Правительство Российской Федерации

Федеральное государственное автономное образовательное учреждение высшего профессионального образования

Национальный исследовательский университет "Высшая школа экономики"

Факультет филологии

Направление подготовки 035800.68
«Фундаментальная и прикладная лингвистика»

КУРСОВАЯ РАБОТА

На тему

«Верификация фреймового подхода к лексической типологии с помощью векторных моделей»

Студент группы № 72л
Кюсева Мария Викторовна

Научные руководители
Рахилина Екатерина Владимировна, д.ф.н.,
Руководитель направления «Фундаментальная и прикладная лингвистика», профессор; заведующая лингвистической лабораторией по корпусным технологиям

Паперно Денис Аронович, PhD,
Научный сотрудник Вычислительной лаборатории по изучению языка и коммуникации (CLIC) университета Тренто

Москва, 2014 г.

Оглавление

1. Введение

Настоящая работа посвящена применению методов векторной семантики в лексической типологии.

Лексическая типология – область лингвистики, которая сравнивает структуру лексических полей в разных языках. Согласно подходу, принятому в настоящей работе, основным инструментом ее исследования является анализ сочетаемости лексем. Сравнивая контексты, которые могут покрывать лексемы со схожими значениями в разных языках, лексические типологи выделяют единицы смысла, которые всегда объединяются в одной лексеме. Эти единицы смысла называются *фреймами*. Фреймы, универсальные для всех языков, являются элементами, организующими структуру любого лексического поля. В задачи лексических типологов входит как определение набора фреймов, так и изучение стратегий их совмещения.

Если одни и те же фреймы стоят в основе лексического состава всех языков, то они должны каким-то образом семантически проявлять себя и на одном языке.

Именно эту гипотезу мы и проверим при помощи статистического метода представления семантики лексики – моделей векторной семантики. Они исходят из дистрибутивной гипотезы, очень близкой лексической типологии. Согласно этой гипотезе, значение слова/словосочетания можно выразить через контексты, в которых оно встречается. В векторных моделях семантика лексической единицы представляется в виде многомерного вектора, значения координат которого представляют собой функцию от количества раз, которое целевое слово встретилось в определенном контексте. Модели дистрибутивной семантики, уже проявившие себя в поиске синонимов и когипонимов слов, никогда раньше не использовались для решения задач лексической типологии.

В настоящей работе при помощи моделей векторной семантики мы проведем серию экспериментов на материале русского языка, чтобы выяснить, проявляется ли в нем деление лексических полей на фреймы. Если результат окажется положительным, он выступит также в качестве верификации фреймового подхода в лексической типологии.

В первой части работы будут подробно описаны общие принципы и методология лексической типологии, а также одно из приложений, которое родилось в ее рамках – Типологическая база адъективной лексики. Во второй части будут описаны базовые понятия векторной семантики и принципы построения векторных моделей. В третьей части будет описан проведенный в рамках исследования эксперимент и оценены его результаты.

2. Лексическая типология: фреймовый подход

2.1 Основные понятия

Лексическая типология – направление в лингвистике, основной задачей которого является сравнение лексических единиц в разных языках (см. обзоры Рахилина, Плунгян 2007, Evans 2010, Koptjevskaja-Tamm 2008). Сформировавшись сравнительно недавно, лексическая типология стремительно набирает силу как самостоятельная область лингвистики, и на данный момент в ее рамках функционирует несколько школ и подходов. Среди них следует отдельно отметить работы психолингвистической школы института им. Макса-Планка в Неймегене и подход, предложенный Анной Вежбицкой (Wierzbicka 1972, 1996).

Первое направление исходит из гипотезы об одинаковом устройстве всей когнитивной деятельности человека, включая язык, и, следовательно, о возможности изучения языка через другие виды когнитивных механизмов. Вслед за классической работой, посвященной семантике цветообозначений, [Berlin, Kay 1969] последователи этой школы собирают языковые реакции носителей разных языков на один и тот же набор внеязыковых стимулов. В качестве таких стимулов выступают таблички разных цветов, характерные образцы с запахами или вкусами, набор видеоклипов или картинок, описывающих разные варианты какой-либо физической ситуации (разрушения [Majid, Bowerman (eds.) 2007], нахождения предмета в пространстве относительно ориентира [Levinson 2006]и т.д).

Представители второго направления (см., например, Goddard, Wierzbicka (eds.) 1994, Godddard 1998, 2008, Гладкова 2010) проводят исследования лексики в рамках так называемой NSM-модели (Natural Semantic Metalanguage Model), согласно которой значение любой языковой единицы можно описать при помощи универсального метаязыка, состоящего из ограниченного списка слов, переводные эквиваленты которых, всегда однозначные, присутствуют в любом языке. Сами толкования опираются при этом на интроспекцию исследователя.

Настоящая работа проведена в рамках принципиально иного подхода, который основан на опыте Московской лексико-типологической группы (MLexT) и который можно назвать *фреймовым* [Рахилина, Резникова 2013]. Согласно этому подходу, значение лексической единицы можно изучить, опираясь лишь на ее языковое поведение. Такая методология связывает этот подход с традицией Московской семантической школы, основной инструмент семантического исследования которой – это сравнение контекстной сочетаемости лексемы с сочетаемостью ее ближайших синонимов (см. Апресян 1974). Главное

открытие этой школы состоит в том, что полных синонимов не бывает: при тщательном исследовании их сочетаемости всегда можно обнаружить контексты, в которых один из синонимов естественен, а другой звучит странно или вовсе невозможен. Подробный анализ таких «зон запретов» позволяет очертить структуру лексического поля в конкретном языке. А допущение о том, что переводные эквиваленты слова функционально равнозначны его синонимам, позволяет расширить этот метод исследования в типологическую перспективу.

При исследовании сочетаемости квази-синонимов в разных языках оказывается, что есть такие семантические зоны, которые «никогда» не разрываются лексемами: если слово описывает какую-то часть такой зоны, оно обязательно описывает и все остальные ее части. Такие семантические зоны в рамках настоящего подхода называются *фреймами*. Описывая типичные смысловые ситуации лексического поля, именно они и организуют его структуру.

Использование анализа контекстной сочетаемости в качестве инструмента типологического исследования лексики, как кажется, весьма плодотворно и дает ряд преимуществ, в том числе, и по сравнению с описанными выше другими подходами к лексической типологии. Действительно, психолингвистический подход, основанный на предъявлении носителю внеязыковых стимулов с трудом применим для изучения лексических зон субъективных реакций, таких как болевые реакции или эмоции. В рамках школы А.Вежбицкой, составляющей определения для слов на семантическом метаязыке, должно быть затруднительно описывать квази-синонимы. Наконец, в обеих школах нельзя исследовать переносные значения слов – возможность, которая реализуется в рамках фреймового подхода и о которой будет сказано подробней ниже.

2.2 Методология исследования

2.2.1 Определение границ поля.

Любое лексико-типологическое исследование в рамках описываемого подхода начинается с первичного определения границ поля. Оно происходит на базе одного (как правило, родного исследователю) языка и состоит в выборе нескольких (квази-)синонимов, которые будут подробно изучаться.

Для выбранных слов составляется список контекстов, в которых они употребляются. Такой список удобнее всего составить при помощи корпуса - в случае русскоговорящего исследователя он обычно составляется на основе Национального корпуса русского языка (НКРЯ).

Структура контекстов разная в зависимости от части речи. Так, чтобы выявить основные свойства прилагательного, в качестве контекста достаточно взять существительное, к которому оно относится; для выявления семантических свойств глагола нужно собрать список существительных, которые выступают в роли его аргументов. С существительным вопрос о том, что именно брать в качестве контекста, решается по-разному в зависимости от его типа. Так, для отглагольных существительных релевантна аргументная структура (*отъезд*: [кого, куда]; *строительство*: [чего, кем]), для конкретных существительных часто важным оказываются зависимые предложные группы со значением места (*щель* **в полу**; *дупло* **в дереве**), не последнюю роль в определении семантики существительных играют и прилагательные, с которыми они сочетаются [Рахилина 2000/2008].

2.2.2 Типологическая анкета и челночный метод исследования лексического поля

Собранный список контекстов для изучаемых квазисинонимов служит основой типологической анкеты, в строках которой содержатся сами контексты, а в столбцах – информация о том, с какими из них сочетается каждая лексема.

Так, фрагмент основы для типологической анкеты, сформированной на базе русского языка для поля ‘ровные поверхности’ [Кашкин 2013] выглядит следующим образом:

	гладкий	скользкий	ровный
Пол		+	+
дорога	+	+	+
Мяч		+	
Щеки	+		
поле			+

Таблица 1

На такой таблице хорошо видны «запретные зоны» - контексты, в которых возможен один из синонимов и невозможен другой. Анализ таких зон уже на материале одного языка позволяет вычленить некоторые противопоставления, релевантные для поля. Так, на приведенном выше фрагменте анкеты видно, что для поля ‘ровных поверхностей’ оказывается релевантным параметр «зрительное» vs. «тактильное восприятие»: слово *ровный* сочетается только с теми предметами, гладкость которых может восприниматься зрительно. Кроме того, объекты, гладкость которых можно оценивать тактильно, подразделяются на ‘собственно гладкие’ и ‘скользкие’. Это противопоставление можно считать подтипом распространенного в признаковых полях параметра ‘функция’ vs. ‘форма’: русское прилагательное *гладкий* описывает лишь внешний вид поверхности предмета, тогда как *скольз-*

кие объекты можно случайно выпустить из рук (*скользкая рыба*), а на *скользких* поверхностях можно поскользнуться и упасть (*скользкая дорога*).

Заметим, что выделению параметров и закреплению прилагательных за отдельными их значениями абсолютно не противоречит отсутствие дополнительной дистрибуции в их сочетаемости. Поверхность многих предметов может описываться с разных сторон. Так, гладкость пола может оцениваться зрительно (1) или функционально (2). В первом случае будет употреблена лексема *ровный*, а во втором – *скользкий*.

(1) Комната, где читал Триродов, его кабинет, была большая, светлая и простая, с белым, некрашеным зеркально-**ровным полом**. [Ф. К. Сологуб. Капли крови (Навьи чары) (1905)]

(2) Директор перебирал ногами, стараясь встать, но подметки скользили по **скользкому полу**. [Андрей Троицкий. Удар из прошлого (2000)]

После составления первоначального варианта анкеты и выявления первых закономерностей лексического поля на материале одного языка начинается следующий этап исследования – привлечение материалов других языков. При уходе в типологическую зону значительно увеличивается типологическая анкета и зачастую расширяются границы поля: возникает необходимость принять во внимание новые, до этого момента не анализируемые, слова родного языка.

Привлечение материала других языков заключается в поиске переводных эквивалентов проанализированных на данный момент русских слов и составлении списков контекстов для них. Такой поиск удобно проводить при помощи словарей и корпусов (особенно параллельных). Безусловно, все данные надо проверять на носителях, так как информация в словарях может быть устаревшей, а в корпусах могут встречаться окказиональные употребления. Однако помощь, которую оказывают эти инструменты на начальном этапе привлечения к исследованию новых языков, очень значительна. Поэтому первыми в исследование обычно добавляются языки с богатой лексикографической традицией и обширными базами текстов. Анализ материала языков без корпусов удобнее проводить позже, при уже сформированной анкете.

При перечислении контекстов переводных эквивалентов русских лексем оказывается, что этот список в большей или меньшей степени пересекается со сформированной на данный момент анкетой: точно так же, как и квази-синонимы, переводные эквиваленты, частично совпадают в сочетаемости с русскими лексемами, а частично нет. Это несовпадение может проявляться в двух вариантах.

В первом случае, сочетаемость переводного эквивалента шире русской лексемы. В этом случае анкета дополняется новыми контекстами и приобретает статус типологической. Примечательно, что, судя по опыту уже осуществленных проектов, такое пополнение анкеты не будет вечным: с каждым новым языком строчек будет добавляться все меньше, а уже после десяти-двенадцати языков можно предполагать, что все ключевые, ядерные контексты поля в анкете уже представлены.

Более широкая сочетаемость переводного эквивалента ведет и к другому важному следствию. Часто она является причиной пересмотра первоначального списка прилагательных в русском языке. Так, при исследовании признакового поля ‘острый’ первоначальный материал русского языка включал только одно это прилагательное. Сочетаемость его сербского когната *оштар*, однако, оказалась более широкой: в отличие от русского слова, *оштар* может описывать ‘колючую бороду’, ‘колючую шерсть’, ‘колючий куст’. Это его свойство послужило причиной включения в список русских (квази-)синонимов прилагательное *колючий*. Этот метод, при котором данные новых языков иногда заставляют возвращаться к уже исследованным, мы называем *челночным* и считаем очень продуктивным для полноценного исследования лексического поля.

Другой вариант несовпадения – более узкая сочетаемость переводного эквивалента по сравнению с русской лексемой. В таком случае, русскому слову часто соответствует несколько слов в другом языке, каждое из которых покрывает только часть его контекстов. Так, русскому *толстый* соответствуют три китайских прилагательных – *hou* для плоских предметов, имеющих большое расстояние между верхней и нижней плоскостями (‘слоев’), *cu* для вытянутых предметов имеющих большую площадь сечения (‘веревок’) и *pang* для людей, имеющих избыточный вес. Такие ситуации крайне важны для выявления новых параметров внутри поля. Благодаря данным китайского языка, например, в поле ‘толстый’ стало возможным выделение противопоставления по топологическому типу объектов (‘слои’ vs. ‘веревки’) и отнесения ‘людей’ к отдельной зоне.

