Д.А. Рыжова

Опыт автоматического построения анкеты для лексико-типологического исследования прилагательных и одноместных глаголов с помощью моделей дистрибутивной семантики[1]

В статье представлен метод автоматического составления анкеты для лексико-типологического исследования двух классов лексики: прилагательных и непереходных глаголов. В работе показано, что алгоритм применим к обоим классам слов, хотя на одном из этапов и нуждается в модификациях, вызванных различиями в частотности и дистрибутивных характеристиках рассматриваемых лексем.

Ключевые слова: лексическая типология, фрейм, типологическая анкета, прилагательные, глаголы, модели дистрибутивной семантики, кластеризация.

1. Фреймовый подход к лексической типологии

Подход к лексической типологии, на который мы будем опираться в настоящей статье[2], восходит к традициям Московской Семантической Школы и основывается на представлении о том, что семантика лексемы отражается на её дистрибуции[3]. Например, в сочетании с наименованиями разного рода контейнеров (*колодец, овраг, тарелка* и т. п.) прилагательное *глубокий* выражает параметрическую семантику большого размера, а в сочетании с существительными, обозначающими эмоции (*горесть, сожаление, обида* и др.), – значение интенсификатора.

Этот принцип обуславливает и формат типологических анкет для сбора данных различных языков. Каждая анкета представляет собой набор минимальных контекстов, в которых могут употребляться слова рассматриваемой семантической области. Для признаковой и глагольной лексики, о которой пойдет речь в настоящей работе, минимальный контекст – это определяемое существительное или набор актантов соответственно. Фрагмент анкеты для изучения прилагательных со значением ‘глубокий’ может выглядеть так:

Таблица 1. Фрагмент анкеты для признака 'глубокий'

Контейнеры	колодец
	овраг
	тарелка
Эмоции	горесть
	сожаление
	обида…

Правый столбец представляет собой собственно примеры контекстов. Поскольку опросник рассчитан на сбор типологических данных, предполагается, что каждая его строка – это не конкретное слово конкретного языка, а ситуация, которая за этим контекстом стоит. Так, например, при сборе данных английского языка исследователь должен будет понять, употребляется ли английский переводной эквивалент прилагательного *глубокий deep* в сочетании с существительными, обозначающими эмоции, и выражает ли он при этом значение интенсификации, т. е. действительно ли словосочетание употребляется в той же ситуации, что и соответствующее ему русское.

Левый столбец таблицы 1 указывает на то, что контексты в анкете-опроснике располагаются не хаотично, а собираются в группы[4]. Эти группы мы называем фреймами. Наш опыт показывает, что именно фреймы в результате становятся удобной базой для сравнения данных различных языков, поскольку переводные эквиваленты отличаются друг от друга тем, какие группы контекстов они способны покрывать, а какие – нет. Например, русскому прилагательному *тонкий* в китайском языке соответствуют две лексемы: *xì* и *báo*. При этом одна из них употребляется в сочетании с названиями длинных вытянутых предметов (*xì gùnzi* – 'тонкая палка'), а вторая описывает тонкие плоские предметы (*báo zhǐ* – 'тонкая бумага')[5]. Тем самым, для поля 'тонкий' релевантно различение фреймов длинных вытянутых vs. плоских объектов. Русское прилагательное *тонкий* их объединяет (равно как и английское *thin*, сербское *tanak* и др.), а в китайском языке (а также в хантыйском, кабардинском и др.) они разводятся лексически[6].

Приведенный пример показывает, что фреймовая структура поля определяется не априорно, а по результатам анализа языковых данных[7]: мы выделяем такие группы контекстов, на которые реагируют языки. Считается, что единый фрейм – это строки анкеты, которые всегда заполняются одинаково. Как только появляется язык, в котором эта группа разделяется лексически, вместо одного фрейма в этом фрагменте анкеты постулируются два разных.

С другой стороны, процесс выделения фреймов не бесконечен. Практика показывает, что многие противопоставления выявляются уже на основе трех-пяти языков (ср., например, статью М.В. Кюсевой

и Д.А. Рыжовой[8]), где фреймовая структура поля ‘острый’, подтвердившаяся впоследствии данными более двадцати языков различных групп и семей, определяется на материале русского, сербского и французского. В этом и заключается сложность и ответственность составления базового варианта анкеты для изучения того или иного семантического поля: необходимо не просто определить круг ситуаций, которые могут описываться лексемами этой зоны, но и выявить те параметры, которые могут оказаться важными для лексикализации рассматриваемого поля в различных языках.

В рамках фреймового подхода к лексической типологии работа по составлению предварительного варианта анкеты-опросника обычно выполняется вручную на материале русского языка и требует очень много времени и сил. Задача исследователя на этом этапе – определить, в каких контекстах могут появляться лексемы рассматриваемого поля, и разделить эти контексты на группы, т.е. постараться предугадать, какие правила сочетаемости окажутся релевантными для этой семантической зоны. Закономерности, которые важны уже для русского языка, выделяются с достаточной степенью уверенности. Те же правила, которые русским языком игнорируются, составляются гипотетически и затем аккуратно проверяются на материале других языков.

В настоящей работе мы покажем, как можно автоматизировать процесс составления предварительного варианта типологической анкеты для прилагательных и одноместных глаголов. В качестве иллюстрации будут использованы признаковые поля ‘острый’ и ‘прямой’ и глагольное поле качания.

2. Алгоритм построения лексико-типологической анкеты

Итак, задача составления предварительного варианта анкеты для лексико-типологического исследования сводится к двум подзадачам: 1) составление списка контекстов употребления лексем рассматриваемого поля; 2) разделение этих контекстов на смысловые группы (= фреймы), которые затем будут положены в основу правил сочетаемости и станут базой для сравнения данных различных языков.

2.1. Составление списка контекстов

Алгоритм, который мы предлагаем, разрабатывался и тестировался на материале признаковой лексики. Семантические поля именно этого типа были выбраны по нескольким причинам. Во-пер-

вых, несколько зон качественных признаков уже было обследовано вручную[9], так что результаты этого анализа можно было использовать в качестве золотого стандарта в процессе тестирования системы. Во-вторых, признаковые семантические поля отличаются сравнительно простой организацией: у прилагательных, как правило, только один актант, который в большинстве случаев и является его ключевым, «диагностирующим» контекстом. К тому же, определяемое слово в атрибутивной конструкции очень часто расположено контактно по отношению к определению, так что задача составления списка возможных контекстов фактически сводится к получению набора биграмм вида «прилагательное» + «существительное».

Задачу составления списка биграмм можно решать разными способами. В первую очередь, для нескольких языков, в том числе и для русского, существует коллекция биграмм Google[10]. Она выгодно отличается объемом собранной в ней информации: вероятность упустить какой-нибудь важный контекст или не набрать данных, достаточных для статистики, очень мала. С другой стороны, высокая степень полноты данных часто влечет за собой низкие показатели точности, из-за чего и в списки биграмм Google попадает очень много шума.

Другой источник двухсловных словосочетаний – биграммы Национального корпуса русского языка (далее – НКРЯ)[11]. Эта коллекция демонстрирует обратное соотношение полноты и точности: ресурс позволяет извлекать необходимые словосочетания, но, как правило, в очень ограниченных количествах.

Наш алгоритм опирается на материалы НКРЯ: мы собираем пары слов, состоящие из интересующих нас прилагательных и стоящих справа от них существительных, по основному подкорпусу НКРЯ, но увеличиваем общий объем данных за счет подключения лемматизации, что позволяет объединить несколько единичных примеров в одну более представительную группу. Затем из уже полученного списка мы удаляем все словосочетания, встретившиеся в корпусе менее 10 раз, чтобы избежать окказиональных употреблений. Такой метод позволяет получить представительный список словосочетаний, содержащий по несколько иллюстраций на каждый фрейм изучаемого поля, однако он ограничивает область исследования достаточно частотными прилагательными. Так, например, он применим для анализа сочетаемости русских лексем *острый* и *прямой*, занимающих позиции 1 452 и 892 соответственно по частотному словарю О.Н. Ляшевской и С.А. Шарова[12], но для изучения прилагательного *тугой* (ранг 7 283) предоставляемых им данных уже явно недостаточно.

Задача составления списка контекстов для глагольной лексики требует модификации процедуры. В настоящей статье мы рассмотрим в качестве примера одно поле: глаголы качания. Эта зона также была исследована вручную на материале нескольких языков[13], что позволит нам оценить качество работы алгоритма. Вслед за автором ручного анализа этой семантической зоны, мы будем рассматривать только один ее фрагмент: одноместные непереходные глаголы (ср. русск. *качаться, шататься, болтаться, колебаться, колыхаться, развеваться*). Тем самым, диагностический контекст для этих лексем будет таким же узким, как и для прилагательных. С другой стороны, в отличие от прилагательных, позиция которых в атрибутивной конструкции, как правило, четко закреплена (в русском языке прилагательное почти всегда находится слева от определяемого слова), глагол может располагаться как слева, так и справа от своего актанта, ср.: *Девочка качается на качелях* и *На качелях качается девочка*. Ещё одна особенность глаголов качания по сравнению с признаковыми словами, которые мы рассматривали, – их относительно низкий уровень частотности (самая частотная русская лексема в этой зоне – глагол *колебаться* – занимает позицию 4 802 по словарю О.Н. Ляшевской и С.А. Шарова, самая редкая – *развеваться* – 14 321).

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

2.2. Выделение фреймов

2.2.1. Модели дистрибутивной семантики

Разделение набранного списка контекстов на группы (будущие фреймы) – классическая задача кластеризации. Однако для того, чтобы можно было применить кластерный анализ, необходимо

определить основание для сравнения словосочетаний. Для наших целей важно, чтобы это основание было семантическим: мы хотим получить группы биграмм, близких по смыслу, описывающих похожие ситуации. Исходя из этих соображений, мы воспользовались аппаратом моделей дистрибутивной семантики[14].

Теория моделей дистрибутивной семантики, как и фреймовый подход к лексической типологии, зиждется на дистрибутивной гипотезе. В рамках этой теории значением языковой единицы (морфемы, слова, словосочетания или предложения) считается сумма контекстов, в которых она употребляется в некотором обучающем корпусе. Сумма контекстов представляется в виде вектора, измерениями которого, в простейшем случае, служат леммы или словоформы, а значением каждого измерения становится количество употреблений единицы, для которой создается вектор, в контексте данной единицы-измерения. При этом контекстом считаются слова, попадающие в окно фиксированного размера, т.е. находящиеся на определенном линейном или синтаксическом расстоянии от искомой языковой единицы.

С целью кластеризации списка словосочетаний, полученного на первом этапе работы алгоритма, мы построили для каждого элемента этого множества векторное представление. Были выбраны следующие параметры дистрибутивной модели: в качестве обучающей текстовой выборки использовался основной подкорпус НКРЯ; в качестве измерений выступали 10 000 наиболее частотных (для этого корпуса) знаменательных лексем; значением каждого измерения считалось количество случаев встречаемости слова-измерения на расстоянии не более пяти знаменательных слов влево или вправо от искомой лексической единицы.

Необходимо, однако, уточнить, что векторное представление для словосочетания можно построить двумя способами. С одной стороны, можно рассматривать словосочетание как единое целое и вычислять значения измерений по контекстам, в которых оно встречается. В этом случае исследователь неминуемо сталкивается с проблемой нехватки данных: словосочетания значительно менее частотные, чем слова, поэтому для качественного представления их сочетаемости нужны корпуса невероятных размеров. К тому же, для глагольной лексики мы вынуждены были бы собирать по два вектора для каждой пары субъекта и предиката, поскольку в русском языке актант может располагаться как слева, так и справа от глагола. С другой стороны, вектор словосочетания можно строить путем композиции векторов его элементов, т.е. сначала собирать отдельные вектора для каждого слова, а затем их объединять. Существует несколько стандартных моделей вычисления результирующего векторного

представления словосочетания на основе векторов его частей[15]. В нашем алгоритме используется одна из самых простых моделей композиции[16]: аддитивная взвешенная, поскольку в предыдущих исследованиях она стабильно демонстрировала хорошие результаты[17]. Эта схема композиции подразумевает сложение векторов прилагательного/глагола и существительного (т.е. попарное суммирование значений по каждому из измерений) с присвоением слагаемым некоторых весов. Значение весового коэффициента вычисляется на основе тренировочного корпуса – набора векторов соответствующих наблюдаемых словосочетаний.

2.2.2. Алгоритмы кластеризации

Теоретически для решения нашей задачи лучше всего подходят алгоритмы, определяющие итоговое число кластеров автоматически: предполагается, что исследователь изначально не знает, сколько фреймов будет в его анкете. Мы протестировали четыре алгоритма такого типа (Affinity Propagation[18], Mean Shift[19], DBScan[20] и алгоритм иерархической кластеризации[21]), из которых три (Affinity Propagation, Mean-shift и DBScan) дали неудовлетворительные результаты[22], поэтому дальнейшие эксперименты были продолжены только с алгоритмом иерархической кластеризации[23].

В ходе наших экспериментов мы провели кластеризацию тестовых данных с помощью этого метода и затем, с целью повышения степени однородности полученных групп и уменьшения уровня шума, мы удалили из каждого кластера периферийные элементы, оставив только по три словосочетания, максимально близких к ядру класса. При этом ядром (центроидом) кластера считался не конкретный объект (то или иное словосочетание), а усредненный вектор, в качестве значения каждого измерения которого выступало среднее арифметическое значений соответствующего измерения векторов всех словосочетаний, вошедших в данную группу. Три словосочетания, максимально близких к ядру, – это объекты, которые находятся ближе всего к центроиду по косинусной мере близости. Кластеры размером меньше трех элементов были совсем исключены из рассмотрения.

На последнем этапе исследования мы провели точную оценку качества работы алгоритма, сравнив итоговые варианты кластеризации с экспертной разметкой тех же словосочетаний. В качестве оценочной метрики мы использовали сбалансированную F-меру, позволяющую определить оптимальное соотношение полноты и точности: $F = 2PR / (P+R)$, где P – точность (т. е. чистота кластеризации), а R – полнота (т. е. доля всех фреймов данного поля, вошедших в результирующую анкету).

Для всех трех наборов тестовых данных значение F-меры оказалось достаточно высоким (см. таблицу 2). Визуальный анализ полученных анкет также показывает их информативность (ср. фрагмент одного из вариантов кластеризации признака *прямой* в таблице 3).

Таблица 2. Оценка качества работы алгоритма

Набор тестовых данных	**Полнота**	**Точность**	**F-мера**
‘острый’	0.882	0.898	0.89
‘прямой’	1.0	0.778	0.875
глаголы качания	0.882	0.762	0.818

Таблица 3. Фрагмент кластеризации контекстов лексемы ‘прямой’

Кластер 1	**Кластер 2**	**Кластер 3**	**Кластер 4**
прямой столб	прямое участие	прямой потомок	прямая необходимость
прямая дорожка	прямая поддержка	прямой предшественник	прямая цель
прямая аллея	прямое руководство	прямое наследие	прямая задача

3. Заключение

Алгоритм автоматического построения анкеты для типологического исследования признаковой и глагольной лексики, который мы предлагаем, состоит из нескольких этапов:

1. Составление списка существительных, с которыми могут сочетаться рассматриваемые прилагательные/одноместные глаголы (на материале основного подкорпуса НКРЯ);
2. Построение векторного представления для каждого словосочетания с помощью аддитивной модели композиции;
3. Кластеризация векторного пространства;
4. Выделение трех центральных элементов из каждого кластера.

Разработанный нами алгоритм демонстрирует достаточно высокую степень полноты и точности: во всех экспериментах лучшее значение F-меры выше 0.8. Наиболее низкое значение F-меры у поля глаголов качания, вероятно, связано с малой частотностью входящих в него лексем и, как следствие, недостаточно высоким качеством векторного представления словосочетаний и точности их кластеризации. Бороться с этой проблемой можно разными способами: например, увеличивая объем тренировочного корпуса или включая в рассмотрение приставочные дериваты анализируемых глаголов (ср. *закачаться, покачаться, раскачаться* и т. д.), т. е. мо-

дифицируя первые два этапа работы алгоритма. Заметим, однако, что только на этих стадиях при переходе от одного класса лексики к другому возникает необходимость разного рода изменений исходной процедуры: все дальнейшие процессы осуществляются по одним и тем же схемам для всех рассмотренных нами полей.

Таким образом, предлагаемый нами метод может использоваться в лексико-типологических исследованиях прилагательных и непереходных глаголов, существенно облегчая работу лексического типолога, сокращая количество ручного труда и позволяя добиться более объективного результата, не зависящего от изначальных установок и опыта исследователя.

Примечания

[1] Настоящее исследование поддержано грантом РФФИ № 14-06-00343а. Автор также выражает благодарность анонимному рецензенту за ценные замечания.

[2] См. *Рахилина Е.В., Резникова Т.И.* Фреймовый подход к лексической типологии // Вопросы языкознания. 2013. №2. С. 3–31.

[3] См., например, *Апресян Ю.Д.* Лексическая семантика: синонимические средства языка. М.: Наука. 1974, а также *Шайкевич А.Я., Андрющенко В.М., Ребецкая Н.А.* Дистрибутивно-статистический анализ языка русской прозы 1850—1870-х гг. Т. 1. М.: Языки славянской культуры, 2013.

[4] Однако это не означает, что вопросы носителю предъявляются также подряд. Техника применения подобной анкеты – это отдельная проблема, которую в рамках настоящей статьи мы затрагивать не будем.

[5] Подробнее см. *Кюсева М.В., Рыжова Д.А., Холкина Л.С.* Лексическая типология: к проблеме определения границ семантического поля (на примере признаков ‘толстый’ и ‘тонкий’) // Tipologia lexica. Гранада : Jizo Ediciones. 2013. С. 255–262.

[6] Таким образом, наши «фреймы» имеют несколько иную природу, нежели «фреймы» Ч. Филлмора (См. *Fillmore C.J.* Frame semantics. Linguistics in the Morning Calm. Seoul, South Korea: Hanshin Publishing Co. 1982. P. 111–137).

[7] Важно отметить, что и границы семантического поля мы определяем на основе языковых данных, считая относящимися к одной семантической зоне такие фреймы, которые во многих языках могут покрываться одним и тем же лексическим средством.

[8] *Кюсