЫЙ</td>
<td>Боеевой</td>
<td>Ловкий</td>
</tr>
<tr>
<td>ГЛУБОКИЙ</td>
<td>Тонкий</td>
<td>Искусный</td>
</tr>
<tr>
<td></td>
<td>Глубокий</td>
<td>Энергичный</td>
</tr>
<tr>
<td></td>
<td>Тонкий</td>
<td>Ловкий</td>
</tr>
<tr>
<td></td>
<td>Силкий</td>
<td>Энергичный</td>
</tr>
<tr>
<td></td>
<td>Острый</td>
<td>Также</td>
</tr>
<tr>
<td></td>
<td>Быстрый</td>
<td>Также</td>
</tr>
<tr>
<td></td>
<td>Пронзительный</td>
<td>Также</td>
</tr>
</tbody>
</table>

Таблица 14. 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 обязательных слов, мы предприняли попытку провести его дополнительную фильтрацию с целью сократить количество абстрактных признаков. Для этого мы модифицировали алгоритм сбора
релевантных лексем по методу обратных переводов. 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, в новом варианте процедуры абстрактные значения отсекаются на этапах как прямого, так и обратного перевода.

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 релевантность. Однако полученный результат существенно превосходит предыдущие с точки зрения отсутствия большого количества «лишних» слов (см. Таблицу 15, аннотация та же): значение полноты – 1, точности – 0.46. 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.

<table>
<thead>
<tr>
<th>Быстрый</th>
<th>Жаркий</th>
<th>Произвольный</th>
</tr>
</thead>
<tbody>
<tr>
<td>Высокий</td>
<td>КОЛЮЧИЙ</td>
<td>РЕЗКИЙ</td>
</tr>
<tr>
<td>Горячий</td>
<td>КРУТОЙ</td>
<td>ЯРКИЙ</td>
</tr>
<tr>
<td>Душистый</td>
<td>ОСТРЫЙ</td>
<td>ОТЧЕТЛИВЫЙ</td>
</tr>
</tbody>
</table>

Таблица 15. Список прилагательных, полученный методом обратных переводов (вариант 2: "острый" => переводы физических значений на английский, немецкий и французский => переводы физических значений на русский).

5. Анализ результатов

Итак, на материале одной семантической области мы опробовали несколько методов определения границ поля: выбор синонимов опорного прилагательного по словарям / электронным базам синонимов; поиск ближайших соседей в пространстве векторов сочетаемости; подбор синонимов, гипонимов и когипонимов по онтологии RuWordNet; выбор релевантных прилагательных методом обратных переводов. Качество работы каждого из методов мы оценили, подсчитав для каждого полученного списка прилагательных полноту и точность относительно оптимального списка, составленного
вручную по материалам уже проведенного типологического исследования. Результаты оценки качества работы алгоритмов обобщены в Таблице 16.

<table>
<thead>
<tr>
<th>Метод</th>
<th>Полнота</th>
<th>Точность</th>
</tr>
</thead>
<tbody>
<tr>
<td>словари / базы синонимов</td>
<td>0.25 / 0.25</td>
<td>0.73 / 0.16</td>
</tr>
<tr>
<td>ближайшие соседи в векторном пространстве (широкое / узкое окно)</td>
<td>0 / 0</td>
<td>0.04 / 0.026</td>
</tr>
<tr>
<td>RuWordNet</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>обратные переводы (вариант 1 / вариант 2)</td>
<td>1 / 1</td>
<td>0.19 / 0.46</td>
</tr>
</tbody>
</table>

Таблица 16. Оценка качества работы методов автоматического определения границ семантического поля.

Из Таблицы 16 видно, что задача составления списка прилагательных, относящихся к изучаемому полю (т.е. фактически задача определения границ поля), очень плохо поддается автоматизации, по крайней мере, нашими методами (ср., однако, аналогичный вывод в работе Шеянова 2016). Самый лучший результат дает метод обратных переводов, подразумевающий не только автоматизированный перевод нужных слов, но и ручной отсев ненужных переводов. Но и в этом случае, при хорошем значении полноты, уровень точности остается очень низким.

Такой результат позволяет сделать несколько теоретических выводов. По-видимому, семантические поля по природе своей могут иметь разное устройство. По замыслу структуралистов (см. Trier 1931), семантическое поле должно было бы представлять собой замкнутую семантическую систему, где все слова были бы объединены наличием некоторого нетривиального общего смысла, а значение каждого элемента поля определялось бы только относительно других участников этой же подсистемы. И, что особенно важно, между разными полями постулировались жесткие границы: каждая лексическая единица должна входить ровно в одно семантическое поле. Мы не будем подробно обсуждать недостатки и достоинства такого определения, но подчеркнем, что лучше всего оно подходит для представления вполне определенных слоев лексики: предметных имен и очень близких к ним относительных прилагательных (ср. прилагательные цвета или материала). Элементы таких полей, как правило, связаны четко выделяющимися парадигматическими отношениями гипо-гиперонимии и синонимии, в их значении довольно легко выделить нетривиальный общий семантический компонент.

Именно такие, «классические», семантические поля хорошо выделяются компьютерными методами. Как правило, эти зоны хорошо проработаны в тезаурусах. А
поскольку слова, относящиеся к таким полям, как правило, употребляются в контекстах одного типа (см., например, Рахлина 2010), они оказываются и ближайшими соседями в пространствах векторов сочетаемости. Ср. первые шесть ближайших соседей (в тех дистрибутивных моделях, которые мы описали в пункте 4.2 текущего раздела) существительного яблоко: груша, виноград, апельсин, орех, яблоня, арбуз, а также несколько ближайших соседей лексемы синий: зеленый, желтый, голубой, лиловый, фиолетовый, серый, черный, розовый, коричневый, белый и т.д.

Поля качественных признаков устроены принципиально иначе, и из-за этого те же самые методы дают для них менее удовлетворительные результаты. Один и тот же признак в применении к объектам разного типа может видоизменяться и, как следствие, маркироваться разными словами (как, например, во французском языке, где инструменты с режущим краем описываются прилагательным tranchant, а инструменты с колющим концом – лексемой pointu). В результате оказывается, что разные лексемы одного поля употребляются в разных контекстах и покрывают разные фрагменты, т.е. разные фреймы, этой семантической зоны. Тем самым, наше исходное предположение о том, что разные значения признаковых лексем реализуются в разных контекстах, заранее противоречит как идее составления списка элементов поля по методу ближайших соседей, так и идее использования словарей и баз синонимов для решения поставленной задачи.

Другая сложность заключается в том, что изменение признака под влиянием контекста приводит к отсутствию четких границ между семантическими полями: качественные признаки плавно перетекают друг в друга (см. пример выше о концептуальной близости признаков ‘тугой’ и ‘тяжелый’), что не позволяет этим зонам укладываться в рамки структураллистского понимания семантического поля и существенно затрудняет задачу представления таких зон в тезаурусах и онтологиях вида WordNet.

примерные границы и ближайшие соседние зоны) часто очень трудно определить по материалу одного языка невооруженным глазом. Так, например, английская доминантная лексема поля ‘острый’ *sharp* употребляется в более широком круге контекстов, чем соответствующее русское слово *острый*, что позволяет создателям англоязычного ворднета представить более широкую периферию данного семантического поля, чем разработчикам русской версии аналогичной онтологии. Степень семантической близости между качественными признаками, на первый взгляд кажущимися очень разными (ср. ‘острый’ и ‘крутой’), часто сложно определить априори, и во многих случаях не только внутреннее устройство, но и примерные границы семантического поля являются не исходными данными, а одним из результатов типологического исследования той или иной области.

Дальнейшие эксперименты мы продолжим со списком, полученным по второму варианту методологии обратных переводов (см. Таблицу 15), но в конце этой главы ещё вернемся к проблеме возможного решения задачи определения границ семантического поля.

§4. Составление списков коллокаций

1. Выбор корпуса

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 в большинстве случаев должно быть достаточно двусловных сочетаний вида «прилагательное + существительное».

Автоматически составить перечень таких словосочетаний для заданной признаковой лексемы на материале русского языка можно несколькими способами. Во-первых, благодаря тому, что у определительной конструкции всего две переменных, можно воспользоваться списками биграмм корпорации Google 23 (именно такой материал используется в работах Абдурашитова 2017; Кюсева и др. 2013а; Орехов & Резникова 2015). Биграммы Google разбиты на файлы и отсортированы по алфавиту по первому

---

23 http://storage.googleapis.com/books/ngrams/books/datasetsv2.html
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 начальной форме. Эта задача требует привлечения морфологического парсера и последующей дизамбигуации, что приводит к появлению достаточно большой порции «мусора» – неправильных разборов. Во-вторых, список биграмм 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 априори, какие словосочетания следует отсечь, а какие оставить, – непростая задача (ср. подобные рассуждения в работе Абдурашитова 2017).

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. Этот ресурс позволяет в несколько кликов получить список всех сочетаний искомой лексемы с существительными, приведенными к начальной форме и отсортированными по частотности. Однако, поскольку эта статистическая информация строится на основе очень объемного корпуса интернет-текстов, результирующий список также требует дополнительной фильтрации. Как и в случае с биграммами Google, просто отобрать определенное количество наиболее частотных словосочетаний – не самое удачное решение: 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 операций, 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 RuWaC. Эти тексты уже снабжены морфологической разметкой, так что достаточно посчитать, в сочетании с какими существительными и сколько раз в рамках каждого корпуса встретилась рассматриваемая лексема.

Мы провели соответствующие эксперименты для опорного прилагательного острый. Их результаты показывают, что жанр и степень сбалансированности корпусов довольно значимо влияют на итоговый список коллокаций. Так, набор существительных, сочетающихся со словом острый в рамках газетного подкорпуса НКРЯ, оказался явно смещен 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 острые иглы) показывают сравнительно низкий уровень частотности (см. Таблицу 17).
Таблица 17. Фрагмент списка существительных, с которыми сочетается прилагательное острый (по газетному подкорпусу НКРЯ).

Аналогичная статистика по корпусу RuWaC дает интуитивно более приемлемые результаты, хотя и здесь, как кажется, сказывается недостаточная сбалансированность текстового материала: например, очень частотны и многочисленны названия разного рода

\[24\text{ ipm (instances per million)}\] – метрика, показывающая количество интересующих нас примеров на миллион словоупотреблений в данном корпусе. Мы используем ее вместо абсолютного количества найденных примеров, чтобы можно было напрямую сравнивать между собой уровни частотности одних и тех же языковых единиц в корпусах разных объемов.

\[25\] Для простоты ранги соответствуют сплошной нумерации коллокаций и не учитывают случаев совпадения уровня частотности.
заболеваний, некоторые из которых известны далеко не каждому носителю русского языка (см. Таблицу 18).

<table>
<thead>
<tr>
<th>Существительное, которое встретилось в сочетании с прилагательным острый</th>
<th>Частотность словосочетания (ipm)</th>
<th>Ранг</th>
</tr>
</thead>
<tbody>
<tr>
<td>Лейкоз</td>
<td>0,73</td>
<td>13</td>
</tr>
<tr>
<td>Пиелонефрит</td>
<td>0,44</td>
<td>32</td>
</tr>
<tr>
<td>Инфаркт</td>
<td>0,44</td>
<td>34</td>
</tr>
<tr>
<td>Отравление</td>
<td>0,43</td>
<td>35</td>
</tr>
<tr>
<td>Воспаление</td>
<td>0,35</td>
<td>40</td>
</tr>
<tr>
<td>Бронхит</td>
<td>0,30</td>
<td>50</td>
</tr>
<tr>
<td>Аппендицит</td>
<td>0,30</td>
<td>51</td>
</tr>
<tr>
<td>Панкреатит</td>
<td>0,28</td>
<td>54</td>
</tr>
<tr>
<td>Гастрит</td>
<td>0,22</td>
<td>64</td>
</tr>
<tr>
<td>Инфекция</td>
<td>0,21</td>
<td>67</td>
</tr>
<tr>
<td>Холецистит</td>
<td>0,14</td>
<td>82</td>
</tr>
<tr>
<td>Гепатит</td>
<td>0,13</td>
<td>87</td>
</tr>
<tr>
<td>Ринит</td>
<td>0,13</td>
<td>88</td>
</tr>
<tr>
<td>…</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Таблица 18. Фрагмент списка существительных, с которыми сочетается прилагательное острый (по корпусу RuWaC).

Наконец, список коллокаций, полученный по основному подкорпусу НКРЯ, выглядит наиболее представительно (см. Таблицу 19).
Существительное, которое встретилось в сочетании с прилагательным **острый** | Частотность словосочетания (ipm) | Ранг |
---|---|---|
Угол | 1,96 | 1 |
Боль | 1,43 | 2 |
Ощущение | 0,98 | 3 |
Нож | 0,97 | 4 |
Глаз | 0,97 | 5 |
Нос | 0,88 | 6 |
Вопрос | 0,82 | 7 |
Взгляд | 0,81 | 8 |
Чувство | 0,80 | 9 |
Форма | 0,77 | 10 |
Проблема | 0,75 | 11 |
Зуб | 0,74 | 12 |
Камень | 0,67 | 13 |
Запах | 0,58 | 14 |
Конец | 0,50 | 15 |
Ум | 0,49 | 16 |
Слово | 0,48 | 17 |
Необходимость | 0,47 | 18 |
Подбородок | 0,45 | 19 |
Период | 0,43 | 20 |
Край | 0,42 | 21 |
Желание | 0,42 | 22 |
Характер | 0,39 | 23 |
Дефицит | 0,39 | 24 |
Момент | 0,39 | 25 |
Бородка | 0,37 | 26 |
Нехватка | 0,36 | 27 |
Жальность | 0,34 | 28 |
Потребность | 0,34 | 29 |
Колено | 0,33 | 30 |
 Вершина | 0,32 | 31 |
Приступ | 0,31 | 32 |
Нужда | 0,31 | 33 |
Игла | 0,30 | 34 |
Шип | 0,29 | 35 |

Таблица 19. 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.

Чтобы не опираться исключительно на интуитивные суждения, мы оценили все три списка более строго. Для каждого словосочетания, уровень частотности которого в каждом корпусе был не ниже 0.05 ipm (что соответствует 10 примерам в основном подкорпусе НКРЯ, 7 примерам в газетном подкорпусе НКРЯ и 50 примерам в корпусе RuWaC), мы
указали фрейм, который это словосочетание иллюстрирует, и затем сверили полученный набор фреймов с собранной вручную анкетой. Оказалось, что в списке коллокаций, построенном на материале основного подкорпуса НКРЯ, фигурируют представители всех фреймов, релевантных для русского прилагательного острый, причем на каждый фрейм есть не менее трех примеров. 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 – острове 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 реальность, знакомая носителям далеко не всех культур, поэтому эта строка анкеты часто оставалась незаполненной (например, она по понятным причинам не может быть заполнена данными коми-зырянского или кабардинского языков).

В списке существительных, сочетающихся с прилагательным острый по данным корпуса RuWaC, также оказалось достаточное количество примеров (не менее трех) на все основные фреймы, которые покрываются этим признаком словом. Непредставленными оказались только два маргинальных метафорических значения, отраженных, тем не менее, достаточно большим количеством примеров в НКРЯ: ‘пронизывающий (о ветре, холоде)’, ср. острий холод, и ‘резкий, неприятный для глаз’, ср. остриая вспышка.

Данные газетного подкорпуса также не фиксируют только что упомянутые метафорические употребления прилагательного острый. Помимо этого, в списке существительных, полученном в результате анализа этого корпуса, менее чем тремя примерами представлены и некоторые более продуктивные метафоры: ‘хорошо функционирующий (об органах чувств)’, ср. острове зрение, и ‘хорошо функционирующий (об уме, интеллекте)’, ср. острий ум.

Таким образом, результаты наших наблюдений показывают, что в качестве текстового материала для автоматического составления списка коллокаций надежнее всего использовать основной подкорпус НКРЯ. Это решение также подкрепляется результатами экспериментов, представленных в Главе 2. Именно с этим материалом мы и продолжим наши дальнейшие эксперименты.
Отметим, однако, что для моделирования такого рода сочетаемости подходит и корпус интернет-текстов RuWaC. Этот вывод очень важен, поскольку корпус такого формата построить значительно проще, чем подготовить объемную, сбалансированную и размеченную выборку текстов, подобную той, что представлена в основном подкорпусе НКРЯ: для целого ряда языков есть корпус WaC, но нет национального корпуса, сопоставимого с НКРЯ.

2. Установление порога частотности

Дальнейшим шагом в обработке списка существительных, сочетающихся с целевым прилагательным, является установление порога частотности: необходимо определить, сколько раз словосочетание должно встретиться в корпусе, для того чтобы его можно было включать в исследование. Такой порог необходим по двум причинам. Во-первых, существительных, сочетающихся с искомым прилагательным, слишком много, что может впоследствии негативно повлиять на объем анкеты. 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.

Для установления оптимального порога частотности мы провели ряд экспериментов с разными признаковыми словами. Мы разметили списки коллажей для лексем острый, тяжелый, хороший, шершавый, грубый (для которых у нас уже были разработаны типологические анкеты), собранные по основному подкорпусу НКРЯ, и отметили, в какой момент происходит «насыщение», т.е. до какого элемента списка нужно дойти, чтобы набрать не менее трех иллюстраций на каждый фрейм, релевантный для данного прилагательного (в том числе, метафорический).

Оказалось, что уровень насыщения зависит от двух факторов: частотности самой признаковой лексемы и степени ее многозначности (см. Таблицу 20). С одной стороны, чем частотнее прилагательное, тем частотнее и включающие его коллажи. С другой стороны, чем больше у лексемы значений, тем больше контекстов требуется для их иллюстрации и тем позднее должен наступать момент насыщения. Так, например, для признака острый достаточно уровня частотности, равного 0.08 ipm (что соответствует 17 примерам по основному подкорпусу), в то время как прилагательному тяжелый, обладающему более широкой полисемией, требуется уровень не выше 0.05 (10 примеров).
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 в НКРЯ не менее 100 раз (0.49 ipm).

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. Кроме того, разные значения многозначного слова могут демонстрировать очень разные уровни частотности: например, прилагательное тяжелый в значении интенсификатора (ср. тяжелый грпп) употребляется очень часто, а примеры, иллюстрирующие его использование в значении 'сытный, жирный, калорийный' (ср. тяжелая пища), встречаются на порядок реже.

Из-за сложного соотношения факторов частотности и многозначности подобрать оптимальную формулу для вычисления порогового значения частотности словосочетания 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 эвристики, 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 0.05 ipm (самый низкий порог насыщения во всех наших предварительных экспериментах) и считаем, что существительное, встретившееся в корпусе в сочетании с опорной лексемой меньше этого количества раз, не является его достаточно устойчивой коллокацией.
Таблица 20. Зависимость порога частотности коллокаций от многозначности и частотности опорной лексемы.

<table>
<thead>
<tr>
<th>Лексема</th>
<th>Количество значений по БТС</th>
<th>Частотность прилагательного по НКРЯ (ipm)</th>
<th>Необходимый уровень частотности коллокаций (ipm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Шершавый</td>
<td>2</td>
<td>5,69</td>
<td>0,09</td>
</tr>
<tr>
<td>Грубый</td>
<td>7</td>
<td>59,71</td>
<td>0,09</td>
</tr>
<tr>
<td>Острый</td>
<td>10</td>
<td>78,63</td>
<td>0,08</td>
</tr>
<tr>
<td>Тяжелый</td>
<td>18</td>
<td>200,10</td>
<td>0,05</td>
</tr>
<tr>
<td>Хороший</td>
<td>5(^{27})</td>
<td>454,58</td>
<td>0,5</td>
</tr>
</tbody>
</table>

Таким образом, мы получили экспериментальный материал, состоящий в общей сложности из 13 прилагательных и 1818 примеров их употреблений (слowoсочетаний вида «прилагательное + существительное»). Следующим шагом необходимо разделить набранные контексты на группы, соответствующие разным значениям прилагательных, а затем из каждой группы выбрать по несколько примеров, чтобы сократить объем итоговой анкеты.

§5. Разделение коллокаций на группы

1. Выбор основания для кластеризации и подготовка векторного пространства

Следующий шаг после составления списка коллокаций – их кластеризация, т.е. разделение на семантически однородные группы (будущие фреймы). Для того, чтобы словосочетания можно было разделить на кластеры автоматически, необходимо подобрать параметры, так или иначе отображающие степень семантической близости между типами контекстов употребления рассматриваемых признаковых слов. В качестве такого основания для кластеризации мы использовали метод моделей дистрибутивной семантики, поскольку их приемлемость для решения подобных задач мы уже доказали в рамках предыдущей серии экспериментов (см. Гл. 2), а также в ходе нашего пилотного исследования Рыжова 2014. Кроме того, эффективность использования подобных моделей для решения задач семантической кластеризации была неоднократно доказана в независимых исследованиях (см., например, Dubossarsky et al. 2016).

\(^{26}\) Большой толковый словарь под редакцией С.А. Кузнецова (Кузнецов 1998).
\(^{27}\) 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 хороший! Хорош ругаться! Или: Ты моя хорошая!
Для решения текущей задачи вектора сочетаемости вычислялись в основном по тем же параметрам, что и для предыдущей серии экспериментов (см. Гл. 2):

- в качестве обучающей выборки использовался текстовый материал трех корпусов: основного подкорpusa НКРЯ, газетного подкорпуса НКРЯ и корпуса RuWaC;
- в роли измерений выступали 10 000 наиболее частотных (по основному подкорпусу НКРЯ) лексем знаменательных частей речи (глаголов, существительных, прилагательных и наречий);
- значением по каждой из 10 000 осей было число раз, когда слово-измерение встречалось в корпусе на расстоянии не более ±5 знаменательных слов от опорной единицы;
- к векторам применялась операция взвешивания (по схеме PPMI);
- для упрощения процесса кластеризации размерность всех пространств сокращалась до 300 измерений (по методу SVD).

Так как наша задача – кластеризация словосочетаний, мы должны подготовить векторные представления для двусловных единиц вида «прилагательное + существительное». Теоретически это можно делать двумя способами: признать словосочетание неделимой языковой единицей и анализировать контексты ее употребления в корпусе или составлять векторное представление для словосочетания из векторов входящих в него существительного и прилагательного по одной из моделей композиции (см. раздел 4.1.2 Главы 2).

Как уже было сказано, для реализации первого способа (представления сочетаемости словосочетаний) нужны очень объемные корпуса текстов. В работе Кюсева 2014 продемонстрировано, что даже имеющихся у нас трёх корпусов недостаточно. Уровень корреляции данных русского языка (по дистрибутивной модели, состоящей из векторов сочетаемости для наблюдаемых словосочетаний) и типологических наблюдений растёт по мере увеличения объема корпуса, однако всё равно остаётся на сравнительно низком уровне (от 0.21 для корпуса объемом около 200 млн словоупотреблений до 0.287 для объединенного корпуса объемом почти 1,4 млрд словоупотреблений). При этом если рассматривать не словосочетания в чистом виде, а соответствующие им варианты композиции двух векторов (прилагательного и существительного), то значение корреляции резко подскакивает до 0.7 и выше, что подтверждают также наши эксперименты, представленные в Главе 2. Это наблюдение даёт нам основание считать,
что метод композиции не искажает информацию, а, напротив, позволяет компенсировать недостаток данных.

Дополнительное преимущество метода композиции заключается в том, что он значительно менее ресурсозатратный. Во-первых, в отличие от наблюдаемых словосочетаний, для каждого из которых нужно собирать отдельный вектор, для композиции достаточно один раз собрать вектора для всех необходимых прилагательных 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 (см. Гл. 2), статистику для композиции можно считать только по нему, без привлечения газетного подкорпуса НКРЯ и корпуса RuWaC, что сокращает количество измерений ещё по крайней мере втрое.

В рамках текущей задачи мы тестируем оба способа моделирования сочетаемости двухсловных словосочетаний. Для подготовки векторных представлений словосочетаний как неделимых языковых единиц мы используем все три корпуса текстов. Векторы сочетаемости для существительных и прилагательных, выступающих в качестве основы для последующей композиции, вычисляются на материале основного подкорпуса НКРЯ. В качестве метода композиции мы используем наиболее простую вычислительно, но, тем менее, наиболее надежную (см. Ryzhova et al. 2016) аддитивную модель.

2. Кластеризация векторного пространства

Все методы кластеризации можно разделить на два типа:
1. Алгоритмы, подбирающие оптимальное количество кластеров автоматически;
2. Алгоритмы, требующие указания числа кластеров, на которые следует разбить все данные.

В следующих двух секциях мы представим наши эксперименты по применению алгоритмов обоих типов: с автоматическим определением итогового количества кластеров (раздел 1) и без него (раздел 2).
1. Алгоритмы с автоматическим определением количества кластеров

Теоретически 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 ряд пробных кластеризаций семантических векторов по четырем алгоритмам, не требующим указания числа кластеров: Affinity Propagation (см. Frey & Dueck 2007), Mean-shift (Comaniciu & Meer 2002), DBScan (Ester et al. 1996) и иерархическая кластеризация (Johnson 1966)\(^{28}\).

Результаты работы первых трех алгоритмов (Affinity Propagation, DBScan и Mean-Shift) 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 сразу отказаться от этих методик и продолжить дальнейшие эксперименты только с алгоритмом иерархической кластеризации.

При реализации этого метода мы использовали дефолтный для иерархической кластеризации в модуле SciPy критерий ‘inconsistent’, а в качестве меры близости выбирали косинусное расстояние, как и в предыдущих наших экспериментах. Варьировалось значение только одного параметра – порога кластеризации, – от которого напрямую зависело результирующее число кластеров и объем каждого из них.

Мы провели серию кластеризаций, изменения пороговое значение от 0.1 до 1.2 с шагом 0.1. Чем ниже порог, тем более дробное разбиение исходного списка мы получаем. Материал поля ‘острый’ показывает, что только на уровне 0.6 в итоговом разбиении появляются кластеры, состоящие более чем из двух элементов (при пороге, равном 0.6, такой кластер ровно один; при 0.7 таких кластеров четыре). И напротив, при пороговом значении, большем или равном 1.2, все входные словосочетания объединяются в один кластер, т.е. распределения на группы не происходит совсем. Таким образом, наиболее

удачным диапазоном порогового значения для кластеризации наших данных оказывается промежуток от 0.8 до 1.1.

2. Алгоритмы с заданным числом кластеров

Количество кластеров в этой группе экспериментов мы определяли следующим образом: считали сумму числа значений всех прилагательных нашей выборки (по Малому академическому словарю, см. Евгеньева 1999) и умножали её на два. Тем самым, для наших 13 прилагательных мы получили число 112 (см. Формулу 14). 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 удалить лишнее при последующей обработке.

\[
[9 \text{ (острый)} + 6 \text{ (резкий)} + 2 \text{ (быстрый)} + 1 \text{ (дущий)} + 6 \text{ (горячий)} + 5 \text{ (яркий)} + 2 \text{ (едкий)} + 4 \text{ (колючий)} + 5 \text{ (крупой)} + 2 \text{ (отчетливый)} + 3 \text{ (пронзительный)} + 7 \text{ (высокий)} + 4 \text{ (жаркий)}] \times 2 = 112
\]

Формула 14. Вычисление количества кластеров на основе словарных данных

Для кластеризации с помощью алгоритмов без автоматического определения количества кластеров мы использовали пакет программ Cluto\(^{29}\). Эта библиотека предлагает несколько методов кластерного анализа. Их названия, наборы параметров и основные принципы работы представлены в Таблице 21.

\(^{29}\) http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview
<table>
<thead>
<tr>
<th>Алгоритм</th>
<th>Параметры</th>
<th>Принцип работы</th>
</tr>
</thead>
<tbody>
<tr>
<td>rb (repeated</td>
<td>● Критерий кластеризации;</td>
<td>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.</td>
</tr>
<tr>
<td>bisections)</td>
<td>● критерий выбора кластера, который будет разбиваться на части следующим</td>
<td></td>
</tr>
<tr>
<td>Rbr</td>
<td>● Функция оптимизации;</td>
<td>Берет результат кластеризации по методу rb и оптимизирует его.</td>
</tr>
<tr>
<td></td>
<td>● критерий выбора кластера, который будет разбиваться на части следующим</td>
<td></td>
</tr>
<tr>
<td>Direct</td>
<td>Критерий кластеризации</td>
<td>Пространство объектов сразу разделяется на нужное количество классов</td>
</tr>
<tr>
<td>Agglo</td>
<td>Функции оптимизации</td>
<td>В основе кластеризации – процесс оптимизации некоторой функции</td>
</tr>
<tr>
<td>Graph</td>
<td>● Мера близости между объектами (косинус угла, Евклидово расстояние,</td>
<td>Из пространства объектов строится граф, который затем разбивается на нужное число фрагментов</td>
</tr>
<tr>
<td></td>
<td>коэффициент корреляции, коэффициент Жаккара);</td>
<td></td>
</tr>
<tr>
<td></td>
<td>● количество ближайших соседей;</td>
<td></td>
</tr>
<tr>
<td></td>
<td>● критерий выбора кластера, который будет разбиваться на части следующим</td>
<td></td>
</tr>
<tr>
<td>Bagglo</td>
<td>Функция оптимизации</td>
<td>Конкатенация алгоритмов rb и agglo: сначала применяется метод rb, затем увеличивается размерность пространства и необходимый результат достигается с помощью алгоритма agglo</td>
</tr>
</tbody>
</table>

Таблица 21. Используемые нами алгоритмы кластеризации с заданным числом кластеров.

В ходе настоящего исследования мы провели кластеризацию наших данных, используя все методы, перечисленные в таблице 21, со всеми допустимыми комбинациями параметров. Помимо этого, любой из алгоритмов (кроме agglo и bagglo) можно модифицировать с помощью критериев agglfrom и agglocrfun. Метод agglfrom 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 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.

Помимо этого, мы применили к нашим данным самый распространенный алгоритм кластеризации с заданным числом кластеров K-Means (метод k-средних, см. MacQueen 1967) с помощью библиотеки Scikit-learn. Упрощенно суть работы этого алгоритма сводится к тому, что из кластеризуемого пространства выбирается (по умолчанию случайно) k элементов (где k – заданное число кластеров), а все остальные элементы разделяются на группы (кластеры) в зависимости от того, к какому из этих k элементов они оказываютближе. Затем в каждом из образовавшихся кластеров выбирается центральный элемент, и вся процедура повторяется. Алгоритм работает до тех пор, пока составы кластеров не перестанут меняться. Параметры работы этого алгоритма мы не варьировали и использовали настройки по умолчанию.

§7. Уменьшение объема анкеты

Следующий шаг после разбиения списка контекстов на кластеры – это собственно составление анкеты, т.е. выбор наиболее представительных примеров из каждого кластера. Для решения этой задачи мы выполняем две операции: 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 (потенциально 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 морфологического парсера или последующей морфологической дизамбигуации.

Сокращение всех кластеров до размера 3, напротив, кажется нам продуктивным. С одной стороны, три примера не перегружают анкету. С другой стороны, вполне вероятно, что и из этих трех примеров 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.

Чтобы выбрать наиболее представительные примеры, для каждого кластера мы вычисляли 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. Словосочетания, которым соответствовали эти три вектора, и включались в итоговую анкету (см. примеры сконструированных таким образом анкет в Приложении 4-5).

§8. Оценка результатов

Для автоматической оценки результатов мы разметили вручную все входные словосочетания, т.е. все коллокации частотностью не ниже 10 употреблений в НКРЯ (см. §4 данной главы) для каждого из отобранных нами 13 прилагательных. Разметка заключалась в том, что для каждого словосочетания мы указали, какой фрейм

30 О том, что близость к центру кластера свидетельствует об относительной семантической устойчивости языковой единицы (в частности, о низком уровне ее многозначности), см. Dubossarsky et al. 2016.
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. В ситуациях, когда словосочетание могло относиться сразу к нескольким фреймам, мы указывали все возможности, разделяя их знаком “|”. Фрагмент такой разметки представлен в Таблице 22.

<table>
<thead>
<tr>
<th>Словосочетание</th>
<th>Номер фрейма</th>
</tr>
</thead>
<tbody>
<tr>
<td>острый_нож</td>
<td>1</td>
</tr>
<tr>
<td>острый_игла</td>
<td>2</td>
</tr>
<tr>
<td>острый_подбородок</td>
<td>3</td>
</tr>
<tr>
<td>острый_предмет</td>
<td>1</td>
</tr>
<tr>
<td>острый_боль</td>
<td>15</td>
</tr>
</tbody>
</table>

Таблица 22. Фрагмент экспертной разметки списка словосочетаний.

Далее оценка проводилась по трем параметрам:
1) Общее количество представленных фреймов (= полнота, R);
2) Чистота кластеризации (= точность, P);
3) Доля словосочетаний, не относящихся к рассматриваемому полю.

Первый параметр соответствует метрике полноты и подразумевает проверку того, все ли ожидаемые фреймы проиллюстрированы в данном варианте анкеты. Фрейм считался представленным, если в анкете было хотя бы одно словосочетание, его иллюстрирующее. 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. При этом если словосочетанию, согласно экспертной разметке, соответствовало более одного фрейма и один из них совпадал с фреймами других словосочетаний из того же кластера, мы считали кластер чистым и не штрафовали систему за выбор неоднозначного словосочетания. 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.

106
Для установления оптимального баланса между полнотой и точностью мы считали F-меру по формуле:

$$F = \frac{2PR}{(P+R)}$$

где $P$ – точность, $R$ - полнота

Формула 15. Вычисление F-меры.

В Таблице 23 приведены лучший результат среди алгоритмов с заданным числом кластеров, показанный алгоритмом K-Means (строка 1) и самый высокий показатель для алгоритма иерархической кластеризации (строка 2), оба результата получены на материале пространства векторов словосочетаний, полученных методом композиции. Для сравнения мы также привели данные кластеризации с помощью лучшего алгоритма с заданным числом кластеров из пакета Cluto (rb_agg168i1) для нескольких моделей:

1) кластеризация пространства векторов, собранных по методу композиции, с выбором трех центральных элементов для каждого кластера (строка 3);

2) кластеризация пространства векторов сочетаемости для наблюдаемых словосочетаний (вместо векторов, подготовленных с помощью модели композиции) с выбором трех центральных элементов для каждого кластера (строка 4);

3) кластеризация векторов словосочетаний, собранных по методу композиции, но с выбором трех случайных элементов из каждого кластера (вместо выбора трех центральных), строка 5;

4) кластеризация векторов наблюдаемых словосочетаний с последующим случайным выбором трех элементов кластера (строка 6).

<table>
<thead>
<tr>
<th>Алгоритм</th>
<th>Задано число кластеров</th>
<th>Композиция</th>
<th>Представители кластера</th>
<th>R</th>
<th>P</th>
<th>Доля ненулевых фреймов</th>
<th>F-мера</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-means</td>
<td>да</td>
<td>да</td>
<td>центроиды</td>
<td>0,848</td>
<td>0,883</td>
<td>0,521</td>
<td>0,865</td>
</tr>
<tr>
<td>Иерарх., порог 0,9</td>
<td>нет</td>
<td>да</td>
<td>центроиды</td>
<td>0,879</td>
<td>0,900</td>
<td>0,488</td>
<td><strong>0,889</strong></td>
</tr>
<tr>
<td>rb_agg168i1</td>
<td>да</td>
<td>да</td>
<td>центроиды</td>
<td>0,941</td>
<td>0,586</td>
<td>0,501</td>
<td>0,723</td>
</tr>
<tr>
<td>rb_agg168i1</td>
<td>да</td>
<td>нет</td>
<td>центроиды</td>
<td>0,882</td>
<td>0,514</td>
<td>0,442</td>
<td>0,649</td>
</tr>
<tr>
<td>rb_agg168i1</td>
<td>да</td>
<td>нет</td>
<td>случайные</td>
<td>0,912</td>
<td>0,557</td>
<td>0,493</td>
<td>0,691</td>
</tr>
<tr>
<td>rb_agg168i1</td>
<td>да</td>
<td>нет</td>
<td>случайные</td>
<td>0,824</td>
<td>0,429</td>
<td>0,435</td>
<td>0,565</td>
</tr>
</tbody>
</table>

Таблица 23. Оценка качества кластеризаций: наиболее показательные результаты.
Из Таблицы 23 видно, что самое лучшее качество демонстрирует алгоритм иерархической кластеризации. Он же является оптимальным с вычислительной точки зрения: он не требует выполнения дополнительных операций и привлечения дополнительных ресурсов для определения количества кластеров, на которые должны быть разбиты входные данные. Заметим при этом, что вычисление трех наиболее близких к центру кластера элементов (вместо простого выбора трех случайных словосочетаний) оказывается оправданным (ср. строки 3 и 5). Сопоставление строки 3 со строками 4 и 6 подтверждает тезис о том, что применение моделей композиции для решения задач в области лексической типологии не только допустимо, но и желательно: такая методология и менее ресурсозатратна, и более эффективна. Подчеркнем, однако, что процент попадания в итоговую анкету «лишних» словосочетаний, относящихся к другим семантическим полям, очень высок для всех моделей (см. низкие показатели доли «нужных» словосочетаний в столбце 4).

Среди алгоритмов с изначально заданным числом кластеров самый высокий результат показывает алгоритм K-Means. Этот результат очень близок к полученному с помощью алгоритма иерархической кластеризации, что говорит о том, что теоретически этот алгоритм можно было бы использовать для наших задач. Однако иерархическая кластеризация обладает несомненным преимуществом: она дает возможность определять число кластеров автоматически. Поэтому метод иерархической кластеризации мы считали оптимальным и именно с ним провели ряд дополнительных экспериментов, которые мы обсудим в следующем разделе.

§9. Эксперименты с другими полями

Все дополнительные эксперименты проводились только для того алгоритма, который показал самое высокое качество на материале поля ‘острый’: вектора словосочетания строились с помощью аддитивной модели композиции из векторов входящих в словосочетание слов, векторное пространство разбивалось на кластеры по методу иерархической кластеризации (значение порога кластеризации было равно 1), из каждого кластера выбиралось три центральных элемента, а все кластеры, размер которых не превышал двух словосочетаний, удалялись из итоговой анкеты.

Процедура отбора коллокаций для слов из признаковых полей оставалась такой же, как и для прилагательных зоны ‘острый’: из основного подкорpusa НКРЯ выбирались существительные, встречающиеся непосредственно справа от опорного прилагательного не менее 10 раз, что соответствует уровню частотности в 0.05 ipm. Для глаголов качания процедура отбора коллокаций была несколько модифицирована. Вслед за автором ручного анализа этой семантической зоны, мы рассматривали только один ее фрагмент: одноместные непереходные глаголы (ср. русск. качаться, шататься, болтаться, колебаться, колыхаться, развеваться). Тем самым, диагностический контекст для этих лексем фактически оставался таким же узким, как и для прилагательных, и включал в себя только одно существительное (главный актант). С другой стороны, в отличие от прилагательных, позиция которых в атрибутивной конструкции, как правило, четко закреплена (в русском языке прилагательное почти всегда находится слева от определяемого слова), глагол может располагаться как слева, так и справа от своего актанта, ср.: Девочка качается на качелях и На качелях качается девочка. Ещё одна особенность глаголов качания по сравнению с признаковыми словами, которые мы рассматривали, – их относительно низкий уровень частотности (самая частотная русская лексема в этой зоне – глагол колебаться – занимает позицию 4 802 по словарю Ляшевская & Шаров 2009, самая редкая – развеваться – 14 321).

С целью учета особенностей этой группы слов мы изменили процедуру следующим образом. Мы по-прежнему исходили из допущения, что актант достаточно часто располагается контактно по отношению к предикату, однако учитывали как случаи, когда существительное находится слева от глагола, так и примеры, в которых актант расположен справа, т. е. искали существительное в окне ±1 относительно глагола. При этом мы ввели дополнительное грамматическое ограничение: диагностирующим контекстом считалось не любое существительное, встретившееся рядом с глаголом, а только слово в именительном падеже. Это позволило не учитывать примеры вида болтался головой вниз,
почва под ногами колебалась, на качелях качался... и т. п. Из результирующего списка также были удалены все редкие примеры, однако в этот раз порог был опущен: исключались только существительные, встретившиеся в контексте искомых глаголов менее 3 раз (0.01 ipm).

<table>
<thead>
<tr>
<th></th>
<th>R</th>
<th>P</th>
<th>F-мера</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘качание’</td>
<td>0.882</td>
<td>0.762</td>
<td>0.818</td>
</tr>
<tr>
<td>‘прямой’</td>
<td>1</td>
<td>0.817</td>
<td>0.899</td>
</tr>
<tr>
<td>‘гладкий’</td>
<td>0.8</td>
<td>0.675</td>
<td>0.732</td>
</tr>
<tr>
<td>‘толстый’</td>
<td>1</td>
<td>0.884</td>
<td>0.938</td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th>Кластер 1</th>
<th>Кластер 2</th>
<th>Кластер 3</th>
<th>Кластер 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>прямой столб</td>
<td>прямое участие</td>
<td>прямой потомок</td>
<td>прямая необходимость</td>
</tr>
<tr>
<td>прямая дорожка</td>
<td>прямая поддержка</td>
<td>прямой предшественник</td>
<td>прямая цель</td>
</tr>
<tr>
<td>прямая аллея</td>
<td>прямое руководство</td>
<td>прямое наследие</td>
<td>прямая задача</td>
</tr>
</tbody>
</table>

Таблица 25. Фрагмент кластеризации контекстов лексемы ‘прямой’.

Результаты, приведенные в Табл. 24, а также фрагмент лучшего варианта кластеризации поля ‘прямой’ в Табл. 25 показывают, что подготовленный нами алгоритм работает достаточно стабильно и демонстрирует высокое качество для всех тестовых зон. Тем не менее, качество работы алгоритма все-таки зависит от некоторых параметров исследуемой области.

Во-первых, важную роль играет частотность анализируемых опорных лексем. Чем лексема частотнее (другими словами, чем больше вхождений лексемы в корпус, который обрабатывается алгоритмом), тем более точным будет результат. Как показали эксперименты, если прилагательное в частотном словаре современного русского языка (Ляшевская & Шаров 2009) занимает позицию ниже четырехтысячной, выводу алгоритма нельзя полностью доверять. Так, он предоставляет достаточно эффективный анализ русских слов ‘острый’ и ‘прямой’, которые занимают 1452 и 892 место в словаре, но совсем не так хорошо справляется, например, с обработкой прилагательного ‘просторный’, которое занимает в словаре 4421 позицию. Это слово в русском языке может описывать разные типы объектов, в том числе пространства (просторный зал) и одежду (просторная рубашка). Если первый тип объектов попадает в автоматически созданную анкету, то второй оказывается не включенным в нее. Все случаи сочетания этого прилагательного с
именами, обозначающими одежду, оказываются слишком редки для того, чтобы быть отобранными алгоритмом.

Во-вторых, на результат влияет число фреймов в поле. Чем оно меньше, тем лучше будет проведена кластеризация контекстов на семантически гомогенные группы. Этим объясняется аккуратная кластеризация поля ‘прямой’: в нем семь фреймов, и каждый представлен большим количеством контекстов.

Наконец, на качество анкеты влияет природа оппозиций, организующих семантическую структуру поля. Наш метод автоматического построения анкет ориентирован на группировку контекстов по таксономическим классам. Так, например, в один кластер контекстов поля ‘прямой’ попадают слова потомок и предшественник, относящиеся к классу людей, а в другой – аллея и дорожка, принадлежащие классу протяженных пространств. В большинстве случаев это ведет к желаемому разделению контекстов на фреймы. Однако, не все фреймы признаковых полей противопоставлены друг другу в соответствии с таксономической классификацией существительных. В некоторых случаях решающую роль играет топология предмета. Например, фреймы поля ‘острый’ ‘инструмент с режущим краем (нож, меч)’ и ‘инструмент с колющим концом (игла, шило)’ предполагают один и тот же таксономический класс существительных (инструмент), но разную топологию предмета: с линейным выраженным сегментом в первом случае и с точечным во втором. Такого рода различия оказываются достаточно сложными для алгоритма.

Сравнение анкет для разных полей показало, что описанные выше факторы не являются равноценными. Так, несмотря на то, что в поле ‘толстый’ важную роль играет топологическая классификация предметов, алгоритм предоставил для него наилучший результат (F-мера = 0.938). Вероятно, это объясняется высокой частотностью прилагательных этого поля и небольшим числом фреймов в его семантической структуре. Помимо этого, часто между топологической и таксономической классификацией предметных имен наблюдается корреляция, что также способствует чистоте кластеризации. Так, например, многие части тела попадают в топологический класс длинных вытянутых объектов (толстые пальцы, руки, ноги), а многие предметы одежды – в топологический класс гибких слоев (толстая куртка, колготки, штаны).

Таким образом, дополнительные эксперименты подтверждают, что мы разработали эффективный метод анализа лексических противопоставлений, который позволяет составлять пилотный вариант лексико-типологической анкеты автоматически на
мateriaле одного языка. Этот алгоритм работает более успешно с одними полями (в первую очередь, частотными) и менее успешно с другими лексическими группами, однако даже в последнем случае качество полученных анкет оказывается довольно высоким (так, для наименее удачно расклассифицированного поля ‘гладкий’ значение полноты составило 0.8, а уровень чистоты кластеризации – 0.675).

§10. Выводы

Метод автоматического составления анкет, который мы представили в этой Главе, включает в себя несколько этапов:

1. Определение границ поля (составление списка лексем, относящихся к изучаемому семантическому полю) по методу обратных переводов (см. §3, раздел 4).

2. Составление списка коллокаций для каждой из набранных лексем (в случае с прилагательными и одноместными глаголами – составление списка существительных, сочетающихся с данным признаковым словом / существительных, выступающих в качестве субъекта при данном глаголе) по материалам основного подкорпуса НКРЯ.

3. Представление каждого словосочетания (вида «прилагательное + существительное» или «глагол + существительное») в виде вектора сочетаемости (методом композиции итогового вектора из векторов прилагательного/глагола и существительного).

4. Кластеризация векторного пространства по методу иерархической кластеризации.

5. Выделение трех центральных элементов из каждого кластера и удаление всех кластеров, размер которых не превышает двух элементов.

Первая задача лучше всего решается методом обратных переводов. Однако, даже несмотря на то, что этот метод не полностью автоматический и требует вмешательства исследователя, результат, который он дает, оказывается недостаточно удачным. Это видно, в частности, по низкой доле словосочетаний, действительно относящихся к полю ‘острый’, в итоговых автоматически сгенерированных анкетах, см. Табл. 23. С теоретической точки зрения это означает, что строгих границ между семантическими полями, по-видимому, действительно не существует (по крайней мере, в зоне качественных признаков, но мы предполагаем, что то же самое будет верно и для многих других областей, в частности, глагольных), и именно поэтому задача их автоматического определения не находит хорошего решения. С практической же точки зрения, на наш взгляд, будет
целесообразно совсем вывести эту задачу за пределы алгоритма автоматического составления лексико-типологических анкет и оставить возможность лингвисту-эксперту самостоятельно определять круг лексем, которые он хотел бы включить в исследование.

Все остальные этапы работы предлагаемого нами алгоритма полностью автоматические. Мы разработали его на материале признаковых слов, но ряд пилотных экспериментов показывает, что он может быть успешно использован при создании анкет для исследования и других классов лексики (см. написанные под нашим руководством курсовые работы Лучина 2015 о применении подобного алгоритма к анализу наречия и частицы прямо и Мельник 2018 о работе с многоместными глаголами).

При переходе к анализу других семантических полей наибольшим изменениям подвергается этап составления списка коллокаций. Прежде всего, менее частотные опорные лексемы (в том числе прилагательные) могут требовать более низкого порога частотности для словосочетаний, включаемых в список рассматриваемых контекстов (см. обсуждение глаголов качания и прилагательного просторный выше).

Помимо порога частотности, изменений потребует и тип учитываемого контекста. Для анализа семантики прилагательных в большинстве случаев достаточно учитывать определяемое ими существительное. Для одноактантных глаголов, согласно нашему пилотному эксперименту, в качестве диагностического контекста можно рассматривать существительное, занимающее позицию субъекта. Для анализа многоактантных глаголов или предметных имен, у которых, наоборот, нет никаких валентностей, определение типа минимального диагностического контекста требует дополнительных исследований. По-видимому, для представления семантики такого типа слов необходимо принимать во внимание не только их лексический контекст, но и морфологическое оформление их актантов – для глаголов – или их собственные грамматические характеристики и конструкции, в которые они встраиваются – для существительных (ср. успешное применение подобных методик для решения задач, близких к нашей, в работах Divjak & Gries 2006; Janda & Solovyev 2009). Все остальные этапы работы алгоритма могут оставаться неизменными.

С теоретической точки зрения, предложенный нами алгоритм составления лексико-типологических анкет, – это следующий шаг после экспериментов, которые мы обсуждали в Главе 2. Если в Главе 2 мы говорили о том, что фреймовая структура поля так или иначе видна уже сквозь призму одного языка, то здесь мы применили это утверждение на практике: раньше мы проверяли, насколько дистрибутивное пространство сопоставимо с
типологическим, то теперь мы смогли воспроизвести типологическое пространство с помощью дистрибутивного, не имея никаких данных, кроме корпуса текстов на одном языке и примерного представления о том, какие слова в русском языке относятся к рассматриваемому полю.

С практической точки зрения это означает, что процесс составления предварительного варианта анкет можно очень существенно ускорить. Если отказаться от этапа автоматического определения круга лексем, относящихся к рассматриваемому полю, то алгоритму нужно будет подавать список таких слов на вход, а на выходе будут получаться группы из трех словосочетаний каждая, иллюстрирующие разные классы употреблений этих слов. Поскольку качество работы алгоритма не абсолютно идеальное, сконструированные таким образом анкеты нуждаются в экспертной проверке: возможно, потребуется убрать ненужные примеры или объединить несколько кластеров в один. Однако все эти действия могут быть выполнены очень быстро, без привлечения каких бы то ни было дополнительных ресурсов. Таким образом, в ближайшем будущем разработанный алгоритм можно будет использовать для массового производства фреймовых анкет, что позволит вывести лексическую типологию на новый уровень.
Глава 4. Методы автоматического сбора данных

§1. Постановка задачи

В работах по лексической типологии, предлагающих тот или иной метод автоматического сбора данных, этот этап исследования, как правило, совмещается с задачей составления анкеты. Так, например, в уже цитируемой нами выше работе Wälchli & Cysouw 2012 данные о дистрибуции базовых глаголов движения собираются автоматически по параллельному корпусу текстов Евангелия от Марка. При этом стихи, в которых может быть употреблен один из базовых глаголов движения, и выступают в качестве своего рода анкеты, т.е. базы для сравнения материала разных языков. В работах, основанных на обработке словарей (ср. Sejane & Eger 2013; Youn et al. 2016), сопоставление полученных данных производится на основе толкований (переводов), извлеченных из тех же самых словарных источников.

У таких методик множество достоинств, в частности, они предполагают полную или, по крайней мере, частичную автоматизацию цикла лексико-типологического исследования. Но одно из их ограничений заключается в том, что они позволяют учитывать информацию только из одного типа источников. Так, например, непонятно, как можно было бы дополнить сведения, почерпнутые из параллельных корпусов по методике Б. Вельхли и М. Сисоу, материалами значительно больших по объему одноязычных корпусов текстов.

Фреймовый подход к лексической типологии, на который мы опираемся в нашем исследовании, напротив, предполагает комплексный анализ лексических единиц, включающий в себя обработку всех доступных ресурсов. Поэтому мы ставим перед собой задачу, несколько отличающуюся от задач, которые решали наши предшественники. А именно, мы стремимся разработать серию алгоритмов, принимающих на вход уже готовую лексико-типологическую анкету и заполняющих ее материалами доступных параллельных и одноязычных корпусов. Тем самым, мы существенно сужаем задачу по сравнению с рассмотренными выше методиками, подразумевающими полный лексико-типологический анализ, но надеемся, что это позволит нам создать инструмент для более детального

---

31 Разработка методов, представленных в этой главе, проводилась преимущественно в рамках курса «Практикум по проектированию и разработке лингвистических систем и компонентов» в НИУ ВШЭ. В исследованиях участвовали студенты НИУ ВШЭ А. Мельник, И. Ершов, М. Соболов, И. Пантелейева, А. Кошевой. Основные результаты совместной работы отражены в статьях Рыжова и др. 2017; Ryzhova et al. 2018.
анализа лексики, в том числе значительно менее базовой и частотной, чем предметные имена из списка Сводеша или основные глаголы движения.

Поскольку в рамках фреймового подхода к лексической типологии анкеты представляют собой наборы прототипических контекстов, в которых могут употребляться слова изучаемого поля, такие анкеты необходимо переводить на каждый язык, включаемый в выборку. Поэтому фактически задача автоматического заполнения анкет данными некоторого нового языка включает в себя два этапа: 1) перевод анкеты, получаемой на вход, и 2) заполнение переведенного опросника.

Эта глава имеет следующую структуру. В §2 дается описание материала для экспериментов и золотого стандарта, с которым мы сравниваем результаты работы компьютерных моделей, а также уточняются задачи, которые мы перед собой ставим. Параграф 3 посвящен описанию алгоритмов автоматического перевода (раздел 3.1) и заполнения готовых анкет (раздел 3.2). В §4 подводятся итоги: обсуждаются достоинства и недостатки предложенного метода и возможные пути его дальнейшего развития.

§2. Материал для экспериментов

Как и в экспериментах, описанных выше, для разработки пилотного варианта алгоритма сбора данных мы выбрали материал качественных признаков.

Поскольку минимальным диагностическим контекстом для признаковой лексики такого типа в большинстве случаев является определяемое существительное, анкета для исследования той или иной признаковой зоны может быть представлена в виде таблицы, в строках которой располагаются существительные, а в столбцах – изучаемые прилагательные (см. Табл. 26). При заполнении анкеты в те ячейки, которые соответствуют допустимым (грамматическим, осмысленным и естественным) сочетаниям «прилагательное + существительное», ставится «1» (или «+»), а ячейки, которые соотносятся с незасвидетельствованными сочетаниями, получают значение «0» (или «-»). Например, в анкете для поля фактуры поверхностей, заполненной материалом русского языка, фрагмент которой представлен в Таблице 26, отмечено, что русск. дорога сочетается с прилагательными скользкий, ровный, гладкий и прямой, а ступенька – с прилагательными скользкий, ровный и гладкий.
Заполнение таких анкет требует очень много усилий со стороны исследователя. Задача нашей работы — предложить набор алгоритмов, который будет позволять обследовать по крайней мере часть источников автоматически и, тем самым, расширять эмпирическую базу таких исследований, одновременно сокращая количество времени, потраченное на сбор релевантных данных.

Таким образом, задачи, которые нам необходимо решить на данном этапе, в применении к материалу признаковой лексики могут быть сформулированы так:

1) Перевод списка прилагательных, относящихся к рассматриваемому полю, и списка существительных, с которыми они потенциально могут сочетаться;
2) Заполнение этой анкеты (т.е. таблицы) сведениями о том, могут ли рассматриваемые прилагательные и существительные в данном конкретном языке сочетаться друг с другом.

В качестве материала для экспериментов мы используем созданные вручную и апробированные на достаточно обширном языковом материале анкеты для четырех признаковых полей: ‘острый’, ‘толстый’, ‘тонкий’ и ‘гладкий’ (см. Kashkin & Vinogradova to appear; Kozlov & Privizentseva to appear; Kyuseva et al. to appear). Изначально эти анкеты состояли из списков русских прилагательных и существительных. Мы дополнительно перевели их вручную на итальянский и английский языки и для этих языков заполнили их традиционным способом: по материалам словарей, корпусов и опросов носителей. Таким образом мы получили золотой стандарт, с которым в дальнейшем сравнивали качество работы наших алгоритмов. Там, где это было возможно, мы использовали в качестве золотого стандарта в том числе и данные, собранные ранее также вручную авторами исследований этих семантических полей и представленные в едином формате Базы данных признаковой лексики (Кюсева, Резникова и Рыжова 2013b). В результате англоязычные анкеты мы использовали в качестве входных данных наравне с русскоязычными. Полный цикл работы алгоритмов с последующей оценкой результатов был реализован для двух пар языков: 1) английский в качестве исходного языка и русский в качестве целевого; 2) английский/русский в качестве исходных языков (в зависимости от доступных ресурсов).

Таблица 26. Фрагмент заполненной анкеты для признакового поля фактуры поверхностей.

<table>
<thead>
<tr>
<th></th>
<th>скользкий</th>
<th>ровный</th>
<th>плоский</th>
<th>гладкий</th>
<th>прямой</th>
</tr>
</thead>
<tbody>
<tr>
<td>дорога</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>ступенька</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Заполнение таких анкет требует очень много усилий со стороны исследователя. Задача нашей работы — предложить набор алгоритмов, который будет позволять обследовать по крайней мере часть источников автоматически и, тем самым, расширять эмпирическую базу таких исследований, одновременно сокращая количество времени, потраченное на сбор релевантных данных.

Таким образом, задачи, которые нам необходимо решить на данном этапе, в применении к материалу признаковой лексики могут быть сформулированы так:

1) Перевод списка прилагательных, относящихся к рассматриваемому полю, и списка существительных, с которыми они потенциально могут сочетаться;
2) Заполнение этой анкеты (т.е. таблицы) сведениями о том, могут ли рассматриваемые прилагательные и существительные в данном конкретном языке сочетаться друг с другом.

В качестве материала для экспериментов мы используем созданные вручную и апробированные на достаточно обширном языковом материале анкеты для четырех признаковых полей: ‘острый’, ‘толстый’, ‘тонкий’ и ‘гладкий’ (см. Kashkin & Vinogradova to appear; Kozlov & Privizentseva to appear; Kyuseva et al. to appear). Изначально эти анкеты состояли из списков русских прилагательных и существительных. Мы дополнительно перевели их вручную на итальянский и английский языки и для этих языков заполнили их традиционным способом: по материалам словарей, корпусов и опросов носителей. Таким образом мы получили золотой стандарт, с которым в дальнейшем сравнивали качество работы наших алгоритмов. Там, где это было возможно, мы использовали в качестве золотого стандарта в том числе и данные, собранные ранее также вручную авторами исследований этих семантических полей и представленные в едином формате Базы данных признаковой лексики (Кюсева, Резникова и Рыжова 2013b). В результате англоязычные анкеты мы использовали в качестве входных данных наравне с русскоязычными. Полный цикл работы алгоритмов с последующей оценкой результатов был реализован для двух пар языков: 1) английский в качестве исходного языка и русский в качестве целевого; 2) английский/русский в качестве исходных языков (в зависимости от доступных ресурсов).
и итальянский язык в качестве целевого. В качестве дополнительного материала в процессе тестирования некоторых этапов работы алгоритма привлекались данные ряда других языков.

§3. Алгоритмы автоматического заполнения типологической анкеты

Как мы уже сказали в предыдущем разделе, задача автоматического заполнения типологической анкеты включает в себя две подзадачи: перевод анкеты с русского (или английского) языка на язык L и собственно заполнение переведенной анкеты. В разделе 3.1 мы предлагаем обзор нескольких способов решения задачи перевода и обсуждаем их достоинства и недостатки. В разделе 3.2 приводится описание используемого нами метода автоматического заполнения анкеты.

3.1. Перевод анкеты

Как уже было сказано ранее, задача перевода анкеты в нашем случае сводится к поиску переводных эквивалентов для двух списков слов: прилагательных и существительных. Традиционно задача перевода (в том числе автоматического) отдельных слов подразумевает один из следующих вариантов решения: (1) выбор самой подходящей лексемы для определенного контекста; (2) подбор наиболее частотного / близкого переводного эквивалента для данной лексемы; (3) выдачу самого удачного эквивалента для каждого из значений исходного слова. Если для перевода существительных нам подошел бы вариант (2), то в случае с прилагательными цель отличается ото всех перечисленных: нам необходимо подобрать такие лексемы, которые могут выступать в качестве переводов исходных слов, но только в контекстах, соответствующих их прямым употреблением. Так, например, в числе английских переводных эквивалентов слова острый мы бы хотели видеть прилагательные sharp и pointed, но не critical или urgent (ср. острые ножи или острые волосы). Иными словами, на этом этапе мы снова сталкиваемся с проблемой определения границ поля, только теперь мы исходим из предположения, что пределы зоны так или иначе очерчены на материале стартового языка, и в нашу задачу входит выделение эквивалентного семантического фрагмента в целевом языке.

В результате выполнения первого этапа работы алгоритма создается csv-таблица: в первый столбец таблицы записываются существительные на языке, с которого осуществлялся перевод, во вторую — на том языке, на который мы переводим анкету, а в первую строку — прилагательные целевого языка. В клетки пересечения между
прилагательным и существительным мы впоследствии, на втором шаге работы алгоритма, записываем информацию о сочетаемости соответствующих единиц (см. Табл. 27).

<table>
<thead>
<tr>
<th></th>
<th>maigre</th>
<th>gros</th>
<th>mince</th>
<th>dense</th>
<th>étroit</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>лист</td>
<td>feuille</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>слой</td>
<td>poussière</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>книга</td>
<td>livre</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Таблица 27. Пример анкеты для признаковых полей ‘толстый’ и ‘тонкий’, переведенной на французский язык.

Мы опробовали три метода решения задачи перевода отдельных слов: (1) с помощью онлайн-переводчиков, предлагаемых Yandex и Google; (2) через машиночитаемые двуязычные словари; (3) с помощью параллельных корпусов. Методы перевода существительных и прилагательных в каждом случае несколько различались. Рассмотрим последовательно каждую методологию.

2.1.1. Онлайн-переводчики компаний Yandex и Google

В рамках нашего исследования мы провели ряд экспериментов по переводу анкеты для нескольких признаковых полей на основе онлайн-переводчиков, реализованных компаниями Google и Yandex.

Алгоритм перевода существительных в этом случае предельно прост: мы подаем переводчику на вход последовательно все существительные из анкеты и для каждого из них извлекаем первый вариант перевода.

Метод перевода прилагательных несколько сложнее. Поскольку количество признаковых лексем целевого языка может отличаться от количества слов, обслуживающих данную зону в исходном языке, простой выбор первого варианта перевода для каждого прилагательного неизбежно повлечет за собой потерю информации. Поэтому, в случае с прилагательными, мы принимаем во внимание все слова того синонимического ряда, который выдает переводчик. Здесь мы сталкиваемся с той же проблемой, которая возникала у нас на этапе определения списка слов, относящихся к изучаемому полю в русском языке (см. §3 Главы 3): синонимы могут относиться как к прямым, так и исключительно к метафорическим значениям исходных слов, в то время как
на данном этапе нас интересуют только прямые. Чтобы из всего множества синонимов отобрать релевантные для нашей задачи, мы вновь прибегаем к методу обратных переводов (ср. пункт 3.4.4 Главы 3). Мы переводим обратно на русский язык все синонимы, полученные при переводе с русского языка на целевой, и оставляем в анкете для целевого языка только те прилагательные, перевод которых на русский дает какое-нибудь из прилагательных исходного списка.

Например, мы хотим узнать, какие прилагательные составляют поле ‘острый’ в финском языке. Тогда, задав онлайн-переводчику компании Yandex в качестве исходного слова русское прилагательное острый, мы получаем такие переводы (8 финских слов): akuutti, äkillinen, terävä, pistävä, kirpeä, pureva, veitsenterävä, tarkka. Далее мы переводим каждое из полученных слов обратно на русский язык и оставляем только те, которые переводятся прилагательным острый. Таким образом мы выбираем 5 финских лексем (akuutti, terävä, pistävä, kirpeä, veitsenterävä) и исключим из рассмотрения 3 оставшихся (äkillinen, pureva и tarkkea), поскольку в первом значении у них указаны переводы внезапный, кусающийся и точный, аккуратный, пунктуальный соответственно. Согласно типологической базе данных признаковой лексики (Кюсева, Резникова и Рыжова 2013b), в финском языке в поле острый входят прилагательные terävä, pistävä, и наш алгоритм верно их выявил, но при этом зафиксировал и три лишних прилагательных akuutti (сравнительно с русск. острый в острый взгляд), kirpeä (острый в значении пикантный), veitsenterävä (окказиональные употребления).

В качестве другого возможного способа отсева лишних прилагательных мы протестировали метод перевода прилагательного в контексте с существительным. Для этого мы попробовали перевести все допустимые в русском языке сочетания прилагательного и существительного из анкеты с помощью онлайн-переводчиков. Для более наглядной иллюстрации этого метода снова приведем пример перевода русского прилагательного острый на финский язык, но уже в сочетании с каждым из рассматриваемых существительных (острый нож, острый каблук, острия стрела). Из каждого полученного словосочетания на финском выберем прилагательное и с помощью обратного перевода аналогичным способом отберем только те, у которых указано прилагательное острый в качестве основного переводного эквивалента. В результате мы получаем только два финских прилагательных, одно из которых оказывается нужным (terävä), а другое (akuutti) – ненужным. Таким образом мы уменьшим количество ненужных прилагательных, но в то же время и уменьшим количество нужных. То же самое
верно и для других языков: в работе Singh 2017 показано, что перевод англоязычных словосочетаний из анкеты для поля ‘острый’ на хинди с помощью онлайн-переводчика корпорации Google позволяет получить только 4 признаковых слова из 8, причем далеко не все словосочетания, выдаваемые автоматическим переводчиком, действительно грамматичны.

2.1.2. Машиночитаемые словари Freedict и Verdict

Наш второй метод перевода анкет основан на использовании машиночитаемых двуязычных словарей, а именно, переводных словарей группы Freedict32 и электронных словарей Verdict, подготовленных компанией Yandex.

Двуязычные словари Freedict хранятся в формате .tei 33 (который является подмножеством языка xml), структура документа определяется парами тегами. Каждое вхождение выделяется тегом <entry>. В него вложены теги <form> и <sense>. В зависимости от количества возможных переводов, может быть несколько вложений с тегом <sense>: в таком случае рядом с тегом записывается показатель n="…", в котором в кавычках указывается номер возможного перевода. Таким образом создается структура, отображающая связь между запрашиваемым словом данного языка и его возможными переводными эквивалентами.

Мы предполагаем, что перевод с показателем n="1" является основным, соответствующим прямому употреблению заданного слова, а остальные переводы связаны с его переносными значениями. Именно поэтому в качестве перевода для исходного прилагательного мы выбираем эквивалент с показателем n="1". Далее, как и в случае с онлайн-переводчиками, производится проверка методом обратного перевода. В качестве исходного прилагательного задается полученный переводной эквивалент и осуществляется перевод на стартовый язык. Если слово входит в начальный список прилагательных, то данный переводной эквивалент включается в итоговый список слов. Для перевода существительных проводится та же операция. Однако, если искомого слова нет в словаре Freedict, то используется соответствующий двуязычный словарь Verdict. Он имеет более простую структуру. Одно вхождение занимает одну строку и имеет всего четыре позиции, разделенные табуляцией: исходное слово, его часть речи, его переводной эквивалент и ссылка на словарь, из которого данная информация получена. Если вариантов

32 Словари доступны по ссылке http://www.freedict.org/ru/.
33 http://www.tei-c.org/index.xml
перевода несколько, то для каждого из них в словаре заводится новая строка. При этом кандидаты на перевод в явном виде никак не ранжируются, поэтому наш алгоритм просто выбирает переводной эквивалент, указанный в строке, располагающейся выше других в файле.

2.1.3. Параллельные корпуса

Метод перевода существительных и прилагательных с помощью параллельных корпусов был разработан Я. Сингом под руководством Д.А. Паперно в процессе решения задачи перевода нескольких анкет для признаковых семантических полей с английского на хинди (Singh 2017). Мы реализовали этот алгоритм на материале английско-русского и русско-итальянского параллельных подкорпусов НКРЯ.

Тексты в параллельных подкорпусах НКРЯ выровнены по предложениям и снабжены морфологической разметкой. Англо-русский (объединенный с русско-английским) корпус содержит 527 782 предложения, русско-итальянский (с итальянско-русским) – 101 814 предложений, т.е. объем второго подкорпуса существенно меньше первого.

Переводные эквиваленты для заданных слов определяются на основе так называемой меры зависимости, или зависимостной вероятности (Dependency Probability, далее D.P.), см. Формулу 16. Для каждой пары «слово на стартовом языке (W1) – кандидат на перевод (W2)» эта метрика отражает вероятность появления в предложении на целевом языке слова W2 в том случае, если в предложении на стартовом языке присутствовало слово W1.

В качестве кандидатов на перевод рассматриваются все слова соответствующей части речи (существительные – при переводе существительных, прилагательные – при переводе прилагательных), зафиксированные в корпусе текстов на целевом языке.

\[ D.P. = \frac{P(W2 \mid W1)^2}{P(W1) \cdot P(W2)} , \]

где W1 – кандидат на перевод, W2 – исходное слово, P(W2|W1) – вероятность наличия слова W2 в предложении на целевом языке, если в предложении на исходном языке есть слово W1.

Формула 16. Мера зависимости (Dependency Probability, D.P.).

---

34 Мы благодарим Д.В. Сичинаву за предоставленный нам доступ к параллельным корпусам.
Далее берется обратное значение меры зависимости (1/D.P.), и все кандидаты на перевод ранжируются в зависимости от значения этой метрики, от самого низкого к самому высокому, т.е. от наиболее вероятного к менее вероятному кандидату. На Рис. 7 представлен график, демонстрирующий изменение значения меры зависимости от ранга к рангу для пятнадцати наиболее вероятных переводных эквивалентов в языке хинди для английского прилагательного *sharp* ‘острый’. По графику видно, что у первых восьми прилагательных очень близкие значения метрики зависимости, после чего происходит резкий скачок. Я. Синг предлагает считать подходящими переводными эквивалентами для прилагательных все слова, соответствующие первым нескольким рангам. Число включаемых рангов зависит от момента, в котором происходит скачок. Так, например, вариантами перевода прилагательного *sharp* на хинди будут считаться первые восемь слов, поскольку ровно столько точек расположено на графике слева от скачка. Существительные переводятся аналогичным образом, только в качестве оптимального переводного эквивалента для стартового слова считается существительное целевого языка, занимающее первый ранг.

![Image](image_url)

Рисунок 7. Значение метрики 1/D.P. для первых пятнадцати кандидатов на перевод в языке хинди для английского прилагательного *sharp* (Singh 2017: 2).

2.1.4. Анализ результатов

Мы рассмотрели три возможных метода перевода анкеты. Метод онлайн-переводчиков хорош тем, что он позволяет сразу охватить большое количество языков, на вход можно сразу подавать как отдельные слова, так и целые словосочетания из исходных анкет, и он довольно прост в использовании. Тем не менее, у этого метода есть ряд существенных недостатков. Во-первых, он непрерывен и непостоянен: технологии перевода у компаний Yandex и Google непрерывно меняются, и мы не можем быть уверены,
что в любой момент времени сможем получить результаты, сопоставимые с теми, что мы
получили в рамках наших пилотных экспериментов. Во-вторых, метод перевода целых
словосочетаний оказывается бесполезным для нашей задачи: почти все пары
«прилагательное + существительное» переводятся с помощью одного, основного элемента
признакового поля, и тем самым сразу несколько периферийных прилагательных
рассматриваемой зоны оказываются неучтенными. Наконец, в-третьих, разные стратегии
перевода с помощью онлайн-переводчиков дают для разных полей разные результаты.

Второй метод – метод машиночитаемых словарей – лишен этих недостатков. Более
того, словари группы Freedict обладают дополнительным преимуществом: эксплицитно
противопоставляют переводные эквиваленты для прямых и переносных значений
исходных слов. Однако данный метод имеет свои недостатки, связанные, в первую очередь,
с небольшим количеством таких словарей и ограниченностью их объёма. Так, например,
не очень частотное прилагательное из рассматриваемого поля в такой словарь не попадет,
а значит, предлагаемый нами алгоритм не сможет его выявить.

Наконец, третий метод – перевод с помощью параллельных корпусов – кажется
наиболее надежным, поскольку он прозрачно отображает статистику употребления
переводных эквивалентов в одинаковых контекстах, но он очень сильно зависит от
доступных ресурсов.

Как видно, у всех трех методов есть свои достоинства и свои недостатки. Первый
метод подлежит дальнейшей проверке на материале большего количества анкет, второй и
третий позволяют получать приемлемые результаты, но только для тех языков, по
которым мы имеем необходимые словари или корпуса. Количественная оценка
результатов перевода существительных (Табл. 28, поля ‘острый’ и ‘гладкий’) и
прилагательных (Табл. 29, зоны ‘гладкий’, ‘толстый’, ‘тонкий’) показывает, что на данном
этапе для перевода предметных слов больше всего подходит метод, опирающийся на
переводчик компании Google, а для перевода признаковой лексики – метод, использующий
параллельные корпуса (в том случае, если они приемлемого объема).

<table>
<thead>
<tr>
<th></th>
<th>Словари</th>
<th>Google.Translate</th>
<th>Параллельный корпус</th>
</tr>
</thead>
<tbody>
<tr>
<td>en-ru</td>
<td>0.38 / 0.5</td>
<td><strong>0.80 / 0.68</strong></td>
<td>0.393 / 0.417</td>
</tr>
<tr>
<td>en/ru-it</td>
<td>0.31 / 0.59</td>
<td><strong>0.64 / 0.63</strong></td>
<td>0.5 / 0.396</td>
</tr>
</tbody>
</table>

Таблица 28. Точность перевода существительных для полей ‘острый’/‘гладкий’
Таблица 29. Точность (P) и полнота (R) для перевода прилагательных, поля ‘гладкий’/ ‘толстый’/ ‘тонкий’.

<table>
<thead>
<tr>
<th></th>
<th>Словари</th>
<th>Google.Translate</th>
<th>Параллельный корпус</th>
</tr>
</thead>
<tbody>
<tr>
<td>en-ru</td>
<td>P</td>
<td>0 / 0.5 / 0</td>
<td>0.3 / 0.67 / 1</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>0 / 0.33 / 0</td>
<td>0.19 / 0.36 / 0.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1 / 1 / 0.67</td>
</tr>
<tr>
<td>en/ru-it</td>
<td>P</td>
<td>0 / 0.5 / 0.66</td>
<td>0.55 / 1 / 0.67</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>0 / 0.16 / 0.4</td>
<td>0.26 / 0.3 / 0.14</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.09 / 0.375 / 0.25</td>
</tr>
</tbody>
</table>

Таблица 29. Точность (P) и полнота (R) для перевода прилагательных, поля ‘гладкий’/ ‘толстый’/ ‘тонкий’.

2.2. Заполнение анкеты

Одна из основных задач исследования в области лексической типологии – это оценка возможности употребления слов описываемого ряда в различных контекстах. Имея исходные русско- или англоязычные списки прилагательных и существительных, с которыми эти прилагательные потенциально могут сочетаться, а также полученные на первом шаге работы нашего алгоритма соответствующие им переводы на другие языки, мы можем автоматически определять дистрибутивные свойства каждого признакового слова по отношению к выделенным предметным именам. Для этого мы используем одноязычные корпуса серии WaC (Baroni et al. 2009). Предложения в корпусах этого типа хранятся в формате .xml и структурируются по принципу xml-разметки. Каждое предложение выделяется тегом <s>. Внутри тега прописываются словоформы, которые содержат данное предложение. Информация о каждой словоформе занимает одну строку и содержит словоформу, лемму, морфологическую информацию о данной словоформе, номер данной словоформы в предложении, номер синтаксической вершины, к которой относится данное слово, и его синтаксическую категорию, разделенные табуляцией.

Поскольку в выбранных нами полях изучаемые слова – прилагательные, а их диагностические контексты – существительные, для того, чтобы оценить вероятность их совместной встречаемости, мы ищем вхождения всех комбинационно возможных биграмм “прилагательное + существительное” из анкеты в корпусе исследуемого языка. Для этого мы ищем такие лексемы, леммы которых совпадают с искомыми (поскольку искомое существительное может встретиться в тексте как, скажем, во множественном числе, так и в единственном), которые находятся на расстоянии один друг от друга.

Для того, чтобы исключить окказиональные и в целом нехарактерные для языка сочетания, мы вычисляем для каждой биграммы характеризующую ее меру взаимной информации (pointwise mutual information, далее PMI), показывающую, насколько совместная встречаемость двух слов статистически значима в данном языке. Значение этой метрики вычисляется по следующей формуле:
Формула 17. Мера взаимной информации (pointwise mutual information).

Введение порога для PMI позволяет оставить контекстно-связанные слова и исключить такие, которые приравниваются к случайно встретившимся рядом в тексте. Мы установили пороговое значение, равное нулю, т.е. фактически мы используем метрику положительной взаимной информации (positive pointwise mutual information, PPMI).

Например, если мы хотим оценить допустимость употребления в итальянском языке существительного corridoio 'коридор' в сочетании с прилагательным stretto 'узкий', то мы будем искать в корпусе два идущих подряд слова с леммами corridoio и stretto. В корпусе ItWaC находится 47 таких вхождений с мерой PMI равной приблизительно 8.7, достаточной для того, чтобы считать сочетание corridoio stretto допустимым при взятом нами пороге, равном нулю. Такой метод позволяет заполнять анкеты для частотных признаковых слов относительно аккуратно (см. Табл. 30). Однако заметим, что качество работы этого модуля нашего алгоритма во многом зависит от объема и степени сбалансированности используемых одноязычных корпусов.

<table>
<thead>
<tr>
<th></th>
<th>‘толстый’</th>
<th>‘тонкий’</th>
</tr>
</thead>
<tbody>
<tr>
<td>RU</td>
<td>P 0.75</td>
<td>0.79</td>
</tr>
<tr>
<td></td>
<td>R 0.87</td>
<td>0.92</td>
</tr>
<tr>
<td>IT</td>
<td>P 0.4</td>
<td>0.54</td>
</tr>
<tr>
<td></td>
<td>R 0.7</td>
<td>0.68</td>
</tr>
</tbody>
</table>

Таблица 30. Точность (P) и полнота (R) для задачи заполнения анкет (поля ‘толстый’ и ‘тонкий’).

§4. Выводы

Итак, мы провели серию пилотных экспериментов в области автоматизации процесса сбора лексико-типологических данных с семантическими полями качественных признаков ‘острый’, ‘гладкий’, ‘толстый’ и ‘тонкий’, которые уже были проанализированы вручную участниками группы MLexT. Мы показали, что эта задача включает в себя две подзадачи: перевод анкеты и ее заполнение. Первая подзадача подразумевает перевод списков существительных и прилагательных. Мы опробовали три метода ее решения, основанные на разных источниках лексикографических данных: 1) перевод с помощью онлайн-
переводчиков компаний Yandex и Google, 2) перевод с помощью машиночитаемых переводных словарей Freedict и Verdict, 3) перевод с помощью параллельных подкорпусов НКРЯ. Последовательный перевод существительных на данный момент эффективнее всего осуществляется с помощью онлайн переводчиков, а искать переведенные эквиваленты для прилагательных надежнее всего с помощью параллельных корпусов достаточного объема. Вторая подзадача решается на основе анализа сочетаемости полученных прилагательных и существительных по одноязычному корпусу серии WaC.

Результаты наших пилотных экспериментов показывают, что лексико-типологические анкеты (по крайней мере, для признаковых семантических полей) могут переводиться и заполняться автоматически данными языков, снабженных необходимыми для этого электронными ресурсами. Процесс перевода с опорой на онлайн-платформы Yandex и Google автоматизирован, поэтому перевод списков слов на любой язык, инкорпорированный в эти системы, не потребует дополнительной ручной работы. Аналогично, все словари серий Freedict и Verdict имеют одинаковую структуру, поэтому любой из них может быть автоматически обработан с помощью наших программ. Полученные нами результаты показывают, что пока эти словари недостаточно представительны, но, по крайней мере, база Freedict продолжает пополняться. Перевод отдельных слов по любому параллельному подкорпусу НКРЯ также может быть осуществлен полностью автоматически без каких бы то ни было дополнительных модификаций наших алгоритмов. То же самое касается и корпусов серии WaC: все они создаются и размечаются по общей схеме, поэтому мы можем считать, что располагаем готовым инструментом обработки для каждого из них. При этом коллекция таких корпусов также пополняется, а наши эксперименты в §4 Главы 3 косвенно свидетельствуют о том, интернет-тексты, которые составляют основу таких корпусов, пригодны для лексико-типологических исследований. Однако подчеркнем, что наши алгоритмы пока не предполагают полной автоматизации процесса сбора материала для лексико-типологических исследований, поскольку на данный момент в лингвистике не накоплено достаточного количества надежных ресурсов: сопоставимых и представительных переводных словарей, объемных и сбалансированных параллельных и одноязычных корпусов. Предложенная нами методика призвана помочь типологу расширить эмпирическую базу исследований путем учета максимального количества доступных ресурсов, упростив и ускорив процесс анализа электронных источников данных.
Преимущество нашего метода заключается в том, что мы отталкиваемся от единой анкеты, которая уже в готовом виде подается на вход модулю, отвечающему за сбор данных. Это позволяет заполнять ее данными из всех доступных источников, будь то одноязычные или параллельные корпуса, вне зависимости от их структур и типов разметки. Однако, безусловно, алгоритмы извлечения данных нуждаются в адаптации под каждый новый словарь или корпус, если только он не относится к уже освоенным нами сериям. Исходя из этого, в дальнейшем мы предполагаем увеличить охват обрабатываемых словарей и корпусов для различных языков, а также разработать алгоритмы сбора типологических данных для глагольных и предметных семантических полей.
Глава 5. Построение семантической карты и анализ типов систем

Разработка семантической карты – важнейший этап типологического исследования, поскольку именно на этом этапе происходит анализ и интерпретация данных, собранных на предыдущих стадиях. Карты представляют собой модели, отражающие закономерности кодирования грамматических или лексических значений в языках мира и позволяющие наглядно сопоставлять материалы различных языков. С этим этапом тесно связан следующий, заключительный – анализ типов систем.

В лингвистической практике применяются два типа семантических карт: так называемые «классические» (графовые) и вероятностные, мы коротко обсудим особенности каждой из них в §1. Фреймовый подход к лексической типологии обычно использует только графовые модели, составляемые вручную. В этой главе мы применим к нашим лексическим данным вероятностную модель картирования, сопоставим ее с графовой и выделим основные достоинства и недостатки каждой из этих методологий с точки зрения наших задач. Этому будет посвящен §2. Затем, в §3, мы представим новую методологию семантического картирования – метод автоматического построения аналогов семантических карт с помощью решеток формальных понятий.

§1. Существующие методы создания семантических карт

1.1. Графовые семантические карты

Классическая семантическая карта представляет собой граф, в узлах которого располагаются значения (грамматические или лексические), а ребра маркируют возможность объединения двух значений в рамках одного лингвистического средства (морфемы, слова или конструкции). В большинстве случаев для построения семантических карт используется модель ненаправленного графа, т.е. отношения между значениями считаются симметричными, однако для отображения путей эволюции языковых единиц в диахронической перспективе может быть применен и направленный граф (см. van der Auwera & Plungian 1998). В недавней работе Georgakopoulos & Polis 2018 вводится также модель смешанного графа, призванная отобразить синхронное состояние и диахронические сведения на одной и той же схеме.

Обычно такие карты строятся вручную (см., однако, Regier et al. 2013) на основе имеющихся в распоряжении исследователя типологических данных. При их разработке
соблюдаются два основных принципа: принцип смежности и принцип экономии. Принцип смежности, или Semantic Map Connectivity Hypothesis (Croft 2001), подразумевает, что любая единица в любом языке выборки должна охватывать только такие значения, узлы которых формируют связный подграф на семантической карте. При этом принцип экономии гласит, что в графе должно быть минимальное количество ребер, достаточное для соблюдения принципа смежности. В большинстве случаев построенные по таким принципам карты представляют собой планарные графы, которые можно изобразить на плоскости так, чтобы никакие из их ребер не пересекались, однако возможны и более сложные схемы, см., например, Cysouw 2007. Таким образом, расстояние между узлами на такой карте незначимо и выбирается произвольно, но их взаимное расположение (прежде всего, в случае если граф планарный) оказывается информативно.

1.2. Вероятностные семантические карты

Вероятностные семантические карты, в отличие от традиционных графовых, строятся автоматически и представляют собой отображение многомерных типологических данных в пространство меньшей размерности. В лингвистической практике используется несколько техник уменьшения количества измерений пространства типологических данных, наиболее распространенная из которых - многомерное шкалирование (Multidimensional Scaling, или MDS, см. Cox & Cox 2001, а также раздел 4.5 Главы 2, где мы использовали именно эту технику для визуализации дистрибутивного и типологического векторных пространств).

Метод многомерного шкалирования подразумевает проекцию многомерного пространства в пространство с меньшим количеством измерений. Новое пространство создается таким образом, чтобы расстояние между точками максимально соответствовало расстоянию между ними в исходном многомерном пространстве. Различия в исходных и итоговых расстояниях контролируются функцией стресса, так что если на классической семантической карте расстояние между узлами не несет никакой смысловой нагрузки, то на вероятностных схемах расстояние между точками играет ключевую роль: чем оно меньше, тем больше вероятности объединения соответствующих значений одним лингвистическим средством (иными словами, тем чаще эти значения объединяются в языках выборки, на основе которой строится карта).

Предполагается, что, если лингвистические данные по той или иной семантической зоне действительно подчиняются некоторым универсальным закономерностям, то по этим
данным можно будет построить пространственную модель с небольшим числом измерений (от одного до трех) и с низким значением функции стресса. Поиск оптимальной модели с минимальным числом измерений можно считать аналогом принципа экономии, соблюдаемого при построении традиционных графовых карт.

Сфера действия конкретных языковых единиц на такой карте определяется возможностью разбиения пространства на две части. Одномерное пространство разбивается точкой, двумерное – прямой линией, трехмерное – плоскостью (ср. cutting line / plane в терминологии Croft & Poole 2008). Объединяться в рамках одного лингвистического средства могут те значения, которые можно отделить ото всех остальных одной точкой, прямой или плоскостью. Это правило можно считать аналогом принципа смежности.

§2. Графовая и вероятностная модели в применении к нашему материалу

В разделе 4.5 Главы 2, посвященном визуализациям многомерных пространств разной природы, мы приводили графовую и вероятностную карты для поля ‘острый’ и подчеркивали, что они моделируют концептуальное пространство рассматриваемого признака примерно одинаково. Здесь мы рассмотрим эти модели подробнее и покажем, что они эквивалентны во многом, но не во всем.

На Рис. 4, который мы для удобства повторяем здесь (Рис. 8), представлена традиционная графовая семантическая карта для поля ‘острый’. Эта карта показывает, что допустимы все возможные комбинации трех правых узлов, т.е. слово со значением ‘острый’ в некотором языке L может иметь все три значения, любые два или любое одно из них. При этом левый узел не связан напрямую с двумя правыми. Это означает, что, если слово X может описывать инструменты с режущим краем (ср. нож, коса) и объекты вытянутой формы (ср. нос, носок ботинка), то оно обязательно будет покрывать и фрейм инструментов с колющим концом (ср. стрела, копье). Аналогично, если слово X описывает режущие инструменты и колющие поверхности, то оно должно использоваться и для описания колющих инструментов.

35 См. Croft & Poole 2008 (с. 11): “If there are language universals in the domain being investigated, we would expect to find a spatial model with few dimensions, with a very good degree of fit to the crosslinguistic data”.

131
Рисунок 8 (=4). Семантическая карта поля ‘острый’.
- Instrument with a sharp functional edge (knife, saw) – инструмент с функциональным острым краем (нож, коса);
- Instrument with a sharp functional end-point (arrow, spear) – инструмент с функциональным острым концом (стрела, копье);
- Object with a sharp form (nose) – Объект острой формы (нос);
- Object/surface that pricks (thorn, blanket) – колючий объект/поверхность (колючка, одеяло).

Такого рода карта позволяет отобразить наиболее общие закономерности, которые можно вывести на имеющемся в распоряжении исследователя типологическом материале: какие комбинации значений допустимы в рамках одного языкового средства, а какие – нет. Сведений о системах конкретных языков такая модель не предоставляет: данные каждого языка могут быть «наложены» на карту, но визуализация будет сохранять иллюстративность только в том случае, если будет отражать данные не более одного-двух языков (ср. иллюстрацию системы лексикализации поля ‘острый’ во французском языке на Рис. 9).

Рисунок 9. Семантическая карта поля ‘острый’ с отображением данных французского языка.
Поскольку графовые семантические карты не отражают сведений о том, как данная семантическая зона оформляется лингвистическими средствами в каждом конкретном языке, они не очень удобны для анализа типов систем. Иными словами, традиционная семантическая карта не отображает, какие комбинации оказываются распространенными, а какие нет. Частично эту задачу решает модель взвешенного графа, вес ребра в котором показывает, насколько часто значения, соответствующие узлам, которые это ребро соединяет, выражаются одним и тем же языковым средством. Однако эти дополнительные сведения, будучи чрезвычайно полезными, все же не позволяют делать выводы о стратегиях оформления исследуемой зоны в разных языках. В частности, вес каждого отдельного ребра не выделяет частотных комбинаций, содержащих более двух узлов. Так, например, модель взвешенного графа, по-видимому, не позволит показать, что почти в половине языков выборки (см. нашу статью Kyuseva, Ryzhova & Parina to appear) в семантическом поле ‘острый’ значения ‘инструмент с острым краем’, ‘инструмент с острым кончиком’ и ‘объект острой формы’ выражаются одним словом, а четвертое – ‘колючий объект/поверхность’ – другим; при том, что вторая распространенная стратегия – объединение инструментов с острым кончиком и объектов вытянутой формы и их лексическое противопоставление режущим инструментам с одной стороны и колючим объектам – с другой.

Помимо того, что графовая семантическая карта не отражает частотности возможных комбинаций, эта модель обычно допускает больше возможных вариантов, чем их на самом деле засвидетельствовано. Например, семантическая карта на Рис. 8 в числе прочих допускает комбинацию ‘Instrument with a sharp functional edge’ + ‘Instrument with a sharp functional end-point’ + ‘Object/surface that pricks’, которая пока не встретилась нам ни в одном из языков выборки.

Вероятностная семантическая карта, построенная на основе предобработанных данных, в которых уже выделены минимальные лексические значения (фреймы), может выступать в качестве аналога графовой карты, поскольку расположение точек на плоскости в ней примерно соответствует взаимному расположению узлов на графовой карте (см. Рис. 10, а также карту неопределенных местоимений, построенную по материалам Haspelmath 1997 в работе Croft & Poole 2008). В этом случае такой формат представления данных скорее прорицает традиционному графовому, поскольку по такого рода визуализациям ещё сложнее определить, какие комбинации значений в рамках одной лексемы маловероятны или полностью запрещены. Так, например, согласно Рис. 10,
объединение фрейма режущих инструментов с фреймом колючих поверхностей почти столь же вероятно, что и объединение первого с фреймом колющих инструментов. При этом вторая комбинация допустима и нередко встречается в языках, в то время как первая, по-видимому, невозможна, что явно отражено на графовой семантической карте (см. Рис. 8).

Вместе с тем, неоспоримое преимущество этого метода заключается в том, что он полностью автоматический, а значит, может быть легко применен и к более обширному языковому материалу. На Рис. 11 (представленном на основе данных по всем пунктам анкеты) представлена вероятностная карта поля ‘острый’, построенная на основе данных по всем пунктам анкеты. Каждой точке соответствует фрейм: зеленый – инструменты с функциональным острым краем (нож, коса); синий – инструменты с функциональным острым концом (стрел, копье); желтый – объекты вытянутой формы (нос); красный – колючие объекты (колючка, одеяло).
включенных в выборку. Значение измерения равнялось 1 в том случае, если слово покрывает данный пункт анкеты, и 0 – если не покрывает.

Рисунок 11 (=1). Карта поля ’острый’, полученная методом многомерного шкалирования. Каждая точка соответствует пункту анкеты.

На этой карте хорошо видно, что, во-первых, контексты, иллюстрирующие один и тот же фрейм, не совпадают в одной точке, а занимают некоторую область, более или менее ограниченную, т.е. представление фрейма в виде неделимого целого – одного узла на дискретной графовой карте – это существенное упрощение реального положения дел (ср. наши рассуждения о фокальной структуре семантических полей в разделе 4.5 Главы 2). Во-вторых, такая карта показывает, что фреймы колющих инструментов и объектов вытянутой формы часто описываются одним и тем же лексическим средством: расстояние между ними совсем небольшое. Напротив, фреймы режущих инструментов и колючих поверхностей отстоят и друг от друга, и от фреймов колющих инструментов и объектов вытянутой формы. На подобные сведения можно опираться в ходе анализа типов систем.

36 Напомним, что таблицы такого же формата мы использовали для вычисления значения метрики типологической близости (см. Главу 2).
В большинстве случаев метод многомерного шкалирования применяют для визуализации больших объемов данных, изначально не структурированных, т.е. в тех случаях, когда построить графовую карту крайне затруднительно. Как правило, такими данными служат контексты параллельного корпуса с информацией о том, как они заполняются в разных языках, при этом языков может быть 100, как в работе Wälchli & Cysouw 2012, или даже 1000, как в работе Östling 2016. В таких ситуациях многомерное шкалирование используется в том числе как метод кластеризации данных и позволяет выявить основные типы контекстов употребления рассматриваемых слов. Такая визуализация не столько иллюстрирует выявленные закономерности, сколько служит дополнительным инструментом для их обнаружения и сама по себе нуждается в интерпретации.

Таким образом, два рассмотренных здесь метода построения семантических карт работают на основе сходных принципов и дают в целом сопоставимые результаты. Основное различие в их функционалах заключается в том, что вероятностные карты лучше приспособлены к большим объемам данных и лучше отображают степени близости между двумя значениями в концептуальном пространстве. В то же время, из традиционной графовой модели легче выводятся ограничения: по ней сразу видно, какие комбинации заведомо недопустимы – далеко не все такие комбинации так же очевидно запрещаются вероятностной моделью. При этом ни графовая, ни вероятностная модели не приспособлены к единовременному отображению данных сразу нескольких языков. Впрочем, расстояния между объектами на вероятностной карте позволяют примерно оценить, насколько часто те или иные значения объединяются в рамках одной языковой единицы.

В следующем параграфе этой главы мы предложим новый метод семантического картирования, отличающийся и от графового, и от вероятностного по своему функционалу. Он основан на математическом аппарате анализа формальных понятий, который мы кратко представим в разделе 1. Далее, в разделе 2, мы покажем, как этот аппарат может быть использован в решении задачи построения семантической карты и чем такая карта будет отличаться от графовой и вероятностной моделей. В разделе мы обсудим возможность применения модели решетки формальных понятий к задаче анализа метафорических значений.
§3. Построение семантических карт с помощью решеток формальных понятий

1. Анализ формальных понятий

Анализ формальных понятий (formal concept analysis, или FCA, см. Ganter & Wille 1999) – математический аппарат, позволяющий анализировать структуру данных, представленных в виде множества объектов, каждый из которых характеризуется определенным набором атрибутов. Множество объектов представляется в виде иерархии формальных понятий (concepts). У каждого формального понятия есть объем (множество относящихся к нему объектов) и содержание (множество атрибутов, необходимых и достаточных для каждого объекта, входящего в данное формальное понятие). Объем понятий служит критерием для их упорядочивания.

Ниже мы приведем несколько определений, необходимых для понимания предлагаемого математического аппарата.

Формальный контекст $K$ – это тройка $(G, M, I)$, где $G$ – это множество объектов, $M$ – множество атрибутов, а $I$ – бинарное отношение $I \subseteq G \times M$, устанавливаемое между объектом и атрибутом в том случае, если данный объект обладает данным атрибутом. Такое отношение называют отношением инцидентности. Для произвольных $A \subseteq G$ и $B \subseteq M$ определены операторы Галуа:

$A' = \{m \in M \mid \forall g \in A : glm\}$,

$B' = \{g \in G \mid \forall m \in B : glm\}$,


Формальным понятием контекста $(G, M, I)$ называются такие пары $(A, B)$, где $A \subseteq G$, $B \subseteq M$, $A = B'$, и $B = A'$. Множество $A$ называется объемом, а $B$ – содержанием понятия $(A, B)$.

Понятие $(A, B)$ считается менее общим, чем понятие $(C, D)$, если $A \subseteq C$. Все множество понятий, структурированное по определенному таким образом отношению быть более или менее общим, формирует решетку, которая называется решеткой формальных понятий для контекста $K$. Обычно решетка формальных понятий визуализируется в виде линейной диаграммы, где узлы соответствуют понятиям и более общие понятия размещаются над менее общими. Два понятия связываются между собой линией, если один из них менее общий, чем другой, и между ними нет других формальных понятий. Объекты,

37 Основные результаты работы в этом направлении отражены в нашей статье Ryzhova & Obiedkov 2017.
соотносящиеся с данным узлом, принято записывать снизу от узла, а атрибуты – сверху от него. Объем понятия определяется по решетке следующим образом: в него входят все объекты, которые приписываются соответствующему узлу, а также всем узлам, находящимся ниже данного, до которых можно дойти, двигаясь из исходного узла по ребрам вниз. Содержание понятия состоит из атрибутов, приписанных соответствующему узлу, а также всем тем узлам, которые находятся выше данного и до которых можно добраться, двигаясь из исходного узла по ребрам вверх.

В следующем разделе мы покажем, что, если объектами считать слова, а атрибутами – фреймы, которые эти слова могут покрывать, то решетку формальных понятий для такого формального контекста можно рассматривать как особый вид семантической карты для данной лексической области.

2. Решетки формальных понятий как лексические семантические карты

Решетка формальных понятий может быть сконструирована на основе того же набора данных, который необходим для создания традиционной графовой или вероятностной семантической карты. Для ее создания достаточно располагать сведениями о том, какие минимальные значения (в нашей терминологии – фреймы) должны быть отражены на карте, и как эти значения покрываются лингвистическими средствами в одном или нескольких языках. Данные представляются в виде таблицы того же формата, что используется для построения вероятностной карты (или для вычисления значения метрики типологической близости): в строках указываются значения (атрибуты), в столбцах – слова разных языков (объекты); на пересечении строки и столбца ставится «1», если слово покрывает данное значение, «0» – если не покрывает. На основе такой матрицы решетка строится автоматически. Для наших экспериментов мы использовали находящуюся в открытом доступе программу Concept Explorer (URL: http://conexp.sourceforge.net/).

Никаких дополнительных сведений для построения такого рода диаграмм не требуется. Это означает, что технически любые типологические данные могут быть представлены в виде решетки формальных понятий. Важно, однако, что не во всех случаях она будет легко интерпретируемой.

Наиболее иллюстративными являются решетки без пересечений, т.е. такие диаграммы, в которых никакие два ребра не пересекаются друг с другом. Такая конфигурация возможна только в том случае, если минимальные значения (фреймы)
организованы линейно, т.е. если на классической графовой карте их можно представить в виде линии. На Рис. 12 представлена решетка формальных понятий, построенная для трех фреймов поля ‘острый’, которые на традиционной семантической карте расположены линейно (см. Рис. 8): ‘инструмент с острым функциональным краем’, ‘инструмент с острым функциональным кончиком’ и ‘объект острой формы’.

Рисунок 12. Решетка формальных понятий для линейного фрагмента поля ‘острый’.

Каждый узел в такой решетке соответствует комбинации фреймов (т.е. включает все фреймы, до которых можно дойти, двигаясь по ребрам вверх) и набору слов, покрывающих эти фреймы (всех, до которых можно дойти, двигаясь по ребрам вниз). В терминах анализа формальных понятий, каждый узел – это понятие, соответствующий набор фреймов – его содержание, а набор слов – его объем. Например, узел, к которому приписан набор слов pance, pointu, spitz, hegyes и jiān, обозначает комбинацию из двух фреймов: ‘инструмент с острым функциональным кончиком’ и ‘объект острой формы’. В его объем входят все слова, покрывающие оба этих фрейма. Лексемы, приписанные непосредственно к этому узлу, объединяют ровно эти два значения, но к объему этого формального понятия относятся в том числе те слова, которые покрывают все три фрейма рассматриваемого поля (такие лексемы приписаны к самому нижнему узлу). Синим выделены узлы, к которым непосредственно приписан хотя бы один атрибут. Размер узла зависит от количества приписанных к нему слов.
Конфигурация синих узлов в решетке полностью соответствует взаимному расположению фреймов на графовой карте: фрейм ‘инструмент с острым функциональным кончиком’ находится между двумя другими. Остальные узлы иерархии вкупе с подписями к узлам передают именно ту информацию, которая отсутствует на графовых и вероятностных семантических картах: узлы показывают, какие комбинации фреймов заведомо допустимы (т. е. засвидетельствованные в том или ином языке выборки), а по подписям можно вывести стратегию лексикализации поля в каждом конкретном языке. В решетке на Рис. 12 к нижнему узлу приписаны слова, у которых есть все три возможных значения, на среднем уровне располагаются узлы, соответствующие возможным комбинациям из двух фреймов, а лексемы, приписанные к узлам верхнего уровня, специализируются только на одном значении каждая.

Кроме того, размер узлов сигнализирует, какие стратегии являются более распространенными в рамках заданного типологического материала. Так, например, в решетке на Рис. 12 есть три узла, размер которых существенно превышает размеры остальных: самый нижний узел, правый в среднем ряду и самый левый в верхнем ряду. Это означает, что на наших данных можно выделить две основные стратегии лексикализации зоны ‘острый’: доминантная, при которой все три фрейма покрываются одним словом (такие слова приписываются самому нижнему узлу), и бинарная, при которой одно слово описывает инструменты с режущим краем (верхний левый узел) и ещё одно объединяет фреймы ‘инструмент с острым функциональным кончиком’ и ‘объект острой формы’ (правый узел в среднем ряду).

Более сложная организация концептуального пространства приводит к пересечениям ребер в решетке и, как следствие, к снижению уровня иллюстративности диаграммы. На Рис. 13 изображена решетка формальных понятий, построенная на полном наборе данных для поля ‘острый’, в который входит дополнительный фрейм колючих поверхностей/объектов, нарушающий линейность семантической структуры. По такой решетке по-прежнему можно определить, какие комбинации фреймов допустимы, а какие нет. В частности, по этой диагrame видно, что комбинация из фреймов ‘инструмент с острым функциональным краем’, ‘инструмент с острым функциональным кончиком’ и ‘колючий объект/поверхность’, допускаемая графовой картой, не засвидетельствована ни в одном языке. Однако взаимное расположение фреймов в концептуальном пространстве, иллюстрирующее степень близости значений, из такой схемы вывести уже затруднительно.
При увеличении количества фреймов в поле и числа допустимых стратегий диаграмма усложняется, вплоть до полной потери иллюстративности.

Рисунок 13. Решетка формальных понятий для полного набора фреймов поля ‘острый’.

Таким образом, решетки формальных понятий отличаются по своему функционалу от графовых и вероятностных семантических карт. В большинстве случаев не самым явным образом визуализируя степень близости значений, решетки позволяют определить, каковы допустимые комбинации узлов карты и какие из возможных стратегий встречаются чаще других в языках выборки. Кроме того, предлагаемый нами формализм может использоваться в качестве удобного инструмента для анализа метафорических значений слов.

3. Представление метафорических значений

На графовых и вероятностных семантических картах пункты, соответствующие минимальным лексическим и грамматическим значениям, представляются независимыми друг от друга. Исключением из этого правила являются так называемые иерархии: линейные карты, иллюстрирующие ситуацию, когда выражение одних значений неизбежно влечет за собой возможность покрывать другие, находящиеся правее или левее в иерархии (ср. широко известную иерархию доступности Э. Кинэна – Б. Комри, Keenan & Comrie 1977). В том случае, если пункты на карте действительно независимы, все синие узлы в решетке формальных понятий располагаются на самом верхнем уровне диаграммы. Однако теоретически аппарат формальных понятий позволяет визуализировать и прямую зависимость одного фрейма от другого, например: если у слова W есть значение X1, то у
него обязательно есть значение X2 (при этом обратное неверно). В таких случаях узел, соответствующий значению X1, располагается в решетке на уровень ниже, чем узел X2. Это свойство решеток формальных понятий делает их очень удобным инструментом для анализа метафорических значений слов.

Опыт работы Московской лексико-типологической группы (MLexT) показывает, что метафорические значения слов с близкой исходной семантикой повторяются от языка к языку, причем одни и те же метафоры могут встретиться как в родственных языках, так и в языках, далеких друг от друга ареально и генетически (Кашкин 2013; Кюсева 2012; Майсак & Рахилина 2007; Резникова и др. 2015). Например, одна из самых распространенных метафор, которую могут развивать лексемы из зоны ‘острый’ – обозначение хорошего функционирования системы органов зрения (ср. русск. острое зрение). Помимо русского, эта метафора встречается в китайском, японском, кабардинском, венгерском, коми-зырянском, малайском и ряде других языков.

В работах Рахилина & Резникова 2013, Rakhilina & Reznikova 2016 формулируется также и более сильное утверждение, согласно которому метафорические значения характерны не для поля (или лексемы) в целом, а ассоциированы с тем или иным фреймом (минимальным исходным значением) или семантическим параметром. Аппарат анализа формальных понятий позволяет автоматически выявить и наглядно отобразить в структуре решетки, с какими именно фреймами, согласно материалам языков выборки, связаны наблюдаемые переносные значения. Так, решетка на Рис. 14 показывает, что значение ‘резкий, четкий (о линии, границе, контрасте)’ формируется только у слов, исходно описывающих острые режущие инструменты: у всех объектов, у которых есть атрибут ‘sharp line / sharp contrast’, обязательно есть и атрибут ‘instrument with a sharp functional edge’. 
Отметим, что существующие методы семантического картирования, насколько нам известно, не позволяют корректно отображать данные о моделях семантических переходов, характерных для изучаемого поля. Все точки на вероятностной карте равноправны, и нет возможности отличить прямые значения от переносных и, тем более, указать, с какими именно исходными значениями связаны метафорические употребления. Графовая модель, напротив, позволяет выделить графически значения разной природы и отметить стрелками направления семантических сдвигов, но и здесь есть два существенных затруднения. Во-первых, такого рода графическая информация загромождает карту (особенно если требуется отличить метафорический сдвиг от диахронического расширения на уровне прямых значений). Во-вторых, данных не всегда достаточно для того, чтобы легко определить, с каким именно исходным значением связано данное переносное, а даже если данных достаточно, их сложно проанализировать вручную. Наконец, в-третьих, будучи значениями другой природы и другого уровня, метафоры часто требуют включения в карту третьего измерения для соблюдения принципа смежности.

Чтобы сохранить наглядность семантической карты, в рамках фреймового подхода к лексической типологии на семантические карты помещаются только прямые значения. В работе François 2008 на карте отражены и прямые, и переносные значения, но разница в их семантической природе никак не маркирована, из-за чего предложенная карта становится очень неоднородной и, как следствие, менее показательной.

Кроме того, традиционная графовая модель только визуализирует уже установленную зависимость. Решетки формальных понятий, напротив, могут служить инструментом для выявления такого рода взаимосвязей.
§4. Выводы

Аппарат анализа формальных понятий открывает новые перспективы в области семантического картирования. Решетка формальных понятий может рассматриваться как аналог графовой или вероятностной семантической карты, однако ее функционал существенно отличается от возможностей уже известных типов визуализации типологических закономерностей. Решетка строится автоматически на основе сведений о том, какие значения какими лингвистическими средствами могут обслуживаться. Для каждого набора данных возможна только одна конфигурация решетки, в которой отображаются все засвидетельствованные комбинации значений, а также фиксируются зависимости между атрибутами (виде «если у объекта есть атрибут X1, то у него обязательно есть и атрибут X2»).

В том случае, если концептуальное пространство организовано линейно, соответствие между решеткой и традиционной графовой картой тривиально. В такой ситуации решетка отличается от графа тем, что позволяет эксплицитно и компактно выразить, какие стратегии оформления зоны языковыми средствами больше распространены в языках выборки, т.е. способствует решению задачи анализа типологии систем. Пространствам более сложной организации соответствуют и более сложные решетки: по мере увеличения количества минимальных значений и усложнения взаимоотношений между ними решетка теряет свою иллюстративность.

Таким образом, решетки формальных понятий могут использоваться в типологии вообще и в лексической типологии в частности в качестве дополнительного инструмента исследования, позволяющего автоматически выявлять и наглядно визуализировать такие закономерности, которые не отображаются ни на одном из известных нам типов семантических карт. В частности, анализ формальных понятий может стать удобным инструментом исследования типологии метафор.
Заключение

В ходе диссертационного исследования получены следующие основные результаты:

(1) Впервые применены квантитативные методики сбора и анализа данных в рамках фреймового подхода к лексической типологии;

(2) Предложено дополнительное, квантитативное обоснование лингвистической значимости лексико-типологического понятия «фрейм»;

(3) Показано, что фреймовая структура поля представляет собой пересекающиеся кластеры с ярко выраженными центрами, т.е. ситуациями-прототипами;

(4) Подтверждено, что первичная фреймовая анкета как исходный пункт для любых самых широких лексико-типологических исследований может быть сформирована на лексических данных одного языка;

(5) Формализована процедура лексико-типологического исследования, выполняемого в рамках фреймовой парадигмы: выделены основные его этапы и сформулированы задачи, которые должны быть решены на каждом шаге;

(6) Предложены алгоритмы автоматизации всех этапов исследования, которые позволяют:
   а) построить предварительный вариант типологической анкеты;
   б) перевести анкету на другие языки и заполнить её, опираясь на материал доступных словарей и корпусов;
   в) построить семантическую карту нового формата, отображающую взаимоотношения между фреймами не только прямых, но и метафорических значений;
   г) получить данные о типичных и периферийных для рассматриваемого лексико-семантического поля типах систем;

(7) Проведен подробный анализ полученных результатов и определена перспектива применения квантитативных методик в лексической типологии.

Таким образом, цели и задачи, поставленные в нашей работе, выполнены: мы показали, что новые методы открывают новые возможности для типологического анализа лексики, позволяя существенно ускорить исследовательский процесс и включить в него более широкий языковой материал и выявить новые закономерности организации лексических значений.
По теме диссертации опубликованы следующие работы:

1. Опыт автоматического построения анкеты для лексико-типологического исследования прилагательных и одноместных глаголов с помощью моделей дистрибутивной семантики // Вестник РГГУ, М.: 2016. № 9 (18). С. 140-150.


Библиография

1. Абдурашитова Г.М. Кластеризация существительных на основе их сочетаемости в разных языках: дипломная работа. Москва, 2017.


7. Брицын В.М. [и др.]. Концепт БОЛЬ в типологическом освещении / под ред. В.М. Брицын, Е.В. Рахилина, Т.И. Резникова, Г.М. Яворская, Киев: Видавничий Дім Дмитра Бураго, 2009.


24. Кюсева М.В., Резникова Т.И., Рыжова Д.А. Совершенствование одноязычных, двуязычных и мультиязычных словарей: автоматизация процесса сбора материала //


34. Минский М. Фреймы для представления знаний / М. Минский, Москва: Энергия, 1979.

35. Наний Л.О. Прилагательные простейших форм и размеров китайского и русского языков в типологическом аспекте: дис. … канд. филол. наук. Москва, 2016.


47. Рыжова Д.А. Опыт автоматического построения анкеты для лексико-типологического исследования прилагательных и одноместных глаголов с помощью моделей дистрибутивной семантики // Вестник РГГУ. Серия «История. Филология. Культурология. Востоковедение». 2016. № 9 (18). С. 140–150.


55. Шеванова М.В. Компьютерные методы установления границ поля в лексико-типологическом исследовании: курсовая работа. Москва, 2016.


Приложение 1. Анкета для поля ‘острый’

<table>
<thead>
<tr>
<th>Фрейм</th>
<th>Микрофрейм</th>
<th>Словосочетание для дистрибутивного представления (русскоязычный корпус)</th>
<th>Словосочетание для дистрибутивного представления (англоязычный корпус)</th>
</tr>
</thead>
<tbody>
<tr>
<td>прямые значения</td>
<td>острый нож</td>
<td>острый_нож</td>
<td>sharp_knife</td>
</tr>
<tr>
<td></td>
<td>острый меч</td>
<td>острый_меч</td>
<td>sharp_sword</td>
</tr>
<tr>
<td></td>
<td>острая сабля</td>
<td>острый_сабля</td>
<td>sharp_saber</td>
</tr>
<tr>
<td></td>
<td>остroe лезвие</td>
<td>острый_лезвие</td>
<td>sharp_blade</td>
</tr>
<tr>
<td></td>
<td>острыя коса</td>
<td>острый_коса</td>
<td>sharp_scythe</td>
</tr>
<tr>
<td></td>
<td>острые ножницы</td>
<td>острый_ножницы</td>
<td>sharp_scissors</td>
</tr>
<tr>
<td></td>
<td>острыя пила</td>
<td>острый_пила</td>
<td>sharp_saw</td>
</tr>
<tr>
<td>острый инструмент с режущим краем</td>
<td>острыя иголка</td>
<td>острый_иголка</td>
<td>sharp_needle</td>
</tr>
<tr>
<td></td>
<td>острыя стрела</td>
<td>острый_стрел</td>
<td>sharp_arrow</td>
</tr>
<tr>
<td></td>
<td>острый коготь</td>
<td>острый_коготь</td>
<td>sharp_claw</td>
</tr>
<tr>
<td></td>
<td>острый кол</td>
<td>острый_кол</td>
<td>sharp_pole</td>
</tr>
<tr>
<td></td>
<td>острыя спица</td>
<td>острый_спица</td>
<td>sharp_knitting_needle</td>
</tr>
<tr>
<td></td>
<td>острыя палка</td>
<td>острый_палка</td>
<td>sharp_stick</td>
</tr>
<tr>
<td></td>
<td>острые копье</td>
<td>острый_копье</td>
<td>sharp_spear</td>
</tr>
<tr>
<td></td>
<td>острый зуб</td>
<td>острый_зуб</td>
<td>sharp_tooth</td>
</tr>
<tr>
<td></td>
<td>острый ноготь</td>
<td>острый_ноготь</td>
<td>sharp_nail</td>
</tr>
<tr>
<td></td>
<td>острый рог</td>
<td>острый_рог</td>
<td>sharp_horn</td>
</tr>
<tr>
<td>предмет, суживающийся к концу</td>
<td>острый нос</td>
<td>острый_нос</td>
<td>sharp_nose</td>
</tr>
<tr>
<td></td>
<td>острый носок</td>
<td>острый_носок</td>
<td>sharp_shoe_tip</td>
</tr>
<tr>
<td></td>
<td>ботинка</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>острыя гора</td>
<td>острый_гора</td>
<td>sharp_peak</td>
</tr>
<tr>
<td></td>
<td>острый локоть</td>
<td>острый_локоть</td>
<td>sharp_elbow</td>
</tr>
<tr>
<td></td>
<td>острый клюв</td>
<td>острый_клюв</td>
<td>sharp_beak</td>
</tr>
<tr>
<td></td>
<td>острый колпак</td>
<td>острый_колпак</td>
<td>pointed_cap</td>
</tr>
<tr>
<td></td>
<td>острый подбородок</td>
<td>острый_подбородок</td>
<td>sharp_chin</td>
</tr>
<tr>
<td></td>
<td>острый каблук</td>
<td>острый_каблук</td>
<td>pointed_heel</td>
</tr>
<tr>
<td></td>
<td>острыя мачта</td>
<td>острый_мачта</td>
<td>pointed_mast</td>
</tr>
<tr>
<td>колючаяся поверхность / колючий предмет</td>
<td>острый шип</td>
<td>острый_шип</td>
<td>sharp_thorn</td>
</tr>
<tr>
<td></td>
<td>колючее одеяло</td>
<td>колючий_одеяло</td>
<td>prickly_blanket</td>
</tr>
<tr>
<td></td>
<td>колючая щетина</td>
<td>колючий_щетина</td>
<td>prickly_bristle</td>
</tr>
<tr>
<td>колючая шерсть</td>
<td>колючий шерсть</td>
<td>prickly_wool</td>
<td></td>
</tr>
<tr>
<td>----------------</td>
<td>---------------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>колючий куст</td>
<td>колючий куст</td>
<td>thorny_bush</td>
<td></td>
</tr>
<tr>
<td>колючий свитер</td>
<td>колючий свитер</td>
<td>prickly_sweater</td>
<td></td>
</tr>
</tbody>
</table>

переносные значения

gеометрическая форма

<table>
<thead>
<tr>
<th>острый угол</th>
<th>острый угол</th>
<th>sharp_angle</th>
</tr>
</thead>
<tbody>
<tr>
<td>кругой поворот</td>
<td>кругой поворот</td>
<td>sharp_turn</td>
</tr>
<tr>
<td>кругой изгиб</td>
<td>кругой изгиб</td>
<td>sharp_bend</td>
</tr>
<tr>
<td>кругая излучина реки</td>
<td>кругая излучина реки</td>
<td>sharp_bend</td>
</tr>
</tbody>
</table>

отвесный, обрывистый, кругой склон

<table>
<thead>
<tr>
<th>кругой подъем</th>
<th>кругой подъем</th>
<th>sharp_rise</th>
</tr>
</thead>
<tbody>
<tr>
<td>кругой склон</td>
<td>кругой склон</td>
<td>steep_slope</td>
</tr>
<tr>
<td>кругой берег</td>
<td>кругой берег</td>
<td>steep_bank</td>
</tr>
<tr>
<td>кругая гора</td>
<td>кругая гора</td>
<td>steep_mountain</td>
</tr>
</tbody>
</table>

четкая линия/изображение с четкими линиями

<table>
<thead>
<tr>
<th>четкая линия</th>
<th>четкий линия</th>
<th>sharp_line</th>
</tr>
</thead>
<tbody>
<tr>
<td>четкая фотография</td>
<td>четкий_фотография</td>
<td>sharp_picture</td>
</tr>
<tr>
<td>резкий контраст</td>
<td>резкий_контраст</td>
<td>sharp_contrast</td>
</tr>
<tr>
<td>четкая картинка</td>
<td>четкий_картинка</td>
<td>sharp_image</td>
</tr>
<tr>
<td>сильные очки</td>
<td>сильный_очки</td>
<td>strong_glasses</td>
</tr>
</tbody>
</table>

умный, хорошо соображающий (о человеке)

<table>
<thead>
<tr>
<th>проницательный человек</th>
<th>проницательный_человек</th>
<th>sharp_man</th>
</tr>
</thead>
<tbody>
<tr>
<td>острый ум</td>
<td>острый_ум</td>
<td>sharp_wit,acute_mind,keen_mind</td>
</tr>
<tr>
<td>острый, проницательный взгляд</td>
<td>проницательный_взгляд</td>
<td>sharp_eye</td>
</tr>
<tr>
<td>проницательный наблюдатель</td>
<td>проницательный_наблюдатель</td>
<td>keen_observer</td>
</tr>
<tr>
<td>умная девочка</td>
<td>умный_девочка</td>
<td>sharp_girl</td>
</tr>
<tr>
<td>умная голова</td>
<td>умный_голова</td>
<td>clever_head</td>
</tr>
</tbody>
</table>

остроумный, язвительный, меткий человек

<table>
<thead>
<tr>
<th>острый журналист</th>
<th>острый_журналист</th>
<th>sarcastic_journalist</th>
</tr>
</thead>
<tbody>
<tr>
<td>острое слово</td>
<td>острый_слово</td>
<td>poignant_word</td>
</tr>
<tr>
<td>острый язык</td>
<td>острый_язык</td>
<td>sharp_tongue</td>
</tr>
<tr>
<td>острое высказывание</td>
<td>острый_высказывание</td>
<td>sharp_utterance</td>
</tr>
<tr>
<td>острый критик</td>
<td>острый_критик</td>
<td>sharp_critic</td>
</tr>
<tr>
<td>острое замечание</td>
<td>острый_замечание</td>
<td>sharp_remark</td>
</tr>
<tr>
<td>острая критика</td>
<td>острая_критика</td>
<td>sharp_criticism</td>
</tr>
<tr>
<td>Острая статья</td>
<td>Острый статья</td>
<td>Harsh article</td>
</tr>
<tr>
<td>---------------</td>
<td>---------------</td>
<td>--------------</td>
</tr>
<tr>
<td>Острый анекдот</td>
<td>Острый анекдот</td>
<td>Harsh anecdote</td>
</tr>
<tr>
<td>Грубый, жестокий, резкий (о человеке)</td>
<td>Резкий человек</td>
<td>Abrupt person</td>
</tr>
<tr>
<td>Резкий взгляд</td>
<td>Резкий взгляд</td>
<td>Abrupt look</td>
</tr>
<tr>
<td>Рекное слово</td>
<td>Резкое слово</td>
<td>Harsh word</td>
</tr>
<tr>
<td>Резкая реакция</td>
<td>Резкая реакция</td>
<td>Harsh reaction</td>
</tr>
<tr>
<td>Резкая брань</td>
<td>Резкий брань</td>
<td>Harsh verbal abuse</td>
</tr>
<tr>
<td>Грубы ответ</td>
<td>Грубый ответ</td>
<td>Harsh answer</td>
</tr>
<tr>
<td>Грубая бесцеремонность</td>
<td>Грубый бесцеремонность</td>
<td>Sharp arrogance</td>
</tr>
<tr>
<td>Грубый цинизм</td>
<td>Грубый цинизм</td>
<td>Sharp cynicism</td>
</tr>
<tr>
<td>Строгий (о человеке)</td>
<td>Строгий закон</td>
<td>Stringent law</td>
</tr>
<tr>
<td>Строгий правитель</td>
<td>Строгий правитель</td>
<td>Strict ruler</td>
</tr>
<tr>
<td>Строгий распорядок</td>
<td>Строгий распорядок</td>
<td>Strict schedule</td>
</tr>
<tr>
<td>Строгое наказание</td>
<td>Строгий наказание</td>
<td>Severe punishment</td>
</tr>
<tr>
<td>Строгая диета</td>
<td>Строгий диета</td>
<td>Strict diet</td>
</tr>
<tr>
<td>Непоседливый (о человеке)</td>
<td>Непоседливый ребенок</td>
<td>Restless child</td>
</tr>
<tr>
<td>Узкоспециализированное исследование</td>
<td>Узкоспециализированное исследование</td>
<td>Specialized research</td>
</tr>
<tr>
<td>Узкоспециализированная работа</td>
<td>Узкоспециализированная работа</td>
<td>Specialized work</td>
</tr>
<tr>
<td>Блюдо с большим количеством специй и пряностей</td>
<td>Острый соус</td>
<td>Hot sauce</td>
</tr>
<tr>
<td>Острый вкус</td>
<td>Острый вкус</td>
<td>Spicy taste</td>
</tr>
<tr>
<td>Острый перец</td>
<td>Острый перец</td>
<td>Hot pepper</td>
</tr>
<tr>
<td>Газированный напиток</td>
<td>Газированный напиток</td>
<td>Sparkling water</td>
</tr>
<tr>
<td>Газированная вода</td>
<td>Газированная вода</td>
<td>Sparkling drink</td>
</tr>
<tr>
<td>Неприятно действующий на органы чувств: звук</td>
<td>Резкий звук</td>
<td>Jarring sound</td>
</tr>
<tr>
<td>Острый визг</td>
<td>Острый визг</td>
<td>Jarring scream</td>
</tr>
<tr>
<td>Резкий голос</td>
<td>Резкий голос</td>
<td>Shrill voice</td>
</tr>
<tr>
<td>Резкий лай</td>
<td>Резкий лай</td>
<td>Jarring bark</td>
</tr>
<tr>
<td>Резкий храп</td>
<td>Резкий храп</td>
<td>Jarring snore</td>
</tr>
<tr>
<td>Неприятно действующий на органы чувств: вкус</td>
<td>Резкий вкус</td>
<td>Sharp taste</td>
</tr>
<tr>
<td>неприятно действующий на органы чувств: свет</td>
<td>резкий свет</td>
<td>резкий_свет</td>
</tr>
<tr>
<td>---------------------------------------------</td>
<td>-------------</td>
<td>-------------</td>
</tr>
<tr>
<td>реальная вспышка</td>
<td>реальный_вспышка</td>
<td>sharp_flash</td>
</tr>
<tr>
<td>острый блеск</td>
<td>острый_блеск</td>
<td>sharp_glint</td>
</tr>
<tr>
<td>неприятно действующий на органы чувств: запах</td>
<td>резкий запах</td>
<td>резкий_запах</td>
</tr>
<tr>
<td>высокий звук</td>
<td>высокий звук</td>
<td>высокий_звук</td>
</tr>
<tr>
<td>высокая нота</td>
<td>высокий_нота</td>
<td>high_note</td>
</tr>
<tr>
<td>яркий цвет</td>
<td>цвет</td>
<td>яркий_цвет</td>
</tr>
<tr>
<td></td>
<td>синева</td>
<td>яркий_синева</td>
</tr>
<tr>
<td></td>
<td>оттенок</td>
<td>яркий_оттенок</td>
</tr>
<tr>
<td>сильный запах</td>
<td>сильный запах</td>
<td>сильный_запах</td>
</tr>
<tr>
<td></td>
<td>сильный аромат</td>
<td>сильный_аромат</td>
</tr>
<tr>
<td>яркий свет</td>
<td>яркий свет</td>
<td>яркий_свет</td>
</tr>
<tr>
<td></td>
<td>яркое солнце</td>
<td>яркий_солнце</td>
</tr>
<tr>
<td>хорошо функционирующий орган чувств</td>
<td>острый слух</td>
<td>острый_слух</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>острое зрение</td>
<td>острый_зрение</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>хорошие глаза</td>
<td>хороший_глаз</td>
</tr>
<tr>
<td></td>
<td>хорошие уши</td>
<td>хороший_ухо</td>
</tr>
<tr>
<td></td>
<td>острый нюх</td>
<td>острый_нюх</td>
</tr>
<tr>
<td>отчетливое ощущение</td>
<td>острое ощущение</td>
<td>острый_ощущение</td>
</tr>
<tr>
<td></td>
<td>острое осознание</td>
<td>острый_осознание</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>сильная эмоция</td>
<td>острое желание</td>
<td>острый_желание</td>
</tr>
<tr>
<td></td>
<td>острая ревность</td>
<td>острый_ревность</td>
</tr>
<tr>
<td></td>
<td>острая зависть</td>
<td>острый_зависть</td>
</tr>
<tr>
<td></td>
<td>острая обида</td>
<td>острый_обида</td>
</tr>
<tr>
<td></td>
<td>острое стремление</td>
<td>острый_стремление</td>
</tr>
<tr>
<td>погодное явление, проявляющееся в высокой степени</td>
<td>острый мороз</td>
<td>острый_мороз</td>
</tr>
<tr>
<td></td>
<td>резкий ветер</td>
<td>резкий_ветер</td>
</tr>
<tr>
<td></td>
<td>сильный холод</td>
<td>сильный_холод</td>
</tr>
<tr>
<td></td>
<td>сильная жара</td>
<td>сильный_жара</td>
</tr>
<tr>
<td>требующий немедленного разрешения</td>
<td>острая проблема</td>
<td>острый_проблема</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>острая нехватка</td>
<td>острый_нехватка</td>
</tr>
<tr>
<td>Острый вопрос</td>
<td>Острый вопрос</td>
<td>Sharp question</td>
</tr>
<tr>
<td>----------------</td>
<td>----------------</td>
<td>----------------</td>
</tr>
<tr>
<td>Острый вопрос</td>
<td>Острый вопрос</td>
<td>Sharp question</td>
</tr>
<tr>
<td>Интенсивное физиологическое ощущение</td>
<td>Острые боль</td>
<td>Sharp pain</td>
</tr>
<tr>
<td>Острые жажда</td>
<td>Excessive thirst</td>
<td></td>
</tr>
<tr>
<td>Острый недосып</td>
<td>Acute sleep deprivation</td>
<td></td>
</tr>
<tr>
<td>Острая дрожь</td>
<td>Severe tremor</td>
<td></td>
</tr>
<tr>
<td>Острая слабость</td>
<td>Severe weakness</td>
<td></td>
</tr>
<tr>
<td>Сильная болезнь</td>
<td>Острые воспаление</td>
<td>Acute inflammation</td>
</tr>
<tr>
<td>Острый грипп</td>
<td>Severe flu</td>
<td></td>
</tr>
<tr>
<td>Острое заболевание</td>
<td>Acute disease</td>
<td></td>
</tr>
<tr>
<td>Острое расстройство</td>
<td>Acute disorder</td>
<td></td>
</tr>
<tr>
<td>Острый аппендицит</td>
<td>Acute appendicitis</td>
<td></td>
</tr>
<tr>
<td>Напряженный (о взаимодействиях)</td>
<td>Острый конфликт</td>
<td>Acute conflict</td>
</tr>
<tr>
<td>Острый момент</td>
<td>Tense moment</td>
<td></td>
</tr>
<tr>
<td>Острое сопротивление</td>
<td>Sharp opposition</td>
<td></td>
</tr>
<tr>
<td>Острый спор</td>
<td>Sharp debate</td>
<td></td>
</tr>
<tr>
<td>Острая перепалка</td>
<td>Sharp squabble</td>
<td></td>
</tr>
<tr>
<td>Интенсивный по степени проявления (о качествах человека)</td>
<td>Острое любопытство</td>
<td>Acute curiosity</td>
</tr>
<tr>
<td>Острая наблюдательность</td>
<td>Acute outsight</td>
<td></td>
</tr>
<tr>
<td>Острая впечатлительность</td>
<td>Acute susceptibility</td>
<td></td>
</tr>
<tr>
<td>Острая гордость</td>
<td>Acute pride</td>
<td></td>
</tr>
<tr>
<td>Резкое изменение</td>
<td>Резкий подъем</td>
<td>Sharp raise</td>
</tr>
<tr>
<td>Резкий спад</td>
<td>Sharp decline</td>
<td></td>
</tr>
<tr>
<td>Внезапное, быстрое движение</td>
<td>Резкий взмах</td>
<td>Abrupt stroke</td>
</tr>
<tr>
<td>Резкий вздох</td>
<td>Abrupt sigh</td>
<td></td>
</tr>
<tr>
<td>Резкое движение</td>
<td>Abrupt movement</td>
<td></td>
</tr>
<tr>
<td>Резкий бросок</td>
<td>Abrupt throw</td>
<td></td>
</tr>
<tr>
<td>Резкий скачок</td>
<td>Abrupt shift</td>
<td></td>
</tr>
<tr>
<td>Быстрый (о скорости)</td>
<td>Быстрый темп</td>
<td>Quick pace</td>
</tr>
<tr>
<td>Быстрый бег</td>
<td>Fast run</td>
<td></td>
</tr>
<tr>
<td>Фрейм</td>
<td>Микрофрейм</td>
<td>Словосочетание для дистрибутивного представления</td>
</tr>
<tr>
<td>-------</td>
<td>------------</td>
<td>--------------------------------------------------</td>
</tr>
<tr>
<td>прямые значения</td>
<td>Скользкая опорная поверхность, покрытая льдом</td>
<td>Сколькакая заделеневшая дорога</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Сколькакие заделеневшие ступеньки</td>
</tr>
<tr>
<td></td>
<td>Скользкая опорная поверхность, не покрытая льдом</td>
<td>Скользкий мокрый пол</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Скользкая полка в бане</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Скользкая лестница</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Скользкий паркет</td>
</tr>
<tr>
<td></td>
<td>Поверхность предмета, высказывающего из рук</td>
<td>Скользкая рыба (только что пойманная)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Скользкое мясо</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Скользкий мокрый мяч</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Скользкие грибы</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Скользкие руки</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Скользкий черенок лопаты (по которому скользят руки)</td>
</tr>
<tr>
<td></td>
<td>Скользкая обувь и подошва обуви</td>
<td>Скользкая подошва</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Скользкие сапоги</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Скользкие ботинки</td>
</tr>
<tr>
<td></td>
<td>Участок местности без зрительно заметных неровностей</td>
<td>ровное место (обширный участок местности)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ровная степь</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ровный луг</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ровная тундра</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ровный покос</td>
</tr>
<tr>
<td></td>
<td>Участок местности без холмов и гор</td>
<td>Плоское место (нет холмов и гор)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Плоская местность</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Плоский ландшафт</td>
</tr>
<tr>
<td>Гладкие поверхности тактильно воспринимаемых предметов: артефакты</td>
<td>гладкая доска</td>
<td>гладкий_доска</td>
</tr>
<tr>
<td>----------------------------------------------------------------</td>
<td>---------------</td>
<td>----------------</td>
</tr>
<tr>
<td>гладкий камень</td>
<td>гладкий_камень</td>
<td></td>
</tr>
<tr>
<td>гладкая дужка кровати</td>
<td>гладкий_дужка</td>
<td></td>
</tr>
<tr>
<td>гладкая миска</td>
<td>гладкий_миска</td>
<td></td>
</tr>
</tbody>
</table>

| Гладкие поверхности тактильно воспринимаемых предметов: части тела | гладкая кожа (у девушки) | гладкий_кожа |
|----------------------------------------------------------------|--------------------------|
| гладкая шерсть (у собаки) | гладкий_шерсть |
| гладкие щеки (чисто выбранные) | гладкий_щека |
| гладкий подбородок (чисто выбранный) | гладкий_подбородок |
| гладкие волосы (хорошо приглаженные) | гладкий_волосы |
| гладкие перышки птицы | гладкий_перышко |

<table>
<thead>
<tr>
<th>Ровная дорога (нет выъёнов, ухабов на поверхности)</th>
<th>Ровная дорога (заасфальтированная)</th>
<th>ровный_дорога</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ровное шоссе</td>
<td>ровный_шоссе</td>
<td></td>
</tr>
<tr>
<td>Ровная тропинка</td>
<td>ровный_тропинка</td>
<td></td>
</tr>
<tr>
<td>Ровный, хорошо накатанный зимник</td>
<td>ровный_зимник</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Поверхность воды в водоёме</th>
<th>Гладкая поверхность воды в озере (при штиле)</th>
<th>гладкий_поверхность</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Зрительно воспринимаемые поверхности артефактов: горизонтальные</th>
<th>Ровный пол</th>
<th>ровный_пол</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ровный потолок</td>
<td>ровный_потолок</td>
<td></td>
</tr>
<tr>
<td>Ровный настил</td>
<td>ровный_настил</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Зрительно воспринимаемые поверхности артефактов: вертикальные</th>
<th>Ровная стена</th>
<th>ровный_стена</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ровный забор</td>
<td>ровный_забор</td>
<td></td>
</tr>
<tr>
<td>Переносные значения</td>
<td>Ненадежность: сфера деятельности</td>
<td>Сколький вопрос (неприятно обсуждать)</td>
</tr>
<tr>
<td>---------------------</td>
<td>----------------------------------</td>
<td>----------------------------------------</td>
</tr>
<tr>
<td></td>
<td>Сколький путь (жизненный)</td>
<td>скользкий_путь</td>
</tr>
<tr>
<td></td>
<td>Сколлькая тема</td>
<td>скользкий_тема</td>
</tr>
<tr>
<td></td>
<td>Ненадежность: человек</td>
<td>Сколький тип (может обмануть, подвести)</td>
</tr>
<tr>
<td></td>
<td>Льстивый человек</td>
<td>льстивый_человек</td>
</tr>
<tr>
<td></td>
<td>Болтливый человек</td>
<td>болтливый_человек</td>
</tr>
<tr>
<td></td>
<td>Быстрота, легкость выполнения действия</td>
<td>Красноречивый человек</td>
</tr>
<tr>
<td></td>
<td>Ловкий человек</td>
<td>ловкий_человек</td>
</tr>
<tr>
<td></td>
<td>Сообразительный человек</td>
<td>сообразительный_человек</td>
</tr>
<tr>
<td></td>
<td>Отсутствие недостатков, затруднений: характеристика человека</td>
<td>Красивая девушка</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Стройная девушка</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Добрый человек</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Мягкий человек</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Аккуратный человек</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Красивый парень</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Стройный парень</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Красивый человек</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Стройный человек</td>
</tr>
<tr>
<td></td>
<td>Отсутствие недостатков, затруднений: вкус</td>
<td>Приятный вкус (+ сочетаемость с названиями естественных для информанта блюд)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Равно расставленные стулья.</td>
</tr>
<tr>
<td></td>
<td>Равные расстояния между селами.</td>
<td>ровный_расстояние</td>
</tr>
<tr>
<td></td>
<td>Ровный загар.</td>
<td>ровный_загар</td>
</tr>
<tr>
<td></td>
<td>Одинаково хорошее качество</td>
<td>стабильный_качество</td>
</tr>
<tr>
<td></td>
<td>Равные промежутки времени.</td>
<td>равный_промежуток</td>
</tr>
<tr>
<td></td>
<td>Равномерность: действия, процессы</td>
<td>Ровный звук (всё время одинаковый).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ровный ритм</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Постоянная скорость</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ровное гудение</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ровное дыхание</td>
</tr>
<tr>
<td></td>
<td>Ровная линия, край</td>
<td>Ровная линия (прямая)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Прямая дорога</td>
</tr>
</tbody>
</table>

168
<table>
<thead>
<tr>
<th>Отсутствие ресурса, отличительных качеств; малая интенсивность</th>
<th>Гладкая ткань (без рисунка)</th>
<th>гладкий_ткань</th>
</tr>
</thead>
<tbody>
<tr>
<td>Простой, легкий для понимания текст (положительная оценка)</td>
<td>Простой_текст</td>
<td></td>
</tr>
<tr>
<td>Простой, поверхностный текст (отрицательная оценка)</td>
<td>Поверхностный_текст</td>
<td></td>
</tr>
<tr>
<td>(Варить) на слабом огне.</td>
<td>Слабый_огонь</td>
<td></td>
</tr>
<tr>
<td>Слабая боль.</td>
<td>Слабый_боль</td>
<td></td>
</tr>
<tr>
<td>Тихий шепот</td>
<td>Тихий_шепот</td>
<td></td>
</tr>
<tr>
<td>Фасад без украшений.</td>
<td>Простой_фасад</td>
<td></td>
</tr>
<tr>
<td>Пустой карман (&quot;нет денег&quot;)</td>
<td>Пустой_карман</td>
<td></td>
</tr>
<tr>
<td>Обычный человек (не выделяется на фоне остальных)</td>
<td>Обычный_человек</td>
<td></td>
</tr>
<tr>
<td>Тихие шаги</td>
<td>Тихий_шаг</td>
<td></td>
</tr>
<tr>
<td>Тихий стук в дверь</td>
<td>Тихий_стук</td>
<td></td>
</tr>
<tr>
<td>Слабый дождь</td>
<td>Слабый_дождь</td>
<td></td>
</tr>
<tr>
<td>Медленное течение реки</td>
<td>Медленный_течение</td>
<td></td>
</tr>
</tbody>
</table>
## Приложение 3. Анкета для поля глаголов качания

<table>
<thead>
<tr>
<th>Фрейм</th>
<th>Микрофрейм</th>
<th>Словосочетание для дистрибутивного представления</th>
</tr>
</thead>
<tbody>
<tr>
<td>специальные приспособления для качания</td>
<td>качели</td>
<td>качели_качаться</td>
</tr>
<tr>
<td></td>
<td>человек на качелях</td>
<td>ребенок_качаться</td>
</tr>
<tr>
<td></td>
<td>кресло-качалка/колыбель (напольная)</td>
<td>колыбель_качаться</td>
</tr>
<tr>
<td>объекты на опоре</td>
<td>поезд на ходу</td>
<td>поезд_качаться</td>
</tr>
<tr>
<td></td>
<td>лодка на воде</td>
<td>лодка_качаться</td>
</tr>
<tr>
<td>мягкие горизонтальные поверхности</td>
<td>поверхность воды</td>
<td>поверхность_воды_колыхаться</td>
</tr>
<tr>
<td></td>
<td>трава на лугу</td>
<td>трава_колыхаться</td>
</tr>
<tr>
<td>мягкие вертикальные поверхности</td>
<td>пламя</td>
<td>пламя_колебаться</td>
</tr>
<tr>
<td></td>
<td>флаг</td>
<td>флаг_развеваться</td>
</tr>
<tr>
<td></td>
<td>занавеска, белье на веревке</td>
<td>занавеска_колыхаться</td>
</tr>
<tr>
<td>гибкие стебли</td>
<td>цветок</td>
<td>цветок_колыхаться</td>
</tr>
<tr>
<td></td>
<td>водоросли</td>
<td>водоросьль_колыхаться</td>
</tr>
<tr>
<td>закрепленные снизу</td>
<td>дерево</td>
<td>дерево_качаться</td>
</tr>
<tr>
<td></td>
<td>высокое здание</td>
<td>здание_качаться</td>
</tr>
<tr>
<td>закрепленные сверху</td>
<td>фонарь</td>
<td>фонарь_качаться</td>
</tr>
<tr>
<td></td>
<td>лампа</td>
<td>лампа_качаться</td>
</tr>
<tr>
<td></td>
<td>маятник</td>
<td>маятник_качаться</td>
</tr>
<tr>
<td>-контроль человека; мешающие ситуации</td>
<td>стул</td>
<td>стул_шататься</td>
</tr>
<tr>
<td>вертикальные</td>
<td>стол</td>
<td>стол_шататься</td>
</tr>
<tr>
<td></td>
<td>зуб перед тем как выпасть</td>
<td>зуб_шататься</td>
</tr>
<tr>
<td></td>
<td>пьяный человек</td>
<td>пьяный_шататься</td>
</tr>
<tr>
<td>-контроль человека; мешающие ситуации</td>
<td>колесо</td>
<td>колесо_болтаться</td>
</tr>
<tr>
<td>плохо закрепленные части целого</td>
<td>топор на рукоятке</td>
<td>топор_болтаться</td>
</tr>
<tr>
<td>(твердые)</td>
<td>брелок, ключи в связке</td>
<td>ключ_болтаться</td>
</tr>
<tr>
<td></td>
<td>сломанная рука</td>
<td>рука_болтаться</td>
</tr>
<tr>
<td>-контроль человека; мешающие ситуации</td>
<td>шнурки на одежде</td>
<td>шнурок_болтаться</td>
</tr>
<tr>
<td>плохо закрепленные части целого</td>
<td>провод от наушников</td>
<td>провод_болтаться</td>
</tr>
<tr>
<td>(мягкие)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Приложение 4. Автоматически сконструированная анкета для поля ‘прямой’

<table>
<thead>
<tr>
<th>Кластер</th>
<th>Содержание</th>
</tr>
</thead>
<tbody>
<tr>
<td>кластер 1</td>
<td>прямой_маркетинг, прямой_но, прямой_реклама</td>
</tr>
<tr>
<td>кластер 2</td>
<td>прямой_текст, прямой_цитата, прямой_перевод</td>
</tr>
<tr>
<td>кластер 3</td>
<td>прямой_предшественник, прямой_продолжатель, прямой_наследие</td>
</tr>
<tr>
<td>кластер 4</td>
<td>прямой_голосование, прямой_предложение, прямой_выбор</td>
</tr>
<tr>
<td>кластер 5</td>
<td>прямой_ущерб, прямой_убыток, прямой_вред</td>
</tr>
<tr>
<td>кластер 6</td>
<td>прямой_голова, прямой_спина, прямой_нога</td>
</tr>
<tr>
<td>кластер 7</td>
<td>прямой_волосы, прямой_прядь, прямой_пробор</td>
</tr>
<tr>
<td>кластер 8</td>
<td>прямой_участие, прямой_поддержка, прямой_руководство</td>
</tr>
<tr>
<td>кластер 9</td>
<td>прямой_умысел, прямой_измена, прямой_виновник</td>
</tr>
<tr>
<td>кластер 10</td>
<td>прямой_следствие, прямой_принципа, прямой_последствие</td>
</tr>
<tr>
<td>кластер 11</td>
<td>прямой_ответственность, прямой_нарушение, прямой_обязанность</td>
</tr>
<tr>
<td>кластер 12</td>
<td>прямой_дорожка, прямой_столб, прямой_аллея</td>
</tr>
<tr>
<td>кластер 13</td>
<td>прямой_мышца, прямой_движение, прямой_кишка</td>
</tr>
<tr>
<td>кластер 14</td>
<td>прямой_противник</td>
</tr>
<tr>
<td>Кластер</td>
<td>Словосочетания</td>
</tr>
<tr>
<td>---------</td>
<td>----------------</td>
</tr>
<tr>
<td>15</td>
<td>прямой атака</td>
</tr>
<tr>
<td></td>
<td>прямой наводка</td>
</tr>
<tr>
<td>16</td>
<td>прямой начальник</td>
</tr>
<tr>
<td></td>
<td>прямой приказ</td>
</tr>
<tr>
<td></td>
<td>прямой начальство</td>
</tr>
<tr>
<td>17</td>
<td>прямой цель</td>
</tr>
<tr>
<td></td>
<td>прямой задача</td>
</tr>
<tr>
<td></td>
<td>прямой необходимость</td>
</tr>
<tr>
<td>18</td>
<td>прямой смысл</td>
</tr>
<tr>
<td></td>
<td>прямой противоположность</td>
</tr>
<tr>
<td></td>
<td>прямой образ</td>
</tr>
<tr>
<td>19</td>
<td>прямой зависимость</td>
</tr>
<tr>
<td></td>
<td>прямой результат</td>
</tr>
<tr>
<td></td>
<td>прямой метод</td>
</tr>
<tr>
<td>20</td>
<td>прямой расход</td>
</tr>
<tr>
<td></td>
<td>прямой затрата</td>
</tr>
<tr>
<td></td>
<td>прямой финансирование</td>
</tr>
<tr>
<td></td>
<td>прямой отношение</td>
</tr>
<tr>
<td></td>
<td>прямой характер</td>
</tr>
<tr>
<td></td>
<td>прямой проявление</td>
</tr>
</tbody>
</table>
Приложение 5. Автоматически сконструированная анкета для поля ‘толстый’

<table>
<thead>
<tr>
<th>Кластер 1</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>толстый_стен</td>
<td></td>
</tr>
<tr>
<td>толстый_столб</td>
<td></td>
</tr>
<tr>
<td>толстый_бревно</td>
<td></td>
</tr>
<tr>
<td>Кластер 2</td>
<td></td>
</tr>
<tr>
<td>широкий_подбородок</td>
<td></td>
</tr>
<tr>
<td>широкий_лоб</td>
<td></td>
</tr>
<tr>
<td>широкий_губа</td>
<td></td>
</tr>
<tr>
<td>Кластер 3</td>
<td></td>
</tr>
<tr>
<td>широкий_воротник</td>
<td></td>
</tr>
<tr>
<td>широкий_брюки</td>
<td></td>
</tr>
<tr>
<td>широкий_пальто</td>
<td></td>
</tr>
<tr>
<td>Кластер 4</td>
<td></td>
</tr>
<tr>
<td>широкий_использование</td>
<td></td>
</tr>
<tr>
<td>широкий_применение</td>
<td></td>
</tr>
<tr>
<td>широкий_определение</td>
<td></td>
</tr>
<tr>
<td>Кластер 5</td>
<td></td>
</tr>
<tr>
<td>широкий_общественность</td>
<td></td>
</tr>
<tr>
<td>широкий_дискуссия</td>
<td></td>
</tr>
<tr>
<td>широкий_обсуждение</td>
<td></td>
</tr>
<tr>
<td>Кластер 6</td>
<td></td>
</tr>
<tr>
<td>широкий_деятельность</td>
<td></td>
</tr>
<tr>
<td>широкий_привлечение</td>
<td></td>
</tr>
<tr>
<td>широкий_организация</td>
<td></td>
</tr>
<tr>
<td>Кластер 7</td>
<td></td>
</tr>
<tr>
<td>широкий_диапазон</td>
<td></td>
</tr>
<tr>
<td>широкий_спектр</td>
<td></td>
</tr>
<tr>
<td>широкий_интервал</td>
<td></td>
</tr>
<tr>
<td>Кластер 8</td>
<td></td>
</tr>
<tr>
<td>широкий_известность</td>
<td></td>
</tr>
<tr>
<td>широкий_популярность</td>
<td></td>
</tr>
<tr>
<td>широкий_поприще</td>
<td></td>
</tr>
<tr>
<td>Кластер 9</td>
<td></td>
</tr>
<tr>
<td>широкий_двери</td>
<td></td>
</tr>
<tr>
<td>широкий_окно</td>
<td></td>
</tr>
<tr>
<td>широкий_подоконник</td>
<td></td>
</tr>
<tr>
<td>Кластер 10</td>
<td></td>
</tr>
<tr>
<td>широкий_кольцо</td>
<td></td>
</tr>
<tr>
<td>широкий_лезвие</td>
<td></td>
</tr>
<tr>
<td>широкий_раструб</td>
<td></td>
</tr>
<tr>
<td>Кластер 11</td>
<td></td>
</tr>
<tr>
<td>широкий_волна</td>
<td></td>
</tr>
<tr>
<td>широкий_струя</td>
<td></td>
</tr>
<tr>
<td>широкий_поток</td>
<td></td>
</tr>
<tr>
<td>Кластер 12</td>
<td></td>
</tr>
<tr>
<td>широкий_понимание</td>
<td></td>
</tr>
<tr>
<td>широкий_понятие</td>
<td></td>
</tr>
<tr>
<td>широкий_смысл</td>
<td></td>
</tr>
<tr>
<td>Кластер 13</td>
<td></td>
</tr>
<tr>
<td>толстый_мальчик</td>
<td></td>
</tr>
<tr>
<td>толстый_девушка</td>
<td></td>
</tr>
<tr>
<td>толстый_девочка</td>
<td></td>
</tr>
<tr>
<td>Кластер 14</td>
<td></td>
</tr>
<tr>
<td>толстый_пачка</td>
<td></td>
</tr>
<tr>
<td>Кластер 15</td>
<td>Толстый_папка</td>
</tr>
<tr>
<td>------------</td>
<td>----------------</td>
</tr>
<tr>
<td>Кластер 16</td>
<td>Толстый_хвост</td>
</tr>
<tr>
<td>Кластер 17</td>
<td>Толстый_шея</td>
</tr>
<tr>
<td>Кластер 18</td>
<td>Толстый_палец</td>
</tr>
<tr>
<td>Кластер 19</td>
<td>Широкий_спин</td>
</tr>
<tr>
<td>Кластер 20</td>
<td>Широкий_улица</td>
</tr>
<tr>
<td>Кластер 21</td>
<td>Широкий_равнина</td>
</tr>
<tr>
<td>Кластер 22</td>
<td>Толстый_достоевский</td>
</tr>
<tr>
<td>Кластер 23</td>
<td>Широкий_ум</td>
</tr>
</tbody>
</table>