2.2.3 Фреймы и семантические карты

В процессе добавления в исследование новых лексем в определенный момент становится видно, что описанное выше «дробление» анкеты на кусочки отдельными словами может происходить в строго определенных местах. В признаковом поле ‘толстый’, например, двумя разными лексемами могут описываться ‘толстые слои’ и ‘толстые веревки’, но не ‘толстые стволы деревьев’ и ‘толстые палки’.

В анкете выделяются строки, которые всегда объединяются одним словом. Эти строки, как правило, представляют собой определенный гештальт, некую ситуацию, которая объ-

единяет в себе конкретные значения выделенных в поле параметров. Такие ситуации мы называем фреймами, и они составляют основные элементы семантической репрезентации поля. Фреймы не могут члениться на более мелкие куски, но могут объединяться лексемами в группы.

Эти объединения происходят не случайным образом – неверно, что любые фреймы могут объединяться с любыми. Одни из них тяготеют к объединению в лексеме с одними и, наоборот, скорее разделены разными словами с другими. Тенденцию к совместной встречаемости фреймов мы считаем свидетельством их когнитивной близости. Эту близость, вслед за грамматическими типологами, мы отражаем на семантических картах. Узлами на таких картах являются фреймы, а их геометрическое расположение значимо: одной лексемой могут описываться только смежные узлы. В отличие от многих грамматических карт, лексические не отражают диахронические процессы, которые могут происходить в словах: расположение узлов слева направо не говорит о том, что левое значение в слове развилось раньше правого. Отражая синхронный срез лексики, семантические карты говорят только о возможных и невозможных объединениях узлов.

Пример простой семантической карты представлен у признакового поля ‘острый’ [Кюсева 2012]. В этом поле выделяются два противопоставления. Первое, ‘линия’ vs. ‘точка’, отделяет острые инструменты с режущим краем (‘острый нож/сабля/коса’) от острых предметов с сужающимся концом (‘острая игла/спица/стрела’, ‘острый нос/колпак/мачта’). Второе, ‘функция’ vs. ‘форма’, различает острые инструменты (‘острый нож/сабля/коса’, ‘острая игла/спица/стрела’) и предметы с острой формой (‘острый нос/колпак/мачта’). Элементами семантической карты признака являются три фрейма, каждый из которых представляет собой определенную реализацию двух описанных выше противопоставлений: ‘острые режущие инструменты: ножи, сабли, косы’ (‘линия’&‘функция’), ‘острые колющие инструменты: иглы, стрелы, копья’ (‘точка’&‘функция’), ‘предметы с острой формой: носы, колпаки, мачты’ (‘точка’&‘форма’).

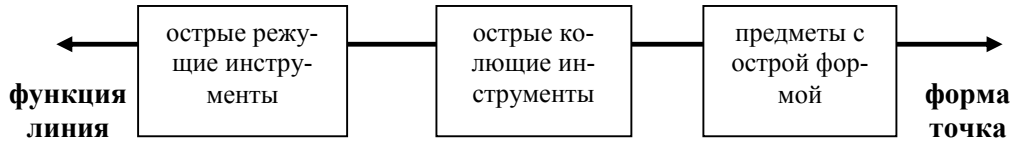

рис.1

Расположение фреймов на карте показывает их близость: одной лексемой могут описываться только смежные узлы. А значит, эта карта предсказывает невозможность существования прилагательного, которое бы описывало острые режущие инструменты и инструменты с острой формой, но не описывало бы острые колющие инструменты.

Заметим, что четвертое логически возможное объединение значений параметров признака ‘острый’ – ‘линия’&‘форма’ – в поле не присутствует: предметов с такой топологией крайне мало (острая крыша длинного дома, острый край стола) и языки никогда не выделяют их отдельной лексемой. Отсутствие в поле этого объединения иллюстрирует еще один важный принцип фреймового подхода лексической типологии, который, например, отличает ее от компонентного анализа: значения параметров в поле не независимы, а образуют устойчивые сочетания.

2.2.4 Исследование метафорических значений

Как уже было отмечено в начале главы, сочетаемостный метод исследования лексики позволяет изучать не только прямые, но и переносные значения лексем: для каждого слова поля анализируются его метафорические значения. Оказывается, что круг метафор, которые могут описывать лексемы одного поля, потенциально ограничен: несмотря на то, что никогда нельзя быть уверенным, что в каком-то языке у слова поля не возникнет незарегистрированное ранее, лингвоспецифичное переносное значение, в целом, переводные эквиваленты демонстрируют сходные переносы, а значит, можно очертить круг ядерных метафор поля, которые будут повторяться во многих языках.

Так, например, слова с прямым значением ‘острый’ в разных языках часто реализуют метафорические переносы на значения ‘хорошо функционирующий ум’ (итал. a*guzzo, acutus*, кит. *jiānrui*), сильная боль (англ. *sharp*, нем. *scharf*, яп. *surudoi*, рус. *острый*), ‘резкий ветер’ (фр. *tranchant*, коми *лэчыд*, итал. *tagliente*) и другие.

Более того, подробный анализ сочетаемости лексем свидетельствует в пользу мотивированности метафор: разные переносные значения связаны с разными фреймами прямых. На практике это проявляется в том, что для каждого переносного значения можно определить один или несколько фреймов прямых значений, которые лексемой, которая это переносное значение описывает, обязательно должны покрываться.

Так, например, переносное значение ‘четкая линия/граница’ может возникнуть только у тех лексем поля ‘острый’, которые описывают в своих прямых употреблениях фрейм ‘хорошо режущие инструменты’, а значение ‘глубокий человек’ – только у тех прилагательных, которые покрывают фрейм ‘хорошо колющие инструменты’.

Обратим внимание, что описываемая связь между двумя типами значений несимметричная. Если данная лексема описывает то или иное переносное значение, то она обязательно описывает и фрейм прямого, от которого это переносное производно. Но неверно, что если лексема описывает тот или иной прямой фрейм, то она обязательно описывает связанные с ним переносные. Мы лишь можем утверждать, что она потенциально может описывать переносные, которые от него производны и не может описывать тех, которые от него не производны.

2.3 Существующие проекты

В рамках описываемого подхода проводятся многочисленные исследования глагольной и признаковой лексики. Наиболее масштабный к настоящему моменту проект - это исследование глаголов движения в воде, в выборку которого вошло пятьдесят языков [Майсак, Рахилина 2007]. Чуть меньшая выборка, в тридцать языков, представлена в работе по изучению глаголов вращения [Круглякова 2010]. Два других глагольных проекта с уже опубликованными результатами – исследование предикатов боли [Брицын и др. 2009] и глаголов качания [Шапиро 2013].

Исследования признаковых полей представлены работами, посвященными прилагательным со значением ‘острый’ vs. ‘тупой’ [Кюсева 2012], ‘ровных’ vs. ‘неровных поверхностей’ [Кашкин 2013], ‘мягкий’ vs. ‘твердый’ [Павлова 2014], ‘легкий’ vs. ‘тяжелый’ [Кюсева и др. 2012], ‘полный’ vs. ‘пустой’ [Тагабилева, Холкина 2010], ‘прямой’ [Лучина и др. 2013] и некоторыми другими.

Наконец, участниками Московской лексико-типологической группы ведутся исследования, посвященные не конкретному полю, а конкретному языку, а точнее, анализу вклада материала одного конкретного языка в лексическую типологию. Такое исследование, посвященное китайскому языку, представлено в [Холкина 2013].

2.4 Типологическая база данных адъективной лексики

Одним из практических приложений, родившихся в рамках MLexT, является Типологическая база данных адъективной лексики (см. [Резникова и др. 2013]. В ней представлены собранные в рамках разнообразных проектов материалы по изучению признаковых слов.

2.4.1 Структура Базы

Вход в Базу – это строка типологической анкеты поля. Для нее указано прилагательное, которое покрывает этот контекст, язык, из которого взято данное прилагательное, фрейм, который контекст реализует в сочетании с данным существительным, а также при-

знаковое поле и таксономический класс прилагательного в этом употреблении. Для каждого контекста указывается также, реализует ли он прямое или переносное значение прилагательного.

Так, несколько строк из Базы для признакового поля ‘прямой’ выглядят следующим образом:

Контекст	Лексема	язык	Фрейм	поле	такс.класс	тип значения
прямая палка	прямой	русский	незакрепленный стержень без изгибов	прямой	физ.свойства: форма	исходное
прямая палка	suora	финский	незакрепленный стержень без изгибов	прямой	физ.свойства: форма	исходное
отвесный склон	suors	финский	отвесный склон	прямой	физ.свойства: форма	исходное
отвесный склон	отвесный	русский	отвесный склон	прямой	физ.свойства: форма	исходное
прямой начальник	прямой	русский	без посредников	Прямой	-	Переносное

Таблица 2

Этот фрагмент, в котором приведены данные из разных языков, иллюстрирует сразу несколько важных особенностей структуры Базы.

Во-первых, в качестве контекста приводится не просто существительное из типологической анкеты, а сочетание существительного с прилагательным. Русский язык в этом столбце служит метаязыком, а сами контексты – «микрофреймами»: минимальными ситуациями, которые затем объединяются во фреймы.

Ср, например, несколько фреймов поля ‘прямой’ с соответствующими им микрофреймами, представленными в типологической анкете Базы:

Микрофрейм	фрейм
прямая палка	незакрепленный стержень без изгибов
прямой гвоздь	
прямой столб	вертикально закрепленный объект без изгибов
прямой забор	
отвесный склон	отвесный склон
отвесный берег	

Таблица 3

Во-вторых, фреймы в базе приписываются микрофреймам не только прямых, но и переносных значений, которые, в свою очередь, имеют несколько иную природу.

Для того чтобы возникла возможность постулировать тот или иной фрейм прямого значения, в поле должно встретиться слово, которое в своих исходных употреблениях покрывает только его. Так, например, стимулом к выделению трех фреймов прямых употреблений в поле ‘острый’ – хорошо режущие инструменты, хорошо колющие инструменты и инструменты с сужающимся кончиком – был материал французского языка, в котором каждый из этих фреймов покрывается отдельной лексемой (*tranchant, aigu* и *pointu*, соответственно). Такая же стратегия по отношению к метафорам кажется довольно затруднительной хотя бы потому, что ни одно переносное значение у слов изучаемого поля не реализуется в одиночку, а совмещается, по крайней мере, с прямым. Поэтому по отношению к таким типам значения была принята несколько смягченная стратегия выделения фреймов. Строки анкеты, выражающие одно переносное значение, объединяются в фрейм в том случае, если в поле находятся прилагательные, которые на них «реагируют» поразному, т.е. одни лексемы эти строки покрывают (причем, покрытие одной строки фрейма влечет за собой покрытие и остальных), а другие – нет. С фреймами прямых значений фреймы переносных объединяет то, что они также описывают некоторую типичную ситуацию. Ср., например, фрагмент типологической анкеты поля ‘мягкий’, иллюстрирующий объединение переносных микрофреймов во фреймы:

Микрофрейм	фрейм
мягкий взгляд	добрый, податливый человек
мягкое сердце	
мягкий человек	
обессилевшие ноги	ослабевший человек
обессилевшее тело	
мягкий свет	несильное воздействие на органы чувств
мягкий звук	

Таблица 4

Наконец, последняя особенность устройства Базы, о которой надо упомянуть, касается ввода в нее отрицательной информации.

Если бы при изучении признакового поля для каждой новой лексемы всегда была бы возможность заполнить всю типологическую анкету, отсутствие в Базе строки, в которой конкретное прилагательное сочетается с конкретным контекстом, однозначно бы говорило о том, что такое сочетание невозможно. Однако на практике оказывается, что заполнение всей анкеты для прилагательных некоторых языков оказывается крайне затруднительным. Несмотря на то, что при составлении типологической анкеты исследователь ста-

рается не включать в нее лингвоспецифичные контексты, иногда оказывается, что у народов некоторых языков просто нет реалий, которые от них требуют определенные строки. Так, например, в типологическую анкету поля ‘острый’ входят примеры колющего и режущего оружия – копья, стрелы, мечи, сабли – предметы, с которыми, например, не знаком народ коми, и которые, поэтому не были заполнены для лексем этого языка. Часто осложняет заполнение всей анкеты и отсутствие у языка корпусов и обширных словарей.

Для того, чтобы такие ситуации «незаполненности» не смешивались с ситуациями запрета лексемы на покрытие данного контекста, было решено вводить в базу и отрицательную информацию. Для каждого прилагательного, которое вводится в Базу, указываются не только те контексты, которые оно покрывает, но и те, которые оно не покрывает. Отсутствие же строки, в которой конкретная лексема покрывает конкретный контекст, говорит лишь о том, что о грамматичности данного сочетания нет ни положительной, ни отрицательной информации.

Для того чтобы не потерять возможность сопоставления лексем и чтобы не получилось так, что для разных слов заполнены разные куски анкеты, часть ее строк (примерно по две для каждого фрейма), наименее культуроспецифичных, имеет статус «обязательных». Для каждого слова поля должно быть заполнено, покрывает ли оно каждый из «обязательных» микрофреймов или нет. «Необязательные» же микрофреймы заполняются по мере наличия данных.

Фрагмент Базы для эстонского слова *kortsus* – ‘морщинистый’ – иллюстрирует заполнение в ней отрицательной информации о сочетаемости лексемы.

контекст	лексема	язык	фрейм	поле	такс.класс	тип значения
морщинистая кожа	*kortsus*	эстонский	морщинистые поверхности	шершавый	физическ. свойства	исходное
морщинистое лицо	*kortsus*	эстонский	морщинистые поверхности	шершавый	физическ. свойства	исходное
смор-щенная картошка	- *kortsus*	эстонский	сморщенные объемные объекты	шершавый	физическ. свойства	исходное
сморщенное яблоко	- *kortsus*	эстонский	сморщенные объемные объекты	шершавый	физическ. свойства	исходное

Таблица 5

Деления на обязательные vs. необязательные фреймы здесь не видно: эта информация, необходимая только для заполняющих Базу, не сохраняется в ее итоговом варианте.

2.4.2 Поисковые запросы

Особенности структуры Базы данных позволяют извлекать из нее данные различных типов.

Два самых простых запроса, которые может задать пользователь базы, - это поиск по лексеме и по фрейму. В первом случае поиск по столбцу, в котором указывается конкретная лемма, позволяет получить полный список фреймов (как прямых, так и переносных значений), покрываемых данной лексемой. В этой ситуации База данных выступает в роли качественного типологически ориентированного одноязычного словаря, в котором представлены все значения прилагательного и особенности его сочетаемости. Второй тип запросов, также как и первый, подразумевающий поиск по одному столбцу таблицы (на этот раз, «фрейму»), позволяет Базе выступать в роли многоязычного словаря нового типа. Для определенного фрейма при этом запросе указываются все прилагательные из разных языков, которые его покрывают.

Полезными для поиска оказываются и более сложные, комбинированные запросы. Так, комбинация двух описанных выше запросов позволяет искать случаи совмещения фреймов, т.е. информацию о том, с какими фреймами в одной лексеме может совмещаться данный фрейм и насколько такое объединение частотно (= типологически релевантно). Запросы такого вида позволяют, например, определить, различаются ли в данном языке в поле ‘тонкий’ класс вытянутых цилиндрических и класс вытянутых плоских объектов.

Наконец, наличие в Базе информации о переносных значениях лексем позволяет прослеживать закономерности в области метафорических переносов. Классификация каждого входа по типу значения (прямое/производное) позволяет искать информацию о том, какие исходные фреймы есть у слова с данным переносным.

Хранение информации об исследованных полях в формате базы данных обеспечивает ее структурированность. Это позволяет, во-первых, сравнивать данные из разных языков и их разных признаковых полей, а во-вторых, проводить статистические подсчеты разных типов. Последнее качество окажется ключевым для возможности провести описанный в настоящей работе эксперимент.

3. Модели дистрибутивной семантики

3.1 Основные понятия

Лексическая типология, хоть и опирается на опыт типологии грамматической, довольно сильно от нее отличается. Так, в грамматическом поле значений, как правило, мало: три лица, два-три числа и т.д. Они хорошо структурированы и всегда непосредственно наблюдаемы в том или ином показателе. В лексическом поле значений обычно значительно больше, и они не имеют никаких специальных формальных показателей, а значит обнаружить их в языке намного труднее (см. Рахилина, Плунгян 2007). Отсюда следует, что для проведения качественного лексико-типологического исследования требуется обработать значительно больший корпус текстов: если для изучения грамматической категории в общем случае достаточно размеченного корпуса в 200-300 тысяч словоупотреблений, то для исследования лексического поля нужен корпус порядка 100 млн словоупотреблений и выше. Анализ корпусных данных такого объема вручную очень трудоемок.

Поэтому с ростом компьютерных технологий стали появляться модели, предлагающие, по крайней мере, частичную автоматизацию процесса исследования лексики. Наиболее распространенными из них в настоящее время являются модели дистрибутивной семантики (Distributional Semantics Models). Разработанные для анализа лексики одного языка, они предлагают автоматический способ извлечения значения слова (см. подробное описание моделей в [Baroni et al. 2013]).

Модели дистрибутивной семантики исходят из дистрибутивной гипотезы, которая гласит, что слова, появляющиеся в похожих контекстах, имеют схожее значение. Эта идея восходит ко многочисленным лингвистическим традициям, включая американскую структурную лингвистику, английскую лексикографию и русскую школу изучения семантики лексики. Она оказала большое влияние на компьютерную лингвистику, так как предложила способ автоматически извлечь семантику слова: если значение лексемы аппроксимируется контекстами, в которых она возникает, то для того, чтобы его зафиксировать, надо просто записать все эти контексты.

В моделях дистрибутивной семантики каждое слово представлено математическим многомерным вектором, то есть упорядоченным списком цифр. Значения этого вектора – функция от количества раз, которое слово встретилось в определенных контекстах в корпусе.

Для большей наглядности приведем небольшой пример. Допустим, мы хотим получить семантические репрезентации слов *банан*, *яблоко* и *каша* на основе их встречаемости

в контексте слов *варить* и *мыть*. Пускай, в нашем корпусе слово *банан* встретилось 1 раз в контексте слова *варить* и 5 раз – *мыть*, *яблоко* – 0 раз в контексте слова *варить* и 7 – *мыть*, а каша – 8 раз в контексте слова *варить* и 3 раза – *мыть*.

Эти данные можно представить в виде таблицы:

	варить	Мыть
Банан	1	5
Яблоко	0	7
Каша	8	3

Таблица 6

Строки этой таблицы – слова, для которых мы собираем семантическую репрезентацию, столбцы – измерения векторов. Вектор каждого слова, таким образом, состоит из двух координат –«мыть» и «варить» – и равен для лексемы *банан* – (1,5), *яблоко* – (0,7), *каша* – (8,3).

Уже на таблице интуитивно видно, что слово *банан* ближе к *яблоку*, чем к *каше*: и *банан*, и *яблоко* часто встречаются в контексте глагола *мыть* и редко – *варить*, а *каша*, наоборот, часто встречается в контексте слова *варить* и редко – *мыть*. Вектора сочетаемости позволяют математически точно оценить эту близость.

На рис. 2 вектора слов *банан*, *яблоко* и *каша* изображены в виде ориентированных отрезков на плоскости:

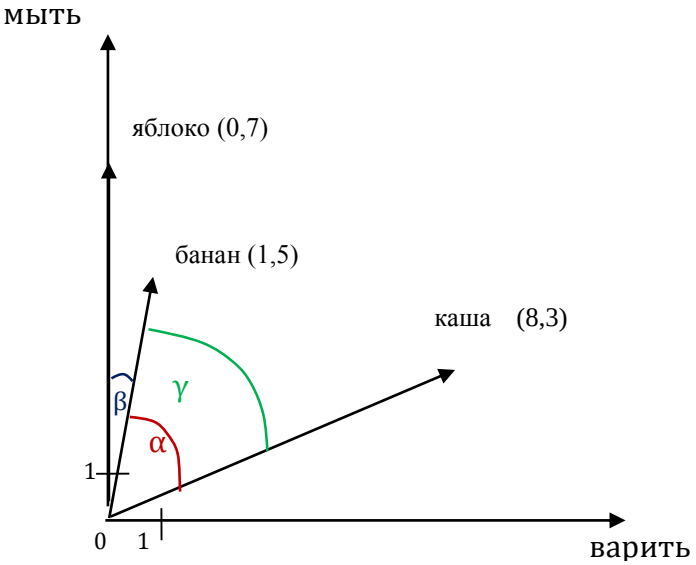

рис. 2

Измерение косинуса угла между каждыми двумя векторами позволит точно оценить их семантическую близость. Чем больше косинус угла, тем ближе вектора (косинус равен 1 у параллельных векторов и 0 у перпендикулярных). Так, на рис.2 $\cos \alpha = 0.53$, $\cos \beta = 0.98$, $\cos \gamma = 0.35$.

В приведенном примере размерность векторов равна 2. Для того, чтобы полностью описать семантику слова, такого размера вектора, конечно, не достаточно. Так, если мы просто поменяем названия координат на «есть» и «любить», различия между тремя словами, так ярко видные при координатах «варить» и «мыть», скорее всего, нивелируются. Поэтому в реальных ds-моделях, во-первых, используются вектора размерностью в несколько тысяч, а, во-вторых, большое значение уделяется выбору координат.

3.2 Параметры модели дистрибутивной семантики

Перечислим основные параметры моделей дистрибутивной семантики.

3.2.1 Корпус

Важный параметр любой модели векторной семантики – это размер и качество корпуса. Если корпус недостаточно велик или если он ориентирован на определенную тематику или жанр, то высока вероятность, что вектора получатся некачественные - они будут представлять семантику слов в искаженном виде. Поэтому в ds-моделях обычно используются сбалансированные (национальные) корпуса от 100млн словоупотреблений.

3.2.2 Контекст

Контекст – это один из ключевых параметров моделей векторной семантики. Понятие контекста определяют по-разному в зависимости от поставленных целей. Так, принятие в качестве контекста появление двух слов в одном документе позволяет поймать «тематические» (topical) отношения. Например, слова *Путин* и *Крым*, в таком случае, будут иметь высокую степень близости, так как они часто встречаются в одних и тех же документах. В случае стремления поймать «таксономические» отношения (когипонимы) чаще всего используется одно из двух понятий контекста. При первом в качестве контекста понимаются (значимые) слова, которые встретились в фиксированном окне относительно целевого слова. Размер окна тоже имеет большое значение. Так, если собрать вектор для прилагательного с узким окном (+/- 1 = одно слово слева и одно справа), то среди самых близких соседей, с большой вероятностью, окажется его антоним. Расширение окна уменьшает степень близости прилагательного с антонимом и опускает его в списке соседних векторов. Например, при сборе векторов на базе основного подкорпуса Национального корпуса русского языка (220 млн словоупотреблений) в окне +/- 1 прилагательное *тупой* оказывается первым в списке соседей слова *острый*, а в окне +/-5 – пятым. Второе понятие кон-

текста, которое часто используют при поиске «таксономической» близости слов – лексемы, находящиеся в определенных синтаксических отношениях к целевому слову. Например, при поиске синонимов или когипонимов к многоместным глаголам часто оказывается полезно в качестве контекста принимать его аргументы.

3.2.3 Трансформация вектора

Как правило, вектора, измерения которых представляют количество раз, которое целевое слово встретилось в разных контекстах, никогда не используются в моделях дистрибутивной семантики в чистом виде. Минимальная обработка, которую они обязательно проходят, - трансформация измерений вектора в ассоциативные коэффициенты (association scores), при которой веса частотных слов-контекстов, понижаются, а редких – повышаются. Так, встречаемость практически любой лексемы в контексте глагола *быть* или *мочь* значительно менее информативна, чем в контексте *мыть* и *варить*. Трансформация координат векторов в ассоциативные коэффициенты, которая обычно осуществляется при помощи таких мер, как Pointwise Mutual Information или Log-Likelihood Ratio, позволяет это учесть.

3.2.4 Уменьшение размерности вектора

Иногда, у векторов, преобразованных в ассоциативные коэффициенты, также уменьшают размерность: похожие координаты вектора объединяются в кластеры, а затем представляются в виде одной усредненной координаты. Например, в одну общую координату могут объединиться такие контексты, как «любить», «обожать», «одобрять». Результирующая размерность может быть во много раз меньше исходной (нередко ее уменьшают с 10тыс, скажем, до 300) и удобна, по крайней мере, в двух аспектах. Во-первых, в векторе с уменьшенной размерностью всегда меньше нулей: если целевое слово легко может ни разу не встретиться в контексте какого-то одного слова-координаты, то ситуации, при которых оно ни разу не встретилось в целом кластере координат, случаются значительно реже. Во-вторых, такое несколько «огрубленное» представление вектора позволяет обнаружить более общие закономерности его семантики, что часто очень полезно для исследования, и, например, практически необходимо для качественной кластеризации векторов.

3.2.5 Композиция векторов

Формирование вектора лексической единицы при помощи прямого подсчета контекстов, в которых она возникает (с возможным последующим взвешиванием и уменьшением размерности) подходит для репрезентации семантики отдельных (значимых) слов. Для того, чтобы представить значение единицы, большей, чем слово (словосочетания, глагола с

аргументами или даже предложения), этот способ, скорее всего, не подходит: в случае его применения понадобился бы огромный, если не бесконечный, корпус.

Поэтому, для представления значения таких единиц были разработаны различные методики композиции векторов – получения вектора большей единицы из векторов меньших единиц, входящих в нее.

Наиболее распространены в настоящее время две модели композиции векторов, предложенные в работе [Mitchell & Lopata 2010], - аддитивная и мультипликативная. Согласно аддитивной модели, вектор составной единицы получается сложением векторов входящих в нее простых, а согласно мультипликативной – умножением (см. таблицу 7).

		варить	мыть
	яблоко	0	7
	каша	8	3
	вкусный	2	0
аддитивная модель	яблоко + вкусный	2	7
	каша + вкусный	10	3
мультипликативная модель	яблоко ⊙ вкусный	0	0
	каша ⊙ вкусный	16	0

Таблица 7

При аддитивной модели композиции вектор словосочетания кумулятивно «наследует» значения измерений входящих в него слов: если какая-то координата у одного из слов нулевая, то ее значение у результирующего вектора будет равно не нулю, а значению координаты вектора второго слова (так, например, координата «мыть» у вектора ‘яблоко+вкусный’ приобретает достаточно высокое значение – «7» – наследуя его от вектора лексемы *яблоко*). Те же координаты, которые у обоих слов больше нуля, будут получать в итоговом векторе достаточно высокие значения. Такая модель увеличивает измерения, важные для обоих слов и оставляет без изменения измерения, важные только для одного.

Мультипликативная модель приписывает нулевые значения всем координатам, которые хотя бы у одного вектора равны нулю и очень сильно увеличивает значения тех измерений, которые больше нуля у обоих слов.

В обеих описанных моделях вектора составляющих слов абсолютно симметричны: они вносят равноправный вклад в репрезентацию результирующей конструкции. Однако кажется вполне интуитивным, что разные части словосочетания вносят разный по вели-

чине вклад в его семантику. Для того чтобы это учесть, были разработаны взвешенные модели композиции. Так, взвешенная аддитивная модель (weighted additive model) перед складыванием двух векторов приписывает каждому из них коэффициент, на который они домножаются. В случае словосочетания прилагательного и существительного, например, больший коэффициент приписывается существительному, чем прилагательному, что отражает более весомый его вклад в семантику словосочетания. Например, домножение нашего вектора слова *вкусный* на коэффициент 0.2, а *яблоко* – 0.8 приведет к вектору словосочетания *вкусное яблоко*, равному (2*0.2 + 0*0.8, 0*0.2 +7*0.8) = (0.4, 5.6) и покажет, например, что *вкусное яблоко* крайне редко возникает в сочетании с глаголом «варить» (первое измерение вектора), более наглядно, чем вектор, полученный простым сложением слов.

Некоторым расширением взвешенной аддитивной модели композиции можно считать модель гомотетии (dilation model), согласно которой вектор словосочетания получается по формуле:

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

где $\vec{p}$ – это результирующий вектор, $\vec{u}$ и $\vec{v}$ – вектора составляющих, а λ – скалярная величина.

Как показано в [Mitchell & Lopata 2010], эти модели очень хорошо себя проявляют в композиции сочетаний прилагательного и существительного, существительного и существительного и глагола-существительного, качественно предсказывая человеческие суждения об их близости. Однако, они не приспособлены для того, чтобы учитывать различие в семантике, вызванное разным порядком слов, а также представлять значение словосочетания, состоящего из знаменательного и функционального слова (артикля, предлога, союза).

Для того чтобы решать эти задачи, разработана модель композиции с дистрибутивными функциями [Baroni et al. 2013]. Согласно этой модели, значение некоторых слов кодируется не в векторах, а в дистрибутивных функциях, которые берут один вектор на входе и возвращают другой, трансформированный, на выходе. Семантика существительных и именных групп при таком подходе представлена в виде векторов, а значение прилагательных, глаголов, артиклей, некоторых местоимений, предлогов и союзов – в виде дистрибутивных функций.

Приведем рисунок, демонстрирующий разницу между этим и аддитивным подходами из [Baroni et al 2013]:

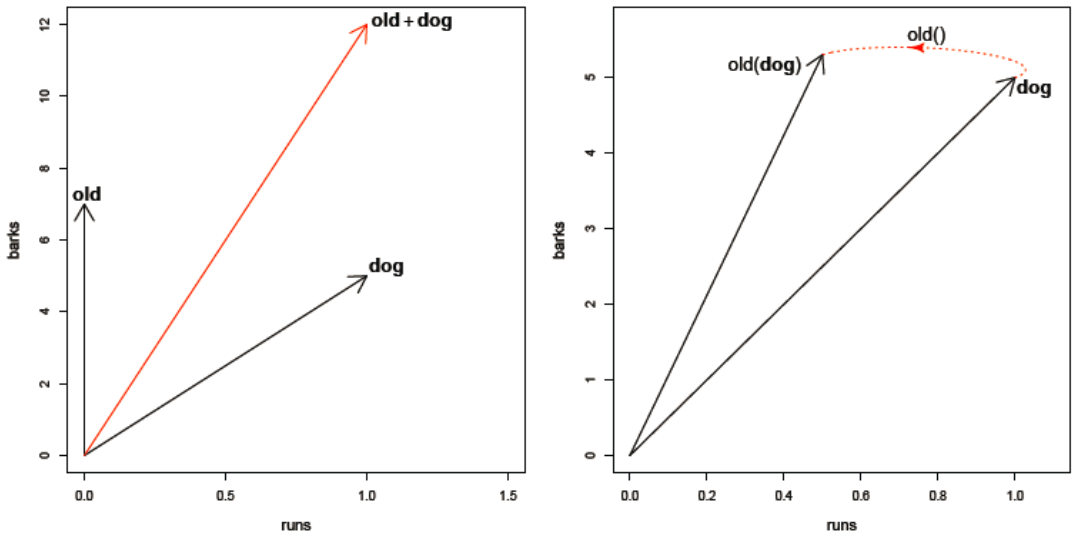

рис.3

Аддитивная модель берет на входе два вектора и складывает попарно их компоненты, чтобы получить вектор словосочетания. При функциональной модели одно из слов представлено не в виде вектора, а в виде функции, которая воздействует на вектор-аргумент и перемещает его в другое место в семантическом пространстве.

Такой подход позволяет оперировать семантикой функциональных слов и учитывать разные значения, возникающие в следствие структурного различия лексических конструкций, таких как *компьютер со столом* vs. *компьютер на столе*.

3.2.6 Мера близости векторов

В большинстве моделей дистрибутивной семантики степень близости векторов вычисляется через косинус угла между ними. В качестве альтернативной меры используют Евклидово расстояние между векторами, которое представляет собой длину отрезка, соединяющего их конечные точки. Евклидово расстояние, однако, чувствительно к длине вектора, которая, в свою очередь, зависит от частотности слова. При сравнении между собой векторов частотного и нечастотного слова Евклидово расстояние будет велико вне зависимости от их семантической близости. Ср, например, разницу между рис. 4а и 4б:

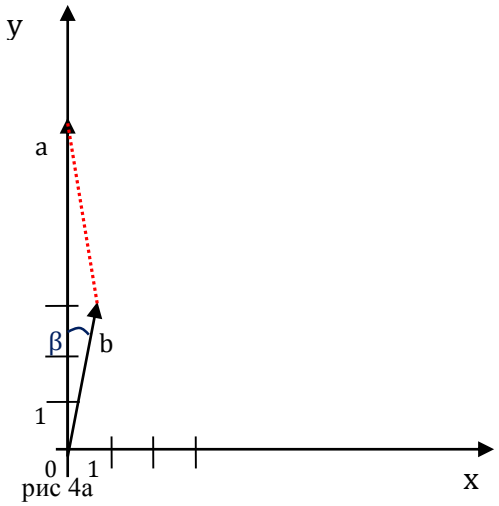

рис 4а

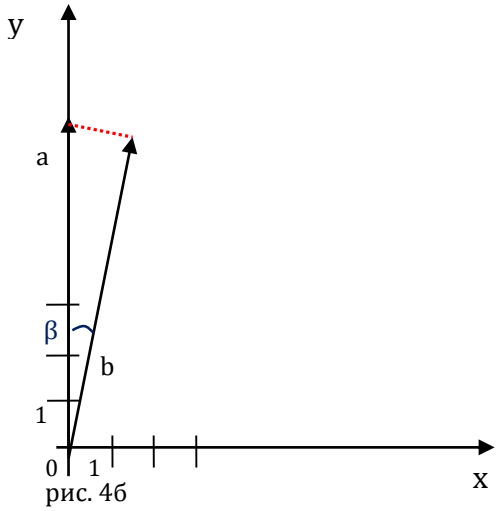

рис. 4б

На рисунке 4а вектор b значительно короче вектора a, на рисунке 4б оба вектора примерно одинаковой длины. Эта разница влечет за собой разные значения Евклидова расстояния при одном и том же угле между векторами.

Для того чтобы избежать такого рода погрешностей, перед вычислением Евклидова расстояния векторá всегда сначала нормализуют по длине.

3.3 Применение векторных моделей

Семантические репрезентации слов, предоставляемые векторными моделями, могут быть применены в самых разных приложениях компьютерной лингвистики, от поиска информации и классификации в вопросно-ответных системах до автоматического построения тезауруса и машинного перевода.

Они эффективно симулируют психологические и лингвистические феномены, связанные с семантическими отношениями между словами. Эксперименты показывают, что модели дистрибутивной семантики могут предсказывать таксономическую классификацию слов, семантический прайминг, суждения людей о семантической близости слов и тематической принадлежности аргументов глагола.

Так, начиная с классической работы [Landauer & Dumais 1997], исследователи показывали, что дистрибутивные модели позволяют выбрать синонимы для целевого слова из списка слов-кандидатов. В [Padó & Lapata 2007] продемонстрировано, что по косинусу угла между векторами двух слов можно определить, будут ли они друг друга «праймить» (будет ли второе слово распознаваться испытуемым быстрее, если первое ему уже предъявлено). Лили Котлерман [Kotlerman et al 2010] использует модели дистрибутивной семантики, для того чтобы предсказать, влечет ли концепт, выраженный в одной лексиче-

ской единице, за собой концепт, заключенный в другой (например, влечет ли понятие ‘собака’ за собой понятие ‘животное’). В [Padó et al 2007] показано, что косинус угла между вектором, представляющим собой «типичный» объект или субъект глагола, и вектором случайного существительного предсказывает суждения людей о возможности этого случайного существительного выступить в роли объекта или субъекта глагола.

В последнее время все большую силу приобретают модели векторной семантики, решающие проблему многозначности слова. Один из алгоритмов автоматического извлечения слов предложен в [Schütze 1998]. Согласно этому алгоритму, значения слова интерпретируются как группы похожих его употреблений. На тренировочном корпусе каждому употреблению слова приписывается так называемый «контекстный» вектор. Этот вектор состоит из суммы векторов, представляющих значения слов, которые встретились в контексте целевого слова в данном его употреблении. Затем все контекстные вектора кластеризуются. Каждый кластер принимается за отдельное значение слова, у которого вычисляется центроид. Этот центроид называется смысловым вектором, см. рис. 5 (из [Schütze 1998]).

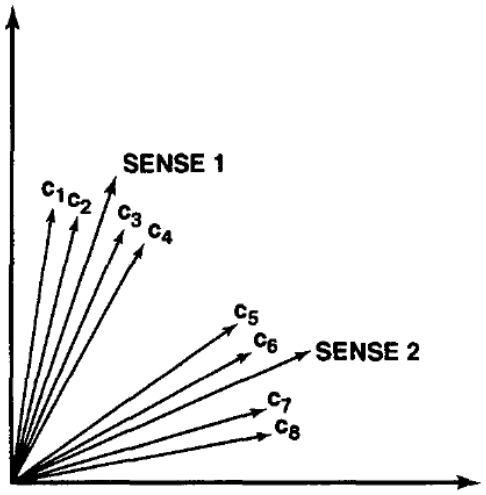

рис. 5

Употребление целевого слова в тестовом корпусе дизамбигуируется при помощи расчета для него контекстного вектора и отнесения этого вектора к кластеру с наиболее близким центроидом.

Итак, векторные модели уже доказали свою применимость для решения целого ряда задач лексической и не только лексической семантики. Однако, никогда раньше они не использовались для целей лексической типологии.

4. Векторные модели в лексической типологии

4.1 Цели эксперимента:

Эксперимент, проведенный в рамках настоящего исследования, преследовал две цели.

Первая цель – проверка гипотезы, которая вполне естественно рождается при принятии фреймового подхода к лексической типологии. Если фреймы – единицы смысла, которые универсальны для всех языков, то они должны проявляться в каждом конкретном языке и обнаруживать себя через сходство / различие в сочетаемости лексических единиц этого языка. Мы предполагаем, что лексические конструкции, которые принадлежат к одному фрейму, ближе друг к другу, чем те, которые принадлежат к разным фреймам.

Например, в поле ‘мягкий’, среди прочих, выделяются фреймы ‘мягкая пища’ (микрофреймы ‘мягкое мясо’, ‘мягкий хлеб’, ‘мягкая картошка’, ‘мягкий персик’) и ‘мягкие вещества’ (‘мягкая глина’, ‘мягкое тесто’, ‘мягкая земля’). Наша гипотеза состоит в том, что в каждом конкретном языке словосочетания, передающие значения, например, ‘мягкий хлеб’ и ‘мягкая картошка’, будут иметь более похожую сочетаемость, чем каждое из них со словосочетанием ‘мягкая земля’.

Вторая цель состоит в необходимости некоторой верификации фреймового подхода в лексической типологии. Презумпция, из которой исходит этот подход, о том, что в каждом лексическом поле можно выделить ограниченное количество фреймов и расположить их определенным образом на карте, так что эта карта будет предсказывать все возможные стратегии лексикализации поля, кажется довольно смелой и часто встречает критику со стороны противников лексической типологии. Действительно, как можно быть уверенным, что ни один новый язык не добавит нового узла на карту, не разделит до этого момента цельный фрейм или не организует поле по совершенно новым правилам? А если мы предполагаем такую возможность, то разве можно хоть в какой-то момент считать, что описание структуры данного лексического поля завершено? Осложняет ситуацию и тот факт, что выборки языков в лексической типологии в силу большой трудоемкости исследования обычно меньше тех, которые используются в типологии грамматической, они менее представительны и разнообразны, а значит – менее убедительны.

В таких условиях полезным оказывается привлечение материала экзотических языков: демонстрация того, что африканские, австронезийские или кавказские языки делят соответствующие лексические поля по тем же законам, что и более привычные нам романские или славянские, кажется весомым аргументом в пользу лексических типологов.

Однако в случае с экзотическими языками проверку фреймового подхода осуществляет сам исследователь. А настоящей, заслуживающей доверия верификацией, как кажется, следует считать некоторый автоматический, «свободный от исследователя» способ проверки.

Модели векторной семантики, основанные на статистической обработке большого массива текстов, представляются наиболее удобным способом как проверить высказанную гипотезу, так и послужить методом верификации фреймового подхода. Действительно, их исходное допущение о возможности определения значения слова/словосочетания через контекст, в котором оно употребляется, очень близко духу лексической типологии. Однако если в лексическо-типологическом исследовании используются данные *нескольких языков*, собранные *вручную* в *ограниченном объеме*, то в традиционных моделях векторной семантики материалы *одного языка* обрабатываются *автоматически* в *любом объеме*, который может обеспечить корпус текстов. Возможность моделей дистрибутивной семантики создавать математическую репрезентацию значений слов и вычислять их близость позволить достичь поставленных целей.

4.2 Данные эксперимента

К настоящему моменту эксперимент проведен на основе одного признакового поля – ‘острый’. По этому признаку База заполнена для 33 слов 15 языков (полный список лексем представлен в приложении 1). Для каждой лексемы заполнена анкета в 150 строк, из которых обязательных – 55. Строки анкеты разделены на 34 микрофрейма (см. анкету в приложении 2).

Эксперимент состоял в сравнении двух типов данных.

Первый тип данных основан на материалах Базы. Для каждых двух микрофреймов поля ‘острый’ вычисляется, насколько часто они покрываются в Базе одной лексемой. Формула такого вычисления проста: количество лексем, которые покрывают два данных микрофрейма складывается с количеством лексем, которые не покрывают ни один из них, и затем эта сумма делится на общее количество лексем, для которых оба микрофрейма заполнены. Математически ее можно выразить так:

$coeff_{a,b} = (nAdj_{[+a,+b]} + nAdj_{[-a,-b]}) / (nAdj_{[-a,+b]} + nAdj_{[+a,-b]} + nAdj_{[+a,+b]} + nAdj_{[-a,-b]})$,

где coeff – вычисляемая частота покрытия двух микрофреймов одной лексемой,

a и *b* – два данных микрофрейма ,

$Adj_{[+/-a,+/-b]}$ – это прилагательное с конкретной реализацией микрофреймов *a* и *b*,

а $nAdj_{[+/-a,+/-b]}$ - это количество таких прилагательных

Так, $nAdj_{[+a,+b]}$ – количество прилагательных в Базе, которые покрывают и микрофрейм *a*, и микрофрейм *b*, $nAdj_{[+a,-b]}$ – тех, которые покрывают микрофрейм *a*, но не *b*; $nAdj_{[-a,+b]}$ – *b*, но не *a*, а $nAdj_{[-a,-b]}$ – количество лексем, которые не описывают ни один из двух микрофреймов.

Нахождение в числителе дроби последней упомянутой составляющей –

$nAdj_{[-a,-b]}$ – объясняется тем, что случаи, в которых прилагательное не описывает ни один из двух данных контекстов для нас равноправны случаям, в которых оно описывает оба контекста. Действительно, мы хотим противопоставить ситуации объединения микрофреймов ситуациям их разделения между разными лексемами. Но непокрытие лексемой обоих микрофреймов – и есть частный случай объединения. Иными словами, когда мы хотим посчитать наш коэффициент частотности для микрофреймов ‘острый нож’ и ‘острый меч’ поля ‘острый’, для нас одинаковый статус имеет и французская лексема *tranchant*, которая описывает остроту обоих предметов, и *pointu*, которая не описывает ни один из них.

Для примера расчета формулы приведем фрагмент Базы:

контекст	Лексема	Язык	фрейм	Поле	такс.класс	тип значения
острая стрела	Острый	Русский	острый инструмент с колющим концом	острый	физ.свойства	исходное
острый нос	Острый	Русский	предмет, сужывающийся к концу	острый	физ.свойства: форма	исходное
острая стрела	Surudoi	Японский	острый инструмент с колющим концом	острый	физ.свойства	исходное
острый нос	-surudoi	Японский	предмет, сужывающийся к концу	острый	физ.свойства: форма	исходное
острая стрела	-лэчыд	Коми	острый инструмент с колющим концом	острый	физ.свойства	исходное
острый нос	-лэчыд	Коми	предмет, сужывающийся к концу	острый	физ.свойства: форма	исходное

Таблица 8

Если обозначить микрофрейм ‘острая стрела’ как *a*, а ‘острый нос’ как *b*, то частотность их совместной встречаемости равна:

$(nAdj_{[+a,+b]} + nAdj_{[-a,-b]}) / (nAdj_{[-a,+b]} + nAdj_{[+a,-b]} + nAdj_{[+a,+b]} + nAdj_{[-a,-b]}) =$

$=(1 + 1) / (0 + 1 + 1 + 1) = 2/3,$

где $Adj_{[+a,+b]}$ – прилагательное *острый*, $Adj_{[-a,-b]}$ – прилагательное *лэчыд*, а

$Adj_{[+a,-b]}$ – прилагательное *surudoi*.

Второй тип данных собирается на основе корпуса текстов. Для списка микрофреймов, составляющего типологическую анкету поля ‘острый’ (того же списка, который участвовал в сборе первого типа данных) по корпусу русского языка собираются вектора сочетаемости, а затем для каждой пары микрофреймов считается близость векторов, представляющих их значения.

В данном случае список микрофреймов выступает в роли заполненной русской анкеты. Это объясняется двумя фактами: во-первых, метаязык типологической анкеты русский, а во-вторых, в поле ‘острый’ в русском языке есть ровно одно доминантное прилагательное – *острый*. Благодаря этому все микрофреймы анкеты, которые его содержат (*острый нож, острый меч, острая пила*), представляют собой контексты, которое русская лексема *острый* покрывает. А все строки, которые его не содержат (*крутой склон, колючая щетина, резкий поворот* и другие) представляют собой контексты, которые прилагательное *острый* не покрывает. В этом случае нет проблемы отрицательной информации и обязательных/необязательных микрофреймов: анкета заполнена для всех микрофреймов, а в роли отрицательной информации выступают те словосочетания, в которых *острый* не присутствует.

Для большей наглядности приведем фрагмент типологической анкеты, заполненной для русского языка.

контекст	Лексема	язык	фрейм	поле	такс.класс	тип значения
острый нож	Острый	Русский	острый инструмент с режущим краем	острый	физ.свойства	исходное
острый меч	Острый	Русский	острый инструмент с режущим краем	острый	физ.свойства	исходное

острый язык	Острый	Русский	остроумный, язвительный, меткий человек	острый	качества человека	переносное
четкая линия	-острый	Русский	четкая линия /изображение с четкими линиями	острый	физ.свойства: форма	переносное
крутая гора	-острый	Русский	отвесный, обрывистый крутой склон	острый	физ.свойства: форма	переносное
резкий бросок	-острый	Русский	внезапное, быстрое движение	острый	высокая скорость	переносное

Таблица 9

Как видно из фрагмента таблицы, все строки анкеты заполнены одним из двух способов: либо «острый», либо «-острый»; при этом вторым ярлыком отмечены те строки, у которых в поле ‘контекст’ прилагательное *острый* не присутствует.

Для того чтобы не ограничивать материал эксперимента только теми микрофреймами, которые покрывает слово *острый*, мы собрали вектора для контекстов всей анкеты. Соответственно, в нашу выборку попали и такие словосочетания как *крутая гора*, *резкий бросок*, *четкая линия* и др. Для каждого из них также вычислялись семантические вектора, а затем считалась близость их векторов с векторами других словосочетаний.

Итак, в первом типе данных список контекстов анкеты представляет собой «минимальные ситуации», которые релевантны для разных языков и которые описываются в них разными лексемами. Для каждой пары таких «минимальных ситуаций» мы вычисляли частотность их покрытия в языках одним прилагательным. Во втором типе данных этот же список контекстов играет роль анкеты, заполненной для русского языка, и для каждой пары русских словосочетаний мы вычисляли степень близости их значений.

Благодаря тому, что оба вычисления производились на материале одного и того же списка контекстов, мы можем провести сопоставление полученных цифр и посмотреть, насколько коэффициент частотности покрытия одной лексемой двух конкретных микрофреймов коррелирует с близостью векторов словосочетаний, репрезентирующих эти микрофреймы, в русском языке. Иными словами, если, например, частотность совместной встречаемости микрофреймов ‘острый нож’ и ‘острый меч’ в Базе равна 0.8, то значит ли

это, что близость векторов словосочетаний *острый нож* и *острый меч* в корпусе русского языка тоже будет высока?

Если да, т.е. если коэффициент корреляции между двумя типами данных окажется высок, то, с одной стороны, наша гипотеза о том, что объединение минимальных ситуаций в один фрейм проявляется и на одном языке, подтвердится, а с другой стороны, это послужит некоторой верификацией фреймового подхода в лексической типологии.

4.3 Параметры модели

Ниже приведены значения параметров примененной в ходе исследования модели дистрибутивной семантики.

4.3.1 Корпус

В разные моменты исследования для вычисления векторов были привлечены следующие корпуса:

- Основной подкорпус Национального корпуса русского языка (~200 млн. словоупотреблений): http://ruscorpora.ru/search-main.html
- Газетный подкорпус Национального корпуса русского языка (~220 млн. словоупотреблений): http://ruscorpora.ru/search-paper.html
- Интернет-корпус RuWac (~1млрд словоупотреблений): http://corpus.leeds.ac.uk/ruscorpora.html

4.3.2 Измерения векторов

В качестве измерений векторов использовался список из 10000 значимых частотных слов русского языка

4.3.3 Контекст

Для всех анализируемых слов/словосочетаний в качестве контекста использовались значимые слова, с которыми они встретились в окне +/- 5.

4.3.4 Метрики сходства

В качестве меры близости векторов во всех проведенных экспериментах вычислялись косинус угла между векторами и Евклидово расстояние между ними. Последняя метрика, однако, не показала значимых результатов, значительно отличавшихся бы от первой, и поэтому ей не уделяется внимания в работе.

4.3.5 Обработка векторов

В ходе проведения экспериментов использовались следующие методы обработки векторов:

1. Нормализация по длине вектора и по частотности его слов-измерений
2. Взвешивание вектора в соответствие с различными схемами (Positive Point-wise Mutual Information, Positive Local Mutual Information, Exponential Point-wise Mutual Information, Positive Log Weighting)
3. Уменьшение размерности векторов с 10000 измерений до 300

4.3.6 Способ сбора векторов

Вектора словосочетаний анкеты собирались по корпусу двумя способами.

При первом способе каждое словосочетание бралось как цельный элемент. Для него собирался вектор в окне +/-5.

При втором способе вектор словосочетания получался за счет композиции векторов соответствующего прилагательного и существительного. При этом были использованы разные модели композиции: аддитивная, мультипликативная, модели гомотетии, с тренировкой и без.

Для некоторых этапов исследования нам были нужны и отдельно вектора существительных, входящих в словосочетания анкеты (*нож*, *нос*, *меч*, *боль* и другие).

Все операции над векторами были произведены при помощи программного обеспечения DISSECT (**DIS**tributional **SE**mantics **C**omposition **T**oolkit).

4.4. Параметры Базы и способы подсчета корреляции

Для подсчета частотности покрытия микрофреймов одной лексемой в Базе были использованы следующие подмножества анкеты:

- все микрофреймы анкеты (150 словосочетаний)
- микрофреймы, описывающие прямые значения (31 словосочетание)
- «частотные» микрофреймы – те, информация о покрытии/не-покрытии которых есть, по крайней мере, для 10 прилагательных (26 словосочетаний)

При сопоставлении двух типов данных вычислялись коэффициенты корреляции Пирсона и Спирмена, а также p-value.

Результаты экспериментов приведены в приложении 3.

4.5. Baseline

В каждом эксперименте мы вычисляли не только вектора словосочетаний анкеты, но и вектора существительных, которые в эти словосочетания входят.

Для этих векторов также вычислялась мера близости, а потом считалась корреляция между степенью близости векторов существительных и частотой совместной встречаемо-

сти в Базе микрофреймов, в которые эти существительные входят. Так, например, вычислялось, насколько высокая частотность покрытия одной лексемой в Базе микрофреймов ‘острый нож’ и ‘острый меч’ коррелирует с близостью векторов русских слов *нож* и *меч*. Соответственно, если эта корреляция оказывалась выше, чем в случае с близостью векторов словосочетаний *острый нож* и *острый меч*, это говорило о том, что значения параметров модели заданы неудачно и что близость векторов двух существительных говорит нам больше о попадании двух словосочетаний в один фрейм, чем близость векторов самих словосочетаний.

4.6 Первый эксперимент

В начальном эксперименте мы посчитали корреляцию между двумя типами данных без предварительной их обработки:

- вектора собирались для словосочетаний как единого элемента на основном подкорпусе Национального корпуса русского языка (200 млн словоупотреблений)

- нормализация, взвешивание, уменьшение размерности не проводились

- в качестве второго типа данных была взята статистика частотности совместной встречаемости **всех** микрофреймов в Базе.

В результате мы получили довольно низкий коэффициент корреляции: 0.025 (Спирмен) и 0.034 (Пирсон).

При оставлении тех же значений параметров модели, но сборе векторов не для словосочетаний, а для существительных, которые в эти словосочетания входят, коэффициент корреляции получается выше: 0.045 (Спирмен и Пирсон).

А значит, без предварительной обработки векторов словосочетаний и при использовании всего множества строк Базы для подсчета корреляции результат эксперимента оказывается скорее отрицательным.

4.7 Изменение параметров модели: нормализация, взвешивание, уменьшение размерности

Безусловно, никакая уважающая себя модель дистрибутивной семантики не должна строиться на необрабатываемых векторах. Поэтому мы применили к собранным векторам нормализацию, разные схемы взвешивания и уменьшение размерности векторов до 300 координат. Лучший результат, которого удалось добиться таким образом - коэффициент корреляции Пирсона 0.068, который был достигнут при нормализованных по длине, взвешенных по схеме Exponential Point-wise Mutual Information векторах с уменьшенной размерностью.

Однако сопоставление статистики из Базы с векторами существительных, обработанных таким же способом, опять дало более высокий результат – 0.092 (Пирсон).

4.8 Изменение параметров Базы

Следующая серия экспериментов была связана с изменениями параметров Базы данных.

Так как в типологической анкете, которая составляет основу Базы, не все микрофреймы обязательны для заполнения, есть такие, которые заполняются довольно редко, поэтому их присутствие в эксперименте загрязняет картину. Так, например, коэффициент частотности покрытия одной лексемой у микрофреймов ‘острый зуб’ и ‘острое сопротивление’ равен единице. Ни один из этих микрофреймов не является обязательным, каждый из них заполнялся в маленьком количестве лексем, а информация об обоих присутствует только в одном языке - русском. А в этом языке оба микрофрейма описываются прилагательным *острый*. Получается, что в единственном случае, в котором указана информация об обоих микрофреймах, они покрываются одной лексемой, а значит, значение частотности покрытия одной лексемой равно:

$(nAdj_{[+a,+b]} + nAdj_{[-a,-b]}) / (nAdj_{[-a,+b]} + nAdj_{[+a,-b]} + nAdj_{[+a,+b]} + nAdj_{[-a,-b]}) =$

$= (1 + 0) / (0 + 0 + 1 + 0) = 1,$

где $Adj_{[+a,+b]}$ – это русское прилагательное *острый*.

Понятно, что такие случаи загрязняют картину и препятствуют высокому коэффициенту корреляции. Поэтому в качестве первого изменения параметров Базы было решено рассчитывать частотность совместной встречаемости только тех микрофреймов, про которые в Базе есть достаточное количество информации. За нижний порог мы взяли 10 лексем: все микрофреймы, о покрытии/не-покрытии которых есть информация про 9 лексем и меньше, в выборку не попали. Таких «частотных» микрофреймов оказалось 26. Это число меньше количества обязательных микрофреймов (55), что связано с тем, что для некоторых лексем на данный момент заполнены только прямые, физические значения.

В таблице 10 представлен полный список микрофреймов, которые попали в эту выборку:

острый_нож	острый_иголка	острый_носок	острый_ум
острый_меч	острый_зуб	острый_локоть	резкий_слово
острый_лезвие	острый_коготь	острый_подбородок	резкий_ветер
острый_коса	острый_ноготь	острый_нос	острый_боль

острый_ножницы	острый_палка	острый_рог	резкий_звук
острый_стрела	острый_клюв	резкий_взгляд	
острый_копье	острый_шип	острый_язык	

Таблица 10

Как видно, большая часть микрофреймов представляет прямые значения (что и связано с тем, что для части прилагательных заполнены пока только они), но ими часто заполняемые микрофреймы не исчерпываются. Кроме них в эту выборку попали и словосочетания, представляющие наиболее частотные переносы – ‘хорошо функционирующий ум’, ‘внезапная сильная боль’, ‘пронизывающий ветер’, и другие.

Сбор частотности совместной встречаемости для такого подмножества Базы и подсчет корреляции этого типа данных с близостью векторов русских словосочетаний дали значительно лучший результат: 0.198 (Спирмен), 0.21 (Пирсон). Такой коэффициент уже говорит о том, что зависимость между двумя случайными величинами (двумя типами данных) есть.

Однако результат сопоставления с векторами существительных при таких же параметрах снова оказался значительно выше: 0.326 (Спирмен), 0.278 (Пирсон). А для того, чтобы считать эксперимент успешным, нам нужно не только добиться высокого коэффициента корреляции, но и «оторваться» от соответствующего коэффициента для векторов существительных.

Другой способ выбрать из 150 микрофреймов Базы подмножество для подсчета статистики совместной встречаемости – принятие в выборку микрофреймов только прямых значений.

Такое сокращение анкеты объясняется уже не техническими трудностями, а теоретическими предпосылками. То, что прямые и переносные значения - две абсолютно разных сущности, известный факт. Когда типологи утверждают, что у лексических полей есть структура, которую можно изучать и сопоставлять в разных языках, речь идет, в первую очередь, о прямых значениях слов. Метафоры значительно менее структурированы. Поэтому они не помещаются на семантические карты, разрабатываемые в рамках фреймового подхода. Именно прямые значения составляют ядро любого семантического поля, определяя стратегии лексикализации в нем.

Часто ставится вопрос, есть ли у переносных значений вообще какая-то структура. Мы склонны отвечать на него положительно, о чем было сказано в первой главе и о чем пойдет речь ниже. Однако эта структура принципиально иного рода, чем у прямых значе-

ний, а поэтому эксперимент с подсчетом совместной встречаемости микрофреймов только прямых значений имеет смысл. Таких микрофреймов в типологической анкете поля ‘острый’ 34 (см. первые 34 строки анкеты в приложении 2).

Наибольший коэффициент корреляции, которого нам удалось достичь при таких параметрах Базы – 0.308 (Спирмен), 0.302 (Пирсон).

4.9 Изменение объема корпуса

Один из параметров DS-модели, которому традиционно уделяется большое внимание в векторной семантике – объем корпуса. Для того чтобы качественно представить дистрибуцию лексической единицы, нужен достаточно большой корпус текстов. Понятно, что если речь идет о векторах словосочетаний, а не отдельных слов, то требуются еще большие его размеры. Поэтому в ходе исследования мы дважды увеличивали объем корпуса, на котором собирались вектора. Сначала был добавлен миллиардный интернет-корпус RuWac, а затем газетный подкорпус Национального корпуса русского языка объемом в 220 млн словоупотреблений.

Результат, которого мы достигли, кажется довольно нетривиальным. Традиционно считается, что для любой векторной модели чем корпус больше, тем лучше. Однако, как оказывается, в этом отношении вектора словосочетаний отличаются от векторов отдельных слов.

Так, в случае со словосочетаниями, увеличение объема корпуса всегда влечет за собой повышение коэффициента корреляции, ср. результаты экспериментов сопоставления с частотными микрофреймами Базы:

объем корпуса	нормализация/ взвешивание/ум. разм	параметры Базы	Корреляция (Spearman)	Корреляция (Pearson)	p-value
200млн	нормализация по длине, plmi-weighting	частотные м/ф	0.198	0.21	0.000128688399869
1200млн	нормализация по длине, plmi-weighting	частотные м/ф	0.24	0.283	0.000000213562692
1420млн	нормализация по длине, plmi-weighting	частотные м/ф	0.245	0.287	0.000000146144194

Таблица 11

С векторами отдельных слов наблюдается несколько другая картина: при увеличении объема корпуса коэффициент корреляции не только не всегда увеличивается, но иногда и, наоборот, идет на убыль. Так, проведение эксперимента с теми же значениями параметров модели, что и в предыдущей таблице, но с векторами существительных, а не словосочетаний показало следующий результат:

объем корпуса	нормализация/ взвешивание/ум. разм	параметры Базы	Корреляция (Spearman)	Корреляция (Pearson)	p-value
200млн	нормализация по длине, plmi-weighting	частотные м/ф	0.326	0.278	0.000000363617996
1200млн	нормализация по длине, plmi-weighting	частотные м/ф	0.351	0.285	0.000000179083025
1420млн	нормализация по длине, plmi-weighting	частотные м/ф	0.349	0.284	0.000000174078897

Таблица 12

Как видно, при первом увеличении корпуса коэффициент корреляции заметно повысился, а при втором стал немного ниже.

Такой результат, вероятно, объясняется тем, что для отдельного слова качественный вектор собрать намного проще, чем для словосочетания. Ведь в любом корпусе слово *нож* встречается намного чаще, чем словосочетание *острый нож*. А значит, «потолок» – такой размер корпуса, при котором дальнейшее его увеличение уже не будет вести к улучшению качества вектора – для отдельных слов наступает значительно раньше. В нашем случае, как кажется, он наступил после первого увеличения корпуса. Второе увеличение его объема уже не дало более высокой цифры. Наоборот, некоторое ее уменьшение, как кажется, можно связать со степенью сбалансированности добавленного корпуса. Последний корпус, который мы добавили, - это газетный подкорпус Национального корпуса русского языка. По своей природе, он несбалансирован. А значит, полученные на его основе вектора могут несколько «загрязнять» общую картину.

Итак, эксперимент с увеличением размера корпуса до полутора миллионов показал, что этого объема (и даже, несколько меньшего) достаточно, чтобы получить качественный вектор слова. Такой же «потолок» для словосочетаний достигнут не был. Понятно, что для качественных векторов словосочетаний нужен существенно больший корпус, но насколь-

ко именно должен быть увеличен его объем, мы на основании настоящего исследования сказать не можем.

Однако в векторной семантике разработан альтернативный метод сбора вектора словосочетания, о котором мы упоминали во второй главе и о котором пойдет речь сейчас. Этот способ – композиция векторов существительного и прилагательного – как кажется, открывает новую перспективу исследования.

4.10 Композиция векторов

Последний параметр DS-модели, значение которого мы меняли в ходе наших экспериментов – способ сбора векторов словосочетаний. Помимо метода, при котором словосочетание трактуется как цельный элемент, для которого собирается список контекстов в окне +/- 5, мы использовали метод композиции. Наша мотивация здесь была проста. Для того чтобы собрать качественный вектор словосочетания нужен корпус, явно больший, чем есть в нашем распоряжении, в то время как для качественных векторов отдельных слов нашего корпуса более чем достаточно.

В ходе экспериментов мы использовали три метода композиции – простую сумму векторов, их умножение и гомотетию.

Несмотря на то, что все три метода композиции показали хорошие результаты, настоящий «прорыв» был достигнут при помощи самого простого метода композиции – суммы.

объем корпуса	нормализация/ взвешивание/ум. разм	метод композиции	параметры Базы	Корреляция (Spearman)	Корреляция (Pearson)	p-value
1420млн	-	сумма	все м/ф	0.502	0.492	0
1420млн	-	умножение	все м/ф	0.31	0.315	0
1420 млн	-	гомотетия	все м/ф	0.352	0.34	0

Однако простая композиция векторов двух слов не учитывает принципиально разный вклад, который каждое из них вносит в значение словосочетания. Так, в случае сочетания прилагательного и существительного, последнее, как кажется, играет большую роль в значении общей конструкции. Для того чтобы это учесть, мы провели тренировку для двух моделей композиции (модель умножения ее проводить не позволяет). В случае с суммой мы вычислили коэффициенты, на которые каждый из двух векторов умножался, а с гомотетией – значение скалярной величины λ.

объем корпуса	нормализация/ взвешивание/ум. Разм	метод композиции	параметры Базы	Корреляция (Spearman)	Корреляция (Pearson)	p-value
1420млн	-	сумма, тренировка	все м/ф	0.619	0.632	0
1420млн	-	сумма, тренировка	прямые м/ф	0.523	0.726	0
1420млн	-	гомотетия, тренировка	все м/ф	0.282	0.299	0
1420 млн	-	гомотетия, тренировка	прямые м/ф	0.226	0.226	0.00000073566282

Таблица 13

Как видно из таблицы, если тренировка модели гомотетии не принесла положительных результатов, то тренировка суммы позволила достичь коэффициента корреляции 0.726 (для прямых значений). Этот результат значительно превышает значение коэффициента корреляции при таких же параметрах модели для векторов существительных (0.326).

А значит, нам удалось найти такие значения параметров модели, при которых:

- коэффициент корреляции настолько высок, что можно утверждать не только наличие зависимости между двумя типами данных, но и достаточно высокую степень этой зависимости;
- результат сопоставления со статистикой из Базы оказывается значительно выше у векторов словосочетаний, чем у векторов существительных, которые в эти словосочетания входят.

4.11 Ближайшие соседи векторов

Более наглядное представление об удаче эксперимента дают списки ближайших соседей. При высоком коэффициенте корреляции мы ожидаем, что ближайшими соседями вектора данного словосочетания будут вектора словосочетаний, относящихся к тому же фрейму. Запросы ближайших соседей для словосочетаний нашей анкеты полностью оправдывают эти ожидания. Так, наиболее близкими к словосочетанию *острая обида* оказываются словосочетания, обозначающие сильные проявления эмоции:

острый_обида

острый_зависть

острый_ревность

острый_любопытство

острый_гордость

А к конструкции *резкий голос* – преимущественно словосочетания, относящиеся к фрейму ‘неприятно воздействующий на органы чувств’:

резкий_голос

резкий_звук

резкий_взгляд

резкий_вкус

резкий_запах

Интересно, однако, посмотреть, какие словосочетания попадают в список ближайших соседей помимо тех, которые относятся к тому же фрейму. Ведь если в соседях у конкретного словосочетания окажется сочетание из другого фрейма, то этот фрейм, вероятно, должен быть близок к данному. Эту гипотезу проще всего проверить на фреймах прямых значений, так как верификацией здесь может послужить семантическая карта.

На семантической карте поля ‘острый’ изображено три фрейма – ‘острые режущие инструменты’, ‘острые колющие инструменты’ и ‘предметы с острой формой’, причем ‘острые колющие инструменты’ располагаются между двумя другими фреймами (семантическая карта признака приведена на рис. 1 и для удобства повторена на рис. 6 ниже):

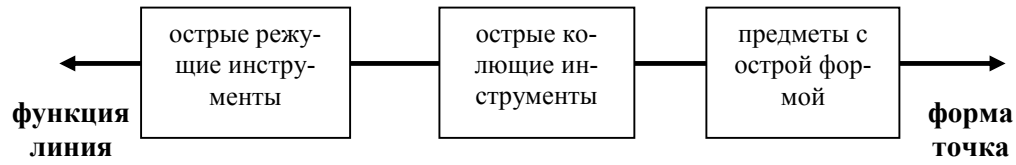

рис.6

Такое взаимное расположение фреймов отражает их когнитивную близость и позволяет делать предсказания о том, какие стратегии лексикализации в поле возможны, а какие нет. Так, поскольку к фрейму ‘острые режущие инструменты’ ближе ‘острые колющие инструменты’, чем ‘предметы с острой формой’, не может быть прилагательного, которое бы описывало ‘острые режущие инструменты’ и ‘предметы с острой формой’, но не описывало бы ‘острые колющие инструменты’.

Как кажется, список ближайших соседей для векторов словосочетаний поля, передающих прямые значения, отражает представленную структуру. Так, среди соседей слово-

сочетания ‘острый нож’, находятся как словосочетания из того же фрейма, так и словосочетания из фрейма ‘острые колющие инструменты’, но не из фрейма ‘предметы с острой формой’:

острый_нож

острый_лезвие

острый_ножницы

острый_палка

острый_стрела

острый_копье

острый_сабля

острый_пила

острый_меч

острый_коготь

А среди соседей словосочетания *острое копье* есть примеры из всех трех фреймов, что опять-таки полностью предсказывается картой: ‘острые колющие инструменты’ близки и к фрейму ‘острые режущие инструменты’, и к фрейму ‘предметы с острой формой’.

острый_копье

острый_стрела

острый_кол

острый_сабля

острый_лезвие

острый_шип

острый_пила

острый_игла

острый_колпак

острый_клюв

Для метафорических значений поиск словосочетаний, не входящих в тот же фрейм, открывает две новых перспективы. Во-первых, несмотря на меньшую структурированность переносных значений, по сравнению с прямыми, некоторые из них явно кластеризуются в группы. Так, очень близки друг к другу разные фреймы, передающие смысл интенсивности – ‘интенсивное физиологическое ощущение’, ‘сильная эмоция’, ‘сильно проявляющееся качество человека’, ‘напряженное взаимодействие’ и др. - что можно увидеть, например, расширив список ближайших соседей для сочетания *острая обида*.

острый_обида

острый_зависть

острый_ревность

острый_любопытство

острый_гордость

острая жажда

острая слабость

острое осознание

острая дрожь

Как видно, пятым, шестым и восьмым соседом *острой обиды* являются словосочетания из фрейма ‘интенсивное физиологическое ощущение’, а седьмым – из фрейма ‘отчетливое, яркое ощущение’.

Во-вторых, нахождение среди ближайших соседей переносного словосочетания лексических конструкций из прямых фреймов могло бы послужить косвенным доказательством мотивированности метафор – связи каждой из них с одним или несколькими конкретными фреймами прямых значений. Такие связи мы находили в процессе ручного анализа языкового материала: при исследовании переносных значений слов, которые в своих прямых употреблениях описывают только один фрейм, мы обнаружили, что эти значения часто повторяются. Так, слова со значением ‘острый’, которые описывают только фрейм ‘острых режущих инструментов’, типологически часто метафорически переносятся на обозначение ‘резкого ветра’ (фр. *tranchant*, коми *лэчыд*, ит. *tagliente*) или ‘четкой границы’ (англ. *razor-sharp*). Более того, на нашей выборке нет ни одного прилагательного, которое бы описывало хотя бы одно из этих переносных значений и при этом не покрывало фрейм острых режущих инструментов. Такие факты позволяют предположить, что переносные значения ‘резкий, порывистый’ (о ветре) и ‘отчетливый, ясно различимый’ производны именно от фрейма ‘острые режущие инструменты’. Некоторые другие метафоры, наоборот, как кажется, тесно связаны с фреймом ‘острые колющие инструменты’. К таким относятся, например, значения ‘хорошо функционирующий’ (об уме, зрении, слухе), ‘остроумный’ (о человеке), ‘очень интенсивный’ (о боли).

Примечательно, что в десяток ближайших соседей к последнему упомянутому словосочетанию – *острой боли* – попала пара словосочетаний из прямых фреймов.

острый_боль

острый_слабость

острый_обида

острый_жажда

острый_дрожь

острый_гордость

острый_зависть

острый_ревность

острый_палка

острый_коготь

И этим прямым фреймом оказался именно ‘острые колющие инструменты’. А значит, модели векторной семантики вновь подтвердили наши теоретические обобщения.

Итак, просмотр списков ближайших соседей для векторов словосочетаний из анкеты признака ‘острый’, во-первых, позволяет убедиться, что соседями любого сочетания являются, преимущественно, словосочетания из того же фрейма, а во-вторых, показывает фреймы, близкие к данному.

5. Заключение

Сопоставив два типа данных – частотность покрытия двух микрофреймов одной лексемой на материале Базы и близость векторов русских словосочетаний анкеты – мы достигли высокого показателя коэффициента корреляции. Это говорит о том, что два типа данных не независимы. Если два микрофрейма часто покрываются одной лексемой в Базе, то и словосочетания, их представляющие в русском языке, будут близки.

Поскольку в частотности покрытия микрофреймов одной лексемой отражается объединение их во фреймы, высокий коэффициент корреляции свидетельствует о том, что, фреймовая структура, универсальная для всех языков, проявляет себя и на одном языке.

Такой результат может послужить и верификацией фреймового подхода: он является статистическим доказательством того, что фреймы действительно организуют структуру лексического поля.

В ходе эксперимента мы обнаружили, что разные параметры модели по-разному влияют на конечный результат.

Так, оказывается очень важным, какое подмножество микрофреймов типологической анкеты использовать в качестве первого типа данных для сопоставления. Например, из-за того, что информация не о всех микрофреймах указана для всех лексем, некоторые контексты, наименее часто заполняемые, могут загрязнять общую картину. Поэтому отбор микрофреймов по частоте заполняемости значительно улучшает результат. Более сильное влияние на итоги сопоставления, однако, оказывает другой метод отбора, основанный не на технических характеристиках Базы, а на теоретических предпосылках. Коэффициент корреляции значительно увеличивается при принятии в качестве первого типа данных микрофреймов анкеты, реализующих прямые употребления прилагательных. Такую зависимость можно объяснить структурностью, которая прямым значениям свойственна в значительно большей мере, чем метафорам.

Увеличение корпуса, с другой стороны, может влиять на результат двояко. Наши эксперименты показали, что качество векторов слов не бесконечно повышается с увеличением объема корпуса. В определенный момент наступает «потолок» - размер корпуса, дальнейшее увеличение которого не приводит к значительному улучшению качества вектора. Наоборот, оно может его даже ухудшить. Для векторов словосочетаний эта же тенденция на материале исследования не подтвердилась. Поскольку вектор слова собрать значительно проще, чем словосочетания, для того чтобы достичь хорошего качества последнего, нужен существенно больший объем текстовых данных.

Однако, увеличивать объем корпуса – задача очень трудоемкая, и существуют другие способы достичь высокого качества векторов словосочетаний. Это различные модели композиции векторов – получения вектора составной конструкции из векторов входящих в нее простых. Эти модели, набирающие все большую силу в последнее время, просто незаменимы для создания качественной семантической репрезентации единиц, больших, чем слово. Их логика довольно проста. Объясним ее на примере сочетания прилагательного с существительным и самой простой модели композиции – аддитивной. Сочетаемость словосочетания *острый нож*, например, довольно сильно «похожа» на сочетаемость существительного *нож*. И то, и то может в текстах *резать, протыкать, вспарывать, колоть*, быть *металлическим, длинным, большим* или *маленьким*, и то, и то можно *занести над головой, поточить* или *вынуть из кармана*. Наличие рядом с этим существительным прилагательного *острый*, однако, сужает его сочетаемость: в текстах может возникнуть *перочинный нож*, но не *перочинный острый нож, столовый нож*, но не *столовый острый нож*. Именно это сужение и обеспечивается присоединением к вектору *ножа* вектора *острого*: в результирующем векторе высокие показатели будут у компонент, которые релевантны как для *острого*, так и для *ножа* (таких, как *резать, колоть*), а низкие показатели будут у тех компонент, которые могут выступать в качестве контекста у *ножа*, но не у *острого* (таких, как *перочинный*), и наоборот.

Такое действие моделей композиции можно сравнить с работой речевого тракта. При порождении звука в легких возникает воздушная струя. Проходя через различные локации речевого тракта, струя видоизменяется. В результате получается тот или иной звук. Вектор существительного, в таком случае, можно метафорически сопоставить с воздушной струей, а вектор прилагательного – с речевым трактом, который меняет ее форму.

В аддитивной и мультипликативной моделях композиции, которые применялись в ходе экспериментов, вклад вектора прилагательного и существительного в семантическую репрезентацию вектора словосочетания принципиально симметричен. Применение к этим моделям тренировки и приписывание двум векторам разных весов является довольно «грубым» способом выразить неравноценную их роль в семантике словосочетания. В случае взвешенных моделей веса векторов оказываются, действительно, разными, но их природа остается одной и той же.

Более элегантное решение этой проблемы представлено в функциональных моделях композиции, согласно которым значение существительного можно представить в виде вектора, а прилагательного – в виде функции, которая видоизменяет вектор существи-

тельного. Поэтому один из планов на будущее исследование по поднятой в работе проблематике – повторение эксперимента с функциональными моделями композиции.

Другой планируемый способ расширения проекта – привлечение к эксперименту других признаковых полей. Несмотря на то, что результаты, достигнутые с одним признаковым полем, выглядят вполне убедительно, мы не можем быть уверены, что с другими полями будет наблюдаться ровно такая же картина. Поэтому мы бы хотели повторить исследование с полями, которые максимально не похожи на ‘острый’: это, во-первых, поля, которые в русском языке представлены не-доминантной системой (не с одним основным прилагательным, покрывающим все ядро поля), а во-вторых, которые имеют другую семантическую ориентацию. Так, интересно посмотреть, так же ли поведут себя прилагательные, описывающие исключительно форму объектов (например, прилагательные размера – *узкий, широкий, высокий, короткий* и др.) или, наоборот, строго экспериенциальную зону (*тяжелый*, *тугой*). Наконец, наиболее убедительно бы прозвучали выводы эксперимента, проведенного на основе не только признаковой, но и глагольной лексики. Поэтому в будущем мы планируем собрать схожую статистику по глаголам.

Векторные модели позволяют также ставить в соответствие друг другу векторные пространства, собранные на основе корпусов разных языков. Реализация этой возможности позволит нам расширить исследование, добавляя в качестве экспериментальных данных текстового материала других языков, в первую очередь, английского.

Литература

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

Berlin, B.; Kay, P. 1969. Basic color terms: Their universality and evolution. Berkeley: Univ. of California press

Evans, N. 2010. Semantic typology // Jae Jung Song (ed.) The Oxford Handbook of Linguistic Typology. Oxford/ New York: Oxford University Press, pp. 504-533

Goddard, Cliff and Wierzbicka, Anna (eds.) 1994. Semantic and Lexical Universals - Theory and Empirical Findings. Amsterdam/Philadelphia: John Benjamins

Goddard, Cliff. 1998. Semantic Analysis: A practical introduction. Oxford. Oxford University Press

Goddard, Cliff (ed.) 2008. Cross-Linguistic Semantics. Amsterdam/Philadelphia: John Benjamins

Koptjevskaja-Tamm, M. 2008. Approaching Lexical typology // M. Vanhove (ed.) From polysemy to semantic change: Towards a typology of lexical semantic associations. Amsterdam/Philadelphia: John Benjamins, 3–52

Kotlerman L.; Dagan I.; Szpektor I.; Zhitomirsky-Geffet, M. 2010. Directional distributional similarity for lexical inference // *Natural language engineering* 16(4): 359-389.

Landauer, T.; Dumais, S. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review* 104(2): 211-240

Levinson, S. C., & Wilkins, D. P. (Eds.) 2006. Grammars of space: Explorations in cognitive diversity. Cambridge: Cambridge University Press

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

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

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

Padó, U.; Padó, S.; Erk K. 2007. Flexible, corpus-based modeling of human plausibility judgments // *Proceedings of EMNLP*: 400-409. Prague, Czech Republic.

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

Wierzbicka, A. 1972. Semantic primitives. Frankfurt (M): Athenäum

Wierzbicka, A. 1996. Semantics: Primes and Universals. Oxford: Oxford University Press

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

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

Гладкова А. Н. 2010. Русская культурная семантика: эмоции, ценности, жизненные установки. М.: Языки славянской культуры

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

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

Кюсева М.В., Рыжова Д.А., Холкина Л.С. 2012. Прилагательные *тяжелый* и *легкий* в типологической перспективе // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Международной конференции «Диалог» (Бекасово, 30 мая – 3 июня 2012 г.) В 2-х томах / Отв. ред. А.Е.Кибрик. Т.1: Основная программа конференции. Вып. 11. М.: РГГУ, с. 247-255

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

Лучина Е. С., Резникова Т. И., Стенин И. А. 2013. Атрибутивы как источник грамматикализации: ‘прямой’ и ‘ровный’ в русском, немецком и финском языках // Guzmán Tirado R., Votyakova I. (eds.) Tipología léxica. Granada. p. 123-129.

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

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

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

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

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

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

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

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

Шапиро М.М. 2013. Глаголы колебательного движения в типологической перспективе. Дипломная работа. М.: МГУ

Приложение 1

Лексемы поля ‘острый’, представленные в Базе

прилагательное	язык
острый	русский
oštar	сербский
sharp	английский
scharf	немецкий
spitz	
aigu	французский
pointu	
tranchant	
miniog	валлийский
llym	
siarp	
pigog	
eles	венгерский
hegyes	
szuros	
terävä	финский
pistävä	
лэчыд	коми
юэс	
orža	мокшанский
pika	
hüt̄e	агульский
жьан	кабардинский
памцIэ	
surudoi	японский
togatta	
jiān	китайский
jianrui	
jianli	
fengli	
ruili	
kuài	
tajam	малайский

Приложение 2

Типологическая анкета признака ‘острый’

(цветом отмечены обязательные микрофреймы)

	микрофрейм	Фрейм
1	острый нож	острый инструмент с режущим краем
2	острый меч	острый инструмент с режущим краем
3	острый коготь	острый инструмент с режущим краем
4	острая сабля	острый инструмент с режущим краем
5	острое лезвие	острый инструмент с режущим краем
6	острая коса	острый инструмент с режущим краем
7	острые ножницы	острый инструмент с режущим краем
8	острая пила	острый инструмент с режущим краем
9	острая иголка	острый инструмент с колющим концом
10	острая стрела	острый инструмент с колющим концом
11	острый коготь	острый инструмент с колющим концом
12	острый кол	острый инструмент с колющим концом
13	острая спица	острый инструмент с колющим концом
14	острая палка	острый инструмент с колющим концом
15	острое копье	острый инструмент с колющим концом
16	острый зуб	острый инструмент с колющим концом
17	острый ноготь	острый инструмент с колющим концом
18	острый рог	острый инструмент с колющим концом
19	острый нос	предмет, суживающийся к концу
20	острый угол	геометрическая форма
21	острый носок ботинка	предмет, суживающийся к концу
22	острая гора	предмет, суживающийся к концу
23	острый локоть	предмет, суживающийся к концу
24	острый клюв	предмет, суживающийся к концу
25	острый колпак	предмет, суживающийся к концу
26	острый подбородок	предмет, суживающийся к концу
27	острый каблук	предмет, суживающийся к концу

28	острая мачта	предмет, суживающийся к концу
29	острый шип	колющаяся поверхность/колющийся предмет
30	колючее одеяло	колющаяся поверхность/колющийся предмет
31	колючая щетина	колющаяся поверхность/колющийся предмет
32	колючая шерсть	колющаяся поверхность/колющийся предмет
33	колючий куст	колющаяся поверхность/колющийся предмет
34	колючий свитер	колющаяся поверхность/колющийся предмет
35	крутой поворот	полоса, резко меняющая направление
36	крутой изгиб	полоса, резко меняющая направление
37	крутая излучина реки	полоса, резко меняющая направление
38	крутой подъем	отвесный, обрывистый, крутой склон
39	крутой склон	отвесный, обрывистый, крутой склон
40	крутой берег	отвесный, обрывистый, крутой склон
41	крутая гора	отвесный, обрывистый, крутой склон
42	четкая линия	четкая линия/изображение с четкими линиями
43	четкая фотография	четкая линия/изображение с четкими линиями
44	резкий контраст	четкая линия/изображение с четкими линиями
45	четкая картинка	четкая линия/изображение с четкими линиями
46	сильные очки	четкая линия/изображение с четкими линиями
47	проницательный человек	умный, хорошо соображающий (о человеке)
48	острый ум	умный, хорошо соображающий (о человеке)
49	острый, проницательный взгляд	умный, хорошо соображающий (о человеке)
50	проницательный наблюдатель	умный, хорошо соображающий (о человеке)
51	умная девочка	умный, хорошо соображающий (о человеке)
52	умная голова	умный, хорошо соображающий (о человеке)
53	острый журналист	остроумный, язвительный, меткий человек
54	острое слово	остроумный, язвительный, меткий человек
55	острый язык	остроумный, язвительный, меткий человек
56	острое высказывание	остроумный, язвительный, меткий человек
57	острый критик	остроумный, язвительный, меткий человек
58	острое замечание	остроумный, язвительный, меткий человек
59	острая критика	остроумный, язвительный, меткий человек

60	острая статья	остроумный, язвительный, меткий человек
61	острый анекдот	остроумный, язвительный, меткий человек
62	резкий человек	грубый, жестокий, резкий (о человеке)
63	резкий взгляд	грубый, жестокий, резкий (о человеке)
64	резкое слово	грубый, жестокий, резкий (о человеке)
65	резкая реакция	грубый, жестокий, резкий (о человеке)
66	резкая брань	грубый, жестокий, резкий (о человеке)
67	грубый ответ	грубый, жестокий, резкий (о человеке)
68	грубая бесцеремонность	грубый, жестокий, резкий (о человеке)
69	грубый циннизм	грубый, жестокий, резкий (о человеке)
70	строгий закон	строгий (о человеке)
71	строгий правитель	строгий (о человеке)
72	строгий распорядок	строгий (о человеке)
73	строгое наказание	строгий (о человеке)
74	строгая диета	строгий (о человеке)
75	непоседливый ребенок	непоседливый (о человеке)
76	узкоспециализированное исследование	узкоспециализированное исследование
77	узкоспециализированная работа	узкоспециализированное исследование
78	острый соус	блюдо с большим количеством специй и пряностей
79	острый вкус	блюдо с большим количеством специй и пряностей
80	острый перец	блюдо с большим количеством специй и пряностей
81	острое лечо	блюдо с большим количеством специй и пряностей
82	газированный напиток	газированный напиток
83	газированная вода	газированный напиток
84	резкий звук	неприятно действующий на органы чувств: звук
85	острый визг	неприятно действующий на органы чувств: звук
86	резкий голос	неприятно действующий на органы чувств: звук
87	резкий лай	неприятно действующий на органы чувств: звук
88	резкий храп	неприятно действующий на органы чувств: звук
89	резкий вкус	неприятно действующий на органы чувств:вкус
90	резкий свет	неприятно действующий на органы чувств:свет

91	резкая вспышка	неприятно действующий на органы чувств:свет
92	острый блеск	неприятно действующий на органы чувств:свет
93	резкая вонь	неприятно действующий на органы чувств:запах
94	резкий запах	неприятно действующий на органы чувств:запах
95	высокий звук	высокий звук
96	высокая нота	высокий звук
97	яркий цвет	яркий цвет
98	яркая синева	яркий цвет
99	яркий оттенок	яркий цвет
100	сильный запах	сильный запах
101	сильный аромат	сильный запах
102	яркий свет	яркий свет
103	яркое солнце	яркий свет
104	острый слух	хорошо функционирующий орган чувств
105	острое зрение	хорошо функционирующий орган чувств
106	хорошие глаза	хорошо функционирующий орган чувств
107	хорошие уши	хорошо функционирующий орган чувств
108	острый нюх	хорошо функционирующий орган чувств
109	острое ощущение	отчетливое ощущение
110	острое осознание	отчетливое ощущение
111	острое желание	сильная эмоция
112	острая ревность	сильная эмоция
113	острая зависть	сильная эмоция
114	острая обида	сильная эмоция
115	острое стремление	сильная эмоция
116	острый мороз	погодное явление, проявляющееся в высокой степени
117	резкий ветер	погодное явление, проявляющееся в высокой степени
118	сильный холод	погодное явление, проявляющееся в высокой степени
119	сильная жара	погодное явление, проявляющееся в высокой степени
120	острая проблема	требующий немедленного разрешения
121	острая нехватка	требующий немедленного разрешения
122	острый вопрос	требующий немедленного разрешения

123	острая боль	интенсивное физиологическое ощущение
124	острая жажда	интенсивное физиологическое ощущение
125	острый недосып	интенсивное физиологическое ощущение
126	острая дрожь	интенсивное физиологическое ощущение
127	острая слабость	интенсивное физиологическое ощущение
128	острое воспаление	сильная болезнь
129	острый грипп	сильная болезнь
130	острое заболевание	сильная болезнь
131	острое расстройство	сильная болезнь
132	острый аппендицит	сильная болезнь
133	острый конфликт	напряженный (о взаимодействиях)
134	острый момент	напряженный (о взаимодействиях)
135	острое сопротивление	напряженный (о взаимодействиях)
136	острый спор	напряженный (о взаимодействиях)
137	острая перепалка	напряженный (о взаимодействиях)
138	острое любопытство	интенсивный по степени проявления (о качествах человека)
139	острая наблюдательность	интенсивный по степени проявления (о качествах человека)
140	острая впечатлительность	интенсивный по степени проявления (о качествах человека)
141	острая гордость	интенсивный по степени проявления (о качествах человека)
142	резкий подъем	резкое изменение
143	резкий спад	резкое изменение
144	резкий взмах	внезапное,быстрое движение
145	резкий вздох	внезапное,быстрое движение
146	резкое движение	внезапное,быстрое движение
147	резкий бросок	внезапное,быстрое движение
148	резкий скачок	внезапное,быстрое движение
149	быстрый темп	быстрый (о скорости)
150	быстрый бег	быстрый (о скорости)

Приложение 3

Результаты основных экспериментов

№ эксп.	объем кор-пуса	словосоч /сущ/ композиция	нормализация/ взвешивание/ уменьшение разм	параметры Базы	корреляция (Spearman)	корреляция (Pearson)	p-value
1	200 млн	словосоч	-	все м/ф	0.025	0.034	0.000346627929037
2	200 млн	сущ	-	все м/ф	0.045	0.0445	0.000002877106643
3	200 млн	словосоч	норм. по длине; взвеш. epmi; уменьш.разм.до 300	все м/ф	0.049	0.068	0.000000000000745
4	200 млн	сущ	норм. по длине; взвеш. epmi; уменьш.разм.до 300	все м/ф	0.073	0.092	0
5	200 млн	словосоч	норм. по длине; взвеш. Plmi	частотные м/ф	0.198	0.21	0.000128688399869
6	200 млн	сущ	норм. по длине; взвеш. Plmi	частотные м/ф	0.326	0.278	0.000000363617996
7	200 млн	словосоч	-	прямые м/ф	0.309	0.302	0.000000000020663
8	200 млн	сущ	норм. по длине; взвеш. epmi; уменьш.разм.до 300	прямые м/ф	0.213	0.23	0.000000173499042
9	1200 млн	словосоч	норм. по длине; взвеш. Plmi	частотные м/ф	0.24	0.284	0.000000213562692
10	1420 млн	словосоч	норм. по длине; взвеш. Plmi	частотные м/ф	0.245	0.287	0.000000146144194
11	1200 млн	сущ	норм. по длине; взвеш. Plmi	частотные м/ф	0.351	0.285	0.000000179083025
12	1420 млн	сущ	норм. по длине; взвеш. Plmi	частотные м/ф	0.349	0.284	0.000000174078897
13	1420 млн	сумма	-	все м/ф	0.502	0.492	0
14	1420 млн	сумма	-	прямые м/ф	0.524	0.693	0
15	1420 млн	сумма	-	частотные м/ф	0.305	0.32	0.000000003622084
16	1420 млн	умножение	-	все м/ф	0.31	0.315	0
17	1420 млн	сумма, трениров-ка	-	все м/ф	0.619	0.632	0

18	1420 млн	сумма, тренировка	-	прямые м/ф	0.523	0.726	0
19	1420 млн	сумма, тренировка	-	частотные м/ф	0.345	0.378	0.000000000001917
20	1420 млн	dilation	-	все м/ф	0.352	0.34	0
21	1420 млн	dilation	-	прямые м/ф	0.462	0.523	0
22	1420 млн	dilation	-	частотные м/ф	0.158	0.18	0.001116945411044
23	1420 млн	dilation, тренировка	-	все м/ф	0.282	0.299	0
24	1420 млн	dilation, тренировка	-	прямые м/ф	0.226	0.226	0.00000073566282