На правах рукописи

Рыжова Дарья Александровна

АВТОМАТИЗАЦИЯ ЛЕКСИКО-ТИПОЛОГИЧЕСКИХ ИССЛЕДОВАНИЙ: МЕТОДЫ И ИНСТРУМЕНТЫ

Резюме
диссертации на соискание ученой степени
кандидата филологических наук НИУ ВШЭ

Москва 2018

ОБЩАЯ ХАРАКТЕРИСТИКА ДИССЕРТАЦИИ

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 (Майсак & Рахилина 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 обширных 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** представляемой 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 квантитативных 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 квантитативное 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.

Все эксперименты, описанные в настоящей работе, проводятся на **материале** нескольких признаковых и глагольных семантических полей, уже исследованных вручную участниками группы MLexT: ‘острый’ (Кюсева 2012), (Kyuseva, Parina, & Ryzhova to appear), ‘гладкий’ (Кашкин 2013), (Kashkin & Vinogradova to appear), ‘прямой’ (Лучина 2014), ‘толстый’ (Kozlov & Privizentseva to appear), ‘качание’ (Шапиро 2015), ‘падение’ (Кузьменко & Мустакимова 2015; Reznikova & Vyrenkova 2015) и некоторые другие.

Апробация результатов исследования. Основные результаты исследования были представлены на XI Конференции по типологии и грамматике для молодых исследователей (г. Санкт-Петербург, 2014), мастер-классе по лексической типологии в Университете Хельсинки (г. Хельсинки, Финляндия, 2014), XVI Апрельской международной научной конференции НИУ ВШЭ (г. Москва, 2015), научном семинаре с профессором Института психолингвистики имени Макса Планка Асифой Маджид (г. Москва, 2015), Международной конференции по компьютерной лингвистике «Диалог 2015» (г. Москва, 2015), конференции «Проблемы компьютерной лингвистики» (г. Воронеж, 2015), I международной научно-практической конференции «Иностранные языки в науке и образовании: проблемы и перспективы» (г. Москва, 2015), международном научном семинаре «Компьютерная лингвистика и наука о языке» (г. Москва, 2016), постерной секции Типологической школы Школы лингвистики НИУ ВШЭ (г. Москва, 2016), X Международной конференции по языковым ресурсам и их оценке LREC’16 (г. Порторож, Словения, 2016), международном семинаре по перцептивной

метафоре (г. Неймеген, Голландия, 2016), международном семинаре по глаголам движения (г. Париж, Франция, 2017), XIV Международной конференции по когнитивной лингвистике (г. Тарту, Эстония, 2017). По теме диссертации опубликовано 9 работ, в том числе 5 в изданиях, рекомендованных ВАК.

Структура работы. Работа состоит из Введения, пяти глав, Заключения, Библиографии из 158 названий и пяти Приложений. Во Введении содержится общая характеристика диссертации. 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.

ОСНОВНОЕ СОДЕРЖАНИЕ ДИССЕРТАЦИИ

Глава 1 «Введение» содержит краткую характеристику существующих методологий типологического анализа лексики. Основной постулат, на котором зиждется лексическая типология, заключается в том, что лексика разных языков системна и сопоставима. Разные подходы в этой области лингвистики отличаются друг от друга прежде всего тем, какие параметры для сравнения слов они выбирают. Эти параметры определяют и выбор источников данных, и методы работы с ними.

На сегодняшний день в лексической типологии можно выделить пять основных направлений:

(1) Экспериментальный подход, начало которому положила работа Berlin, Kay 1969, посвященная типологии цветообозначений. Эта методология опирается на метод элицитации с использованием анкет, состоящих из экстралингвистических стимулов, т.е. в качестве параметров для сравнения выбираются характеристики объектов или ситуаций, которые воспринимаются органами чувств.

(2) Теория семантических примитивов, разрабатываемая А. Вежбицкой и К. Годдардом (Wierzbicka 1985). Согласно этому подходу, значение любого слова естественного человеческого языка складывается из очень узкого набора универсальных семантических примитивов (таких, как «я», «ты», «что-то», «большой», «думать» и т.п.). Значению разных слов соответствуют разные комбинации универсальных смыслов.

(3) Серия подходов, опирающихся на словарные данные. Выделяют словарные подзначения и ищут закономерности в способах их «колексификации» (в терминах François 2008), т.е. объединения в рамках конкретных лексических средств.

(4) Серия подходов, опирающихся на данные параллельных корпусов (см., например, Viberg 2006, Wälchli, Cysouw 2012). Используют общий набор контекстов в качестве аналога лексико-типологической анкеты и определяют разницу между словами по тому, в каких контекстах они могут или не могут употребляться.

(5) Фреймовый подход к лексической типологии, основывающийся на предположении о существовании некоторого универсального набора минимальных лексических значений (фреймов). Предполагается, что каждое семантическое поле характеризуется своим набором фреймов, а разные слова покрывают разные их комбинации.

В первом разделе Главы 1 обсуждаются первые четыре методологии, выделяются их сильные и слабые стороны, подчеркивается общая тенденция к внедрению компьютерных методов сбора и анализа данных в процесс исследования.

Подробному обсуждению пятого подхода, который мы принимаем за основу в своем исследовании, посвящен раздел 2 Главы 1. Этот подход восходит к традициям Московской Семантической Школы (см. Апресян 1974) 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* и т.п.
- …

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

Такие ситуации называются **фреймами**. Предполагается, что фреймы – это минимальные лексические значения, т.е. каждая конкретная лексема покрывает ту или иную их комбинацию. При этом не все комбинации фреймов одинаково вероятны: какие-то значения часто объединяются в рамках одного лексического средства, а какие-то, напротив, в большинстве случаев оказываются лексически противопоставлены. Закономерности объединения фреймов в рамках одной лексемы представляются графически в виде лексико-семантических карт.

Такой метод изучения семантики слов был опробован на обширном лексическом материале, ср. Майсак, Рахилина 2007, Круглякова 2010, Кашкин 2013, Холкина 2014, Кюсева 2012. В ходе проведенных исследований доказано, что этот подход действительно

позволяет выявлять фреймовую структуру каждого семантического поля и сопоставлять лексику разных языков.

Исследование лексических единиц в этой парадигме включает несколько основных этапов:

1. Составление анкеты (т.е. предварительное определение набора фреймов) на основе анализа сочетаемости лексем выбранного поля в 3-5 языках.

2. Сбор данных других языков выборки для уточнения набора фреймов.

3. Составление семантической карты для описания системы каждого языка и ее визуализации.

4. Анализ типов систем, реализованных в разных языках

Для определения набора фреймов, релевантных для рассматриваемого поля, необходимо провести подробный анализ словарных и корпусных данных, дополнив их в ходе опросов носителей. Поскольку основная задача исследования – определить правила сочетаемости для каждой лексемы, относящейся к данному полю, анкеты для работы с носителями содержат контексты, в которых могут употребляться изучаемые слова, а это означает, что итоговую анкету необходимо переводить на каждый из языков, включаемых в выборку.

До сих пор практически все работы в рамках этого подхода осуществлялись вручную и требовали долгой, кропотливой и согласованной работы специалистов по всем языкам, включенным в выборку. Трудоемкость 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.

Глава 2 «Верификация понятия фрейма с помощью моделей дистрибутивной семантики» описывает 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 фреймы F1 и F2, а лексема L2 – значения F2 и F3, но при этом нет ни одного слова, которое означало 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.

Опираясь на пилотное исследование Кюсева 2014, мы разработали формулу более точного, численного определения типологической близости между фреймами на основе данных о частоте колексификации минимальных значений. Каждый фрейм представляется

в виде вектора *w*, в качестве измерений которого выступают лексемы изучаемого поля из всех языков выборки. В случае, если лексема l_i может описывать данный фрейм, соответствующее ей измерение w_i принимает значение 1, а если не может – 0. Типологическое расстояние между фреймами определяется с помощью косинусной меры близости между представляющими их векторами (ср. похожую метрику близости в недавней работе Youn et al. 2016).

Однако известны и другие методы определения семантических расстояний между лексическими значениями, в частности, представление значения лексической единицы (слова или словосочетания) с помощью вектора его сочетаемости (т.н. модели дистрибутивной семантики, см. Baroni et al. 2013). Такие семантические представления используются для решения широкого круга задач, в том числе, близких к нашей (например, для семантической дизамбигуации или выбора из ряда квазисинонимов 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 минимальные лексические значения, то дистрибутивные расстояния между ними должны соответствовать типологическим.

Мы провели серию экспериментов на материале признаковых полей ‘острый’ и ‘гладкий’. Наборы фреймов, а также данные для вычисления типологических расстояний между ними были взяты из Типологически ориентированной базы данных признаковой лексики (см. Кюсева и др. 2013а).

Для каждого фрейма было взято по несколько иллюстраций (или «микрофреймов», например, для фрейма ‘инструмент с колющим концом’ из поля ‘острый’ – ‘острая игла’, ‘острая стрела’, ‘острое копье’). Каждому микрофрейму было поставлено в соответствие двусловное русское сочетание (*острая игла, острая стрела, острое копье*), и для каждого такого словосочетания построен вектор сочетаемости.

Для построения дистрибутивных моделей использовались следующие параметры, которые подробно обсуждаются в тексте диссертации:

1. **Измерения:** 10 000 самых частотных лемм самостоятельных частей речи (по основному подкорпусу НКРЯ)
2. **Значения измерений:** частота встречаемости в окне ±5 слов самостоятельных частей речи от единицы, для которой строится вектор
3. **Расстояния** между векторами определяются с помощью косинусной меры близости
4. **Обучающий корпус:** основной подкорпус НКРЯ (около 220 млн словоформ), газетный подкорпус НКРЯ (около 200 млн словоформ) и корпус интернет текстов ruWaC (около 1 млрд словоформ) в разных комбинациях
5. **Дополнительная обработка векторов:** взвешивание (нет vs. PPMI vs. PLMI vs. PLOG vs. EPMI) и сокращение размерности (нет vs. SVD до 300 измерений)
6. **Тип вектора словосочетания:** наблюдаемый (словосочетание, для которого строится вектор, принимается за единую лексическую единицу) vs. комбинированный (вектор для словосочетания составляется из векторов входящих в него слов по одной из следующих моделей композиции: аддитивная (additive), аддитивная взвешенная (weighted additive), мультипликативная (multiplicative), расширение (dilation), лексическая функция (lexical function), практическая лексическая функция (practical lexical function, PLF)).

Таким образом, для каждого микрофрейма мы получили по два векторных представления: типологическое и дистрибутивное. Далее для всех возможных пар микрофреймов внутри каждого поля были вычислены типологическое и дистрибутивное расстояния и подсчитан коэффициент корреляции Пирсона между этими двумя метриками.

Для обоих полей коэффициент корреляции получился очень высоким (0.766 для поля ‘острый’ и 0.946 для поля ‘гладкий’). Поскольку некоторые параметры дистрибутивных моделей варьировались, важно отметить, что наилучшие результаты для двух признаковых зон были получены на одних и тех же настройках: в качестве обучающего корпуса использовался основной подкорпус 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, см. Paperno et al. 2014). 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 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.

Таким образом, вопреки распространенному мнению (см., например, Bullinaria and Levy, 2012), качество модели не растет пропорционально увеличению объема обучающего корпуса: в нашем случае небольшой, но хорошо сбалансированный основной подкорпус НКРЯ дает более высокий результат, чем объединенный корпус общим объемом около 1,44 млрд словоформ, включающий основной и газетный подкорпусы НКРЯ и корпус ruWaC (ср. аналогичное наблюдение в работе Kutuzov, Kuzmenko 2015).

Для создания качественного векторного представления отдельных лемм основного подкорпуса 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.

Поскольку результатов, полученных на материале двух семантических полей, недостаточно для того, чтобы делать содержательные выводы с высокой степенью уверенности, мы провели два дополнительных эксперимента. В одном из них мы использовали параметры дистрибутивных моделей, которые дали самые лучшие результаты в первых двух экспериментах, но в качестве тестового материала выбрали не признаковое, а глагольное семантическое поле (‘качание’)[1]. Второй дополнительный эксперимент был проведен на базе поля ‘острый’, но с другим обучающим корпусом – англоязычным ukWaC (тем самым, типологические расстояния, которые остались неизменными, сопоставлялись с дистрибутивными расстояниями между соответствующими английскими словосочетаниями, например, *sharp needle* ‘острая игла’,

[1] Мы выражаем благодарность Марии Шапиро, которая предоставила нам типологический материал для этого эксперимента.

sharp spear ‘острое копье’, *sharp arrow* ‘острая стрела’ и т.п.). В обоих случаях было получено высокое значение коэффициента корреляции Пирсона: 0.7 для поля качания и 0.668 в эксперименте на базе англоязычного корпуса. Эти результаты дополнительно подтверждают гипотезу о том, что фреймовая структура поля может быть примерно очерчена уже на материале одного языка, причем неважно, какого именно.

Наконец, соответствие между типологическим и дистрибутивным пространствами хорошо иллюстрируется их визуализациями. Для каждого поля мы отобразили каждое пространство на плоскость с помощью техники многомерного шкалирования, обозначив одним цветом точки, относящиеся к одному и тому же фрейму. На Рис. 1-3 представлены проекции типологического и дистрибутивного пространств поля ‘острый’. На всех картах зеленым цветом обозначены ситуации, относящиеся к фрейму ‘острый инструмент с режущим краем’, синим – ‘острый инструмент с колющим концом’, желтым – ‘объект вытянутой формы’, красным – ‘объект с колючей поверхностью’. Подчеркнем, что эти кластеры выделялись не на основе полученных карт, а были заданы изначально результатами типологических исследований группы MLexT, т.е. в этом разделе мы будем говорить о «зеленом», «синем», «желтом» и «красном» кластерах, имея в виду соответствующие четыре фрейма поля ‘острый’.

Рисунки 1-3 демонстрируют интересный эффект. Визуализация типологического пространства (Рис. 1) наглядно отображает фреймовую структуру поля[2]. Визуализация дистрибутивного пространства, напротив, отражает только те противопоставления, которые лексикализованы в данном языке. Так, например, карта на Рис. 2 построена на основе данных русского языка, и по ней четко выделяются контексты для прилагательного *колючий*, а фреймы, 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.

[2] Заметим, что метод многомерного шкалирования успешно применяется в типологии как раз для автоматического построения семантических карт (см. Croft and Poole 2008, Wälchli and Cysouw 2012 и др.). Мы будем говорить об этом подробнее в Главе 6.

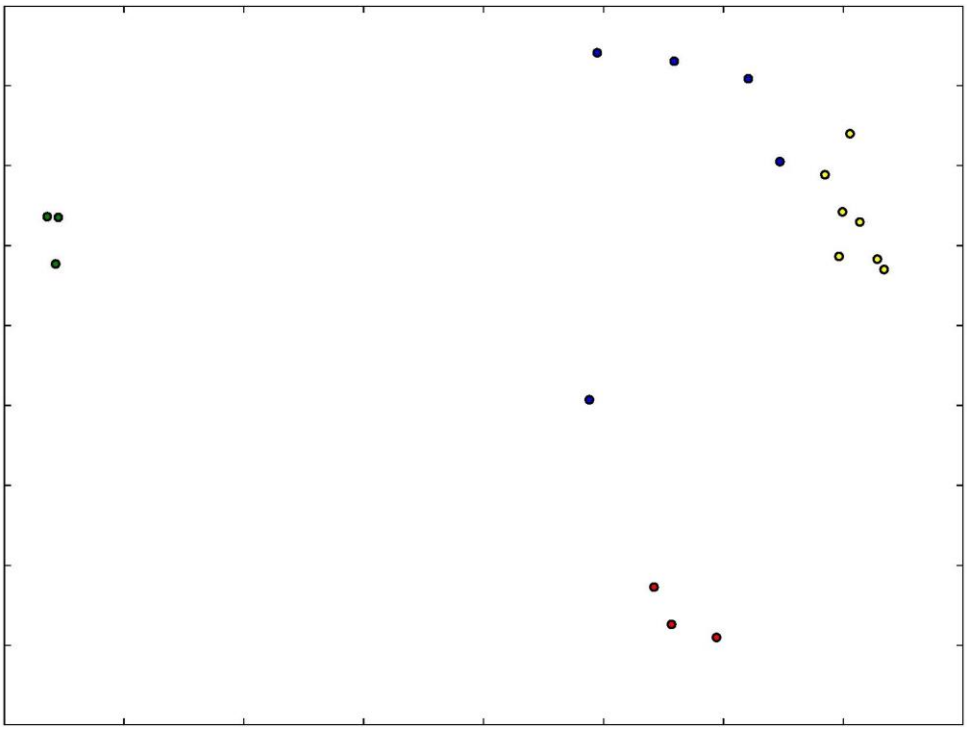

Рис. 1. Визуализация типологического пространства поля ‘острый’

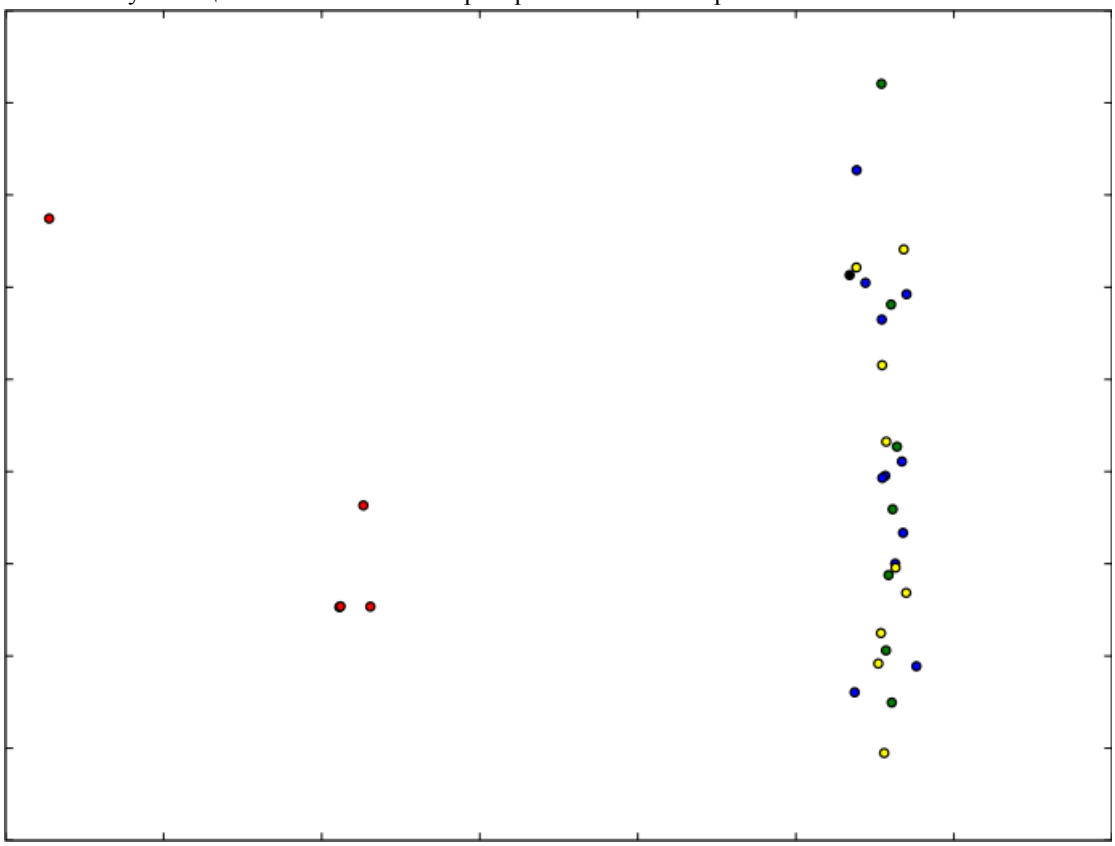

Рис. 2. Визуализация дистрибутивного пространства поля ‘острый’, построенного на основе русскоязычного корпуса

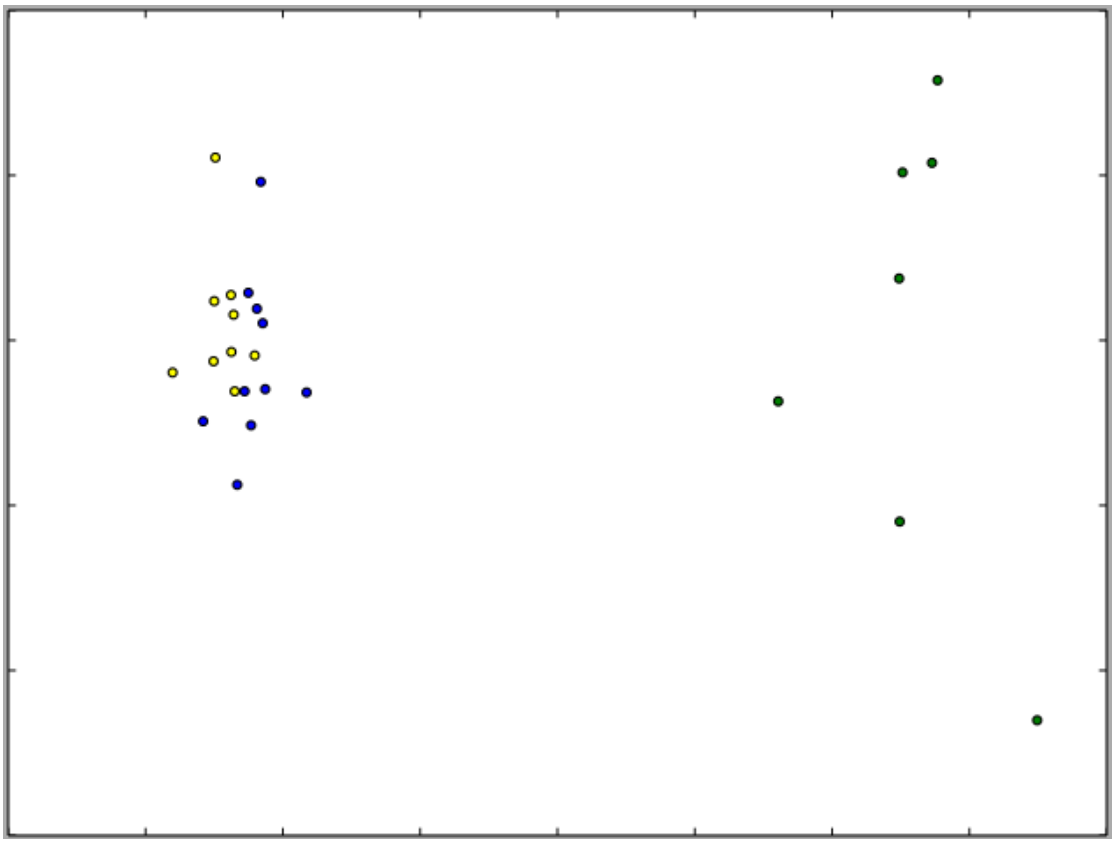

Рис. 3. Визуализация дистрибутивного пространства поля ‘острый’, построенного на материале франкоязычного корпуса (без учета фрейма ‘объект с колючей поверхностью’)

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.

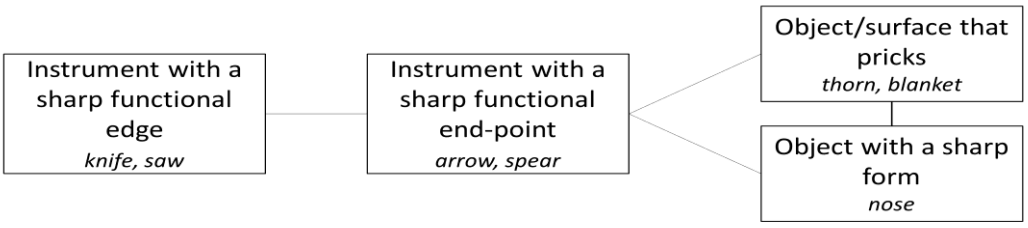

Рис. 4. Семантическая карта поля ‘острый’, составлена вручную на основе типологических данных

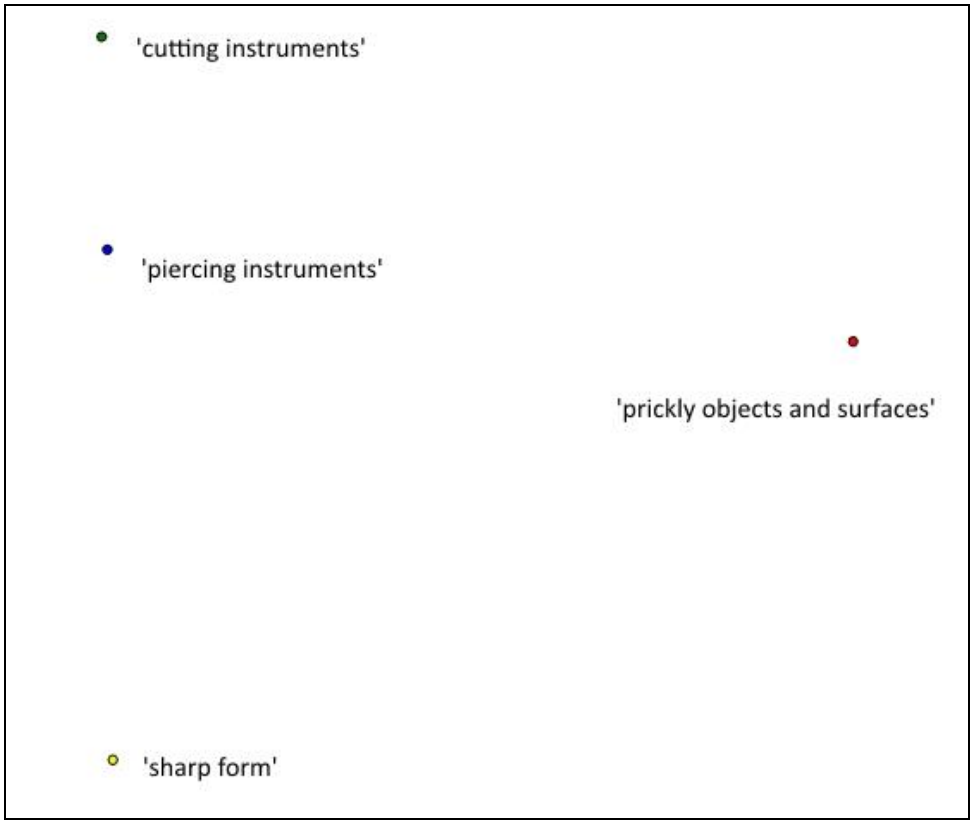

Рис. 5. Семантическая карта поля ‘острый’, составленная автоматически путем отображения на плоскость векторного пространства, состоящего из центральных представителей каждого фреймового кластера.

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. 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. 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 2012) – 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.

Теперь, когда мы привели дополнительное обоснование теоретической и практической состоятельности понятия фрейма, мы переходим к обсуждению возможных способов автоматизации этапов исследования в рамках фреймового подхода к лексической типологии.

Опираясь на наблюдение (3) из предыдущей главы, мы разработали метод построения лексико-типологической анкеты на материале одного языка. Этот метод мы описываем в Главе 3 «Автоматическая разработка анкеты с помощью моделей дистрибутивной семантики».

Алгоритм, который мы предлагаем, позволяет создать предварительный вариант анкеты для типологического исследования признаков слов или других одноместных предикатов, таких как глаголы движения, звука, состояний. Он включает несколько этапов:

1. Составление списка существительных, с которыми могут сочетаться рассматриваемые прилагательные/глаголы (по основному подкорпусу НКРЯ);
2. Представление каждого словосочетания в виде вектора его сочетаемости;
3. Разбиение полученного векторного пространства на кластеры методом иерархической кластеризации;
4. Выделение трех центральных элементов из каждого кластера и удаление всех кластеров, включающих менее трех элементов.

Обратим внимание, что этапы 2-4 остаются неизменными при работе с любыми классами лексики. Этап 1 зависит от формата минимального диагностического контекста для исследуемых слов. Мы исходили из допущения, что для определения значения признакового слова или другого одноместного предиката достаточно существительного, занимающего позицию его единственного актанта. Соответственно, для признаковой лексики мы учитывали существительные, которые встречаются в основном подкорпусе НКРЯ справа от опорного слова, а для глагольной – существительные, которые стоят справа или слева от заданного глагола и имеют при этом форму именительного падежа.

Алгоритм разрабатывался и тестировался на материале четырех признаковых (‘острый’, ‘гладкий’, ‘прямой’, ‘толстый’) и одного глагольного поля (‘качание’). Для каждого поля мы оценивали полноту и точность итоговой анкеты. Полнота определялась по доле фреймов, представленных в анкете хотя бы одной иллюстрацией, а метрика точности отображала чистоту полученных кластеров.

	Полнота	**Точность**	**F-мера**
‘острый’	0,733	0,827	0,777
‘прямой’	1	0,817	0,899
‘гладкий’	0,8	0,675	0,732
‘толстый’	1	0,884	0,938
‘качание’	0.882	0.762	0.818

Таблица 1. Количественная оценка работы алгоритма

Из Таблицы 1, в которой представлены значения метрик оценки качества работы алгоритма для каждого тестового поля, видно, что в целом метод работает достаточно хорошо, однако для одних семантических зон он дает значительно более высокие результаты, чем для других. По-видимому, это связано с несколькими факторами.

Во-первых, важную роль играет частотность анализируемого прилагательного. Чем лексема частотнее (другими словами, чем больше вхождений лексемы в корпус, который обрабатывается алгоритмом), тем более точным будет результат. По-видимому, низкое значение F-меры у поля глаголов качания связано именно с малой частотностью входящих в него лексем и, как следствие, недостаточно высоким качеством векторного представления словосочетаний и точности их кластеризации.

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, в-третьих, на качество анкеты влияет природа оппозиций, организующих семантическую структуру поля. Метод автоматического построения анкет, который мы предлагаем, 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 алгоритм фиксирует с меньшим успехом.

Эти факторы не являются равноценными. Так, несмотря на то, что в поле ‘толстый’ важную роль играет топологическая классификация предметов, алгоритм предоставил для него наилучший результат. 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, пальто, свитер).

Глава 4 «Методы автоматического сбора данных» упрощает этап сбора материала, который сводится к решению двух задач: (1) перевода анкеты, состоящей из минимальных контекстов; (2) заполнения этой анкеты материалом соответствующих языков. Эксперименты в этой области проводились на материале качественных признаков (‘острый’, ‘гладкий’, ‘толстый’), поэтому задача (1) заключалась в переводе списка прилагательных, относящихся к рассматриваемому полю, и списка существительных, с которыми они потенциально могут сочетаться.

Проблема перевода прилагательных очень нетривиальна. Традиционно задача перевода (в том числе автоматического) подразумевает либо выбор самой подходящей лексемы для определенного контекста, либо подбор наиболее частотного / близкого переводного эквивалента для данной лексемы, либо выдачу самого удачного эквивалента для каждого из значений исходного слова. Наша цель отличается ото всех перечисленных: нам необходимо подобрать такие прилагательные, которые могут выступать в качестве переводов исходных слов, но только в контекстах, соответствующих их прямым употреблениям. Так, например, в числе английских переводных эквивалентов слова *острый* мы бы хотели видеть прилагательные *sharp* и *pointed*, но не *critical* или *urgent* (ср. *острая нехватка, острый вопрос*).

Проработав несколько разных алгоритмов (каждый из которых подробно описывается в основном тексте диссертации), мы остановились на методе, который опирается на машиночитаемые словари группы FreeDict. Преимущество этих словарей заключается в том, что возможные переводы в них размечены по тому, к какому значению исходного слова они относятся. Наш алгоритм выбирает переводные эквиваленты только для первого значения исходного прилагательного, а потом проводит дополнительную проверку по методу обратного перевода: найденное слово переводится обратно на исходный язык и включается в итоговый список только в том случае, если прилагательное, являющееся его эквивалентом в первом значении, входит в стартовый список признаковых слов. Существительные переводятся по той же методологии, но с небольшой модификацией: в случае, если существительного нет в словаре FreeDict, перевод для него подбирается по соответствующему машиночитаемому словарю компании Яндекс.

Далее анкета переводится в табличный формат, где в качестве столбцов выступают прилагательные, а в качестве строк – существительные. Полученная таблица заполняется по материалам доступных корпусов: если прилагательное встречается в комбинации с тем или иным существительным в корпусе, для этой пары подсчитывается значение взаимной информации. Комбинации с отрицательным значением взаимной информации считаются случайными и исключаются из итоговой заполненной анкеты.

Автоматизация заключительного этапа рассматривается в Главе 5 «Автоматическое построение семантической карты с помощью решеток формальных понятий». Особое внимание уделяется теории анализа формальных понятий (Ganter, Wille 1999), которая

позволяет строить особого рода диаграммы – решетки формальных понятий (РФП). Мы утверждаем, что такие диаграммы могут использоваться в лингвистических исследованиях в качестве семантических карт нового типа.

РФП строятся на основе так называемых формальных контекстов. Формальный контекст *K = (G, M, I)* – это множество объектов *(G)*, множество признаков *(M)* и бинарное отношение *(I)*, связывающее объекты и признаки, которыми они обладают. Формальное понятие – это такая пара *(A, B)*, где *A* является подмножеством *G*, а *B* – подмножеством *M*, причем в *B* содержатся все признаки, которыми характеризуются объекты из *A*, а в *A* – все объекты, обладающие признаками из *B*, в рамках данного формального контекста. РФП представляет данные в виде иерархии формальных понятий, где понятия упорядочиваются от более общих (охватывающих большее количество объектов) к менее общим (покрывающим меньшее число объектов).

В нашем случае в качестве объектов выступают лексемы, в качестве признаков – фреймы. Между лексемой и фреймом устанавливается отношение инцидентности, если эта лексема покрывает данный фрейм. Эксперименты проводились на материале 10 признаковых полей (‘острый’, ‘мягкий’, ‘гладкий’, ‘шершавый’, ‘твердый’, ‘пустой’, ‘толстый’, ‘тонкий’, ‘высокий’ и ‘низкий’) и глагольного поля падения.

Насколько нам известно, ранее этот метод практически не использовался лингвистами (одно из немногочисленных исключений – работа Priss 2005). Наши эксперименты позволили определить границы его применимости для решения лексико-типологических задач. Было выявлено, что РФП сама по себе, без дополнительных модификаций, может использоваться в качестве семантической карты для полей с линейной структурой - таких, у которых конфигурация фреймов на традиционной семантической карте представляет собой цепочку (Фрейм 1 - Фрейм 2 - Фрейм 3). В таких случаях расположение узлов решетки соответствует расположению фреймов на традиционной карте, однако решетка строится автоматически, а традиционная семантическая карта – вручную.

Более того, РФП не только автоматизирует процесс построения карт, но и предоставляет исследователю новые возможности. Так, во-первых, благодаря иерархической организации узлов решетки становится возможным отобразить на одной схеме одновременно все допустимые стратегии лексикализации поля (в то время как традиционная техника семантического картирования предполагает отдельное изображение каждой конкретно-языковой реализации совмещения фреймов в рамках данного поля), см. Рис. 6. Это, в свою очередь, значительно упрощает задачу анализа типологии систем: оказывается, некоторые комбинации, допускаемые традиционной семантической картой, не реализовываются никогда или реализовываются очень редко, а другие, наоборот, очень частотны. Так, например, РФП для поля ‘острый’ показывает, что в нашей выборке чаще всего встречаются две стратегии: доминантная, при которой все основные фреймы покрываются одной лексемой, и бинарная, при которой одна лексема описывает инструменты с режущим краем (ножи, пилы, бритвы), а вторая объединяет инструменты с колющим концом (копья, стрелы) и объекты вытянутой формы (нос, носок ботинка). Интересно, что традиционная семантическая карта этой тенденции не отображает: на ней фрейм ‘острый (об инструментах с колющим концом)’ просто расположен между двумя

другими, из чего следует, что его объединения с фреймом ‘острый (об объектах вытянутой формы)’ и с фреймом ‘острый (об инструментах с режущим краем)’ равновероятны (см. Рис. 4 выше).

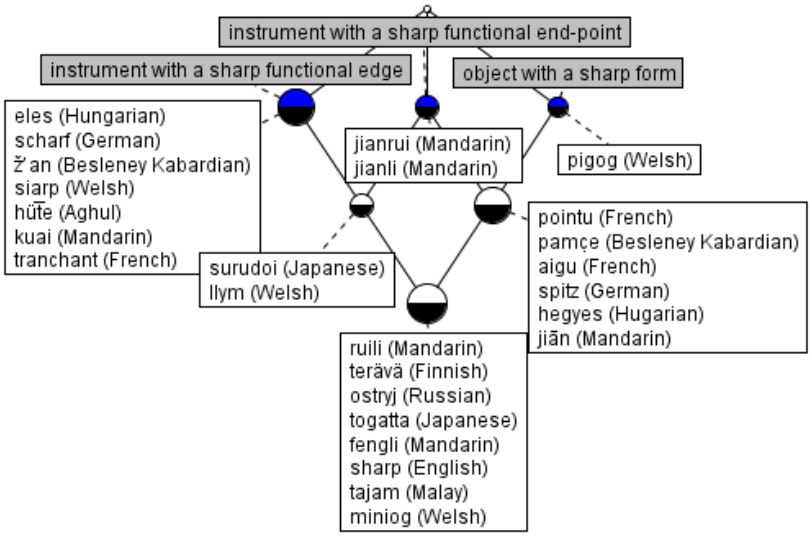

Рис. 6. РФП для семантического поля ‘острый’

Во-вторых, иерархическая организация узлов позволяет помещать на карту одновременно прямые и переносные значения. При этом схематическое изображение степени близости прямых употреблений сохраняется, поскольку все соответствующие узлы находятся на одном уровне решетки, а узлы, отражающие фреймы переносных значений, занимают следующий уровень. Такая конфигурация позволяет проследить взаимосвязь между исходной и результирующей семантикой лексем, подвергшихся метафоризации, и наглядно изобразить модели типологически релевантных семантических сдвигов (см. Рис. 7), что полностью соответствует наблюдениям о природе противопоставлений прямых и переносных значений, высказанным в работе Рахилина, Резникова 2013.

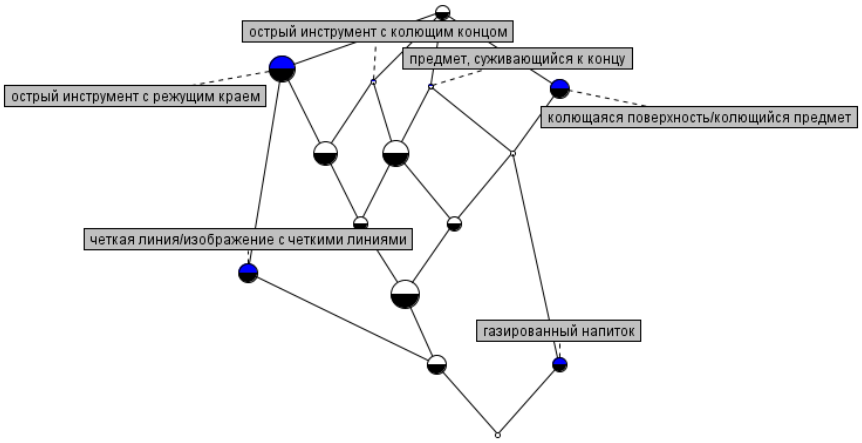

Рис. 7. РФП, отображающая некоторые связи прямых и переносных значений в семантическом поле ‘острый’

В тех же случаях, когда топология семантического поля сложнее линейной, отображающая такое поле РФП становится неудобочитаемой – в связи с тем, что решетки формальных понятий более информативны, чем традиционные семантические карты.

Таким образом, эта методология может:

(1) использоваться для автоматического построения (усовершенствованных) семантических карт для полей с линейной организацией или же для линейных участков полей с более сложной структурой;

(2) служить основой для нового алгоритма автоматического построения семантических карт, превращающего РФП в традиционный граф. Для реализации такого алгоритма, однако, необходимо определить, какая информация должна быть сохранена, а какого рода закономерности, улавливаемые с помощью анализа формальных понятий, можно не учитывать при переходе к графовой модели.

В Заключении формулируются основные результаты работы:

(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.

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.
2. *Фантастическая конференция, чудовищный доклад*: формирование оценочных значений на базе русской признаковой лексики // Вестник Московского университета. Серия 9: Филология. 2016. № 4. С. 178–192.
3. Грамматическая полисемия сквозь призму лексики: инструменталис в бесленеевском диалекте кабардино-черкесского языка // В кн.: ACTA LINGUISTICA PETROPOLITANA. Труды Института лингвистических исследований РАН / Отв. редактор Н. Н. Казанский. Т. XII. Ч. 2. Материалы Десятой конференции по типологии и грамматике для молодых исследователей (2013 г.) / Отв. ред. тома Д.В. Герасимов. СПб.: Наука, 2016. С. 665–678. В соавторстве с П.М. Аркадьевым и М.В. Кюсевой.
4. Выражение локативных значений в кубанском диалекте кабардино-черкесского языка (на примере глаголов падения) // Вестник Воронежского государственного университета. Серия: Лингвистика и межкультурная коммуникация. 2016. № 2. С. 79–85. В соавторстве с М.В. Кюсевой.
5. Построение лексико-типологической анкеты с помощью моделей дистрибутивной семантики // Вестник Воронежского государственного университета. Серия: Лингвистика и межкультурная коммуникация. 2015. № 3. С. 127–132.
6. Formal concept lattices as semantic maps // Proceedings of the 1st International Workshop on Computational Linguistics and Language Science (CLLS 2016), CEUR-WS.org, Eds. D. Ilvovsky, E. Chernyak, A. Vybornova, D. Skorinkin. 2017. Co-authored with S. Obiedkov.
7. Глаголы звучания как материал к теории семантических моделей // В кн. «Глаголы звуков животных – типология метафор» / Ред. Т. И. Резникова, А.С. Выренкова, Б. В. Орехов, Д.А. Рыжова. Языки славянской культуры, 2015. С. 325–343. (в соавторстве с Е.В. Рахилиной и М.В. Кюсевой)
8. Глаголы звуков животных в хинди // Глаголы звуков животных – типология метафор / под ред. Т. И. Резниковой, А.С. Выренковой, Б. В. Орехова, Д.А. Рыжовой. Языки славянской культуры, 2015. С. 141–155. (в соавторстве с Е.В. Бессоновой (Козловой))
9. Глаголы звуков животных в бжедугском диалекте адыгейского языка// Глаголы звуков животных – типология метафор / под ред. Т. И. Резниковой, А.С. Выренковой, Б. В. Орехова, Д.А. Рыжовой. Языки славянской культуры, 2015. С. 233-244. (в соавторстве с М.В. Кюсевой)
10. Typology of Adjectives Benchmark for Compositional Distributional Models // Proceedings of the Language Resources and Evaluation Conference, 2016. P. 1253–1257. Co-authored with M. Kyuseva and D. Paperno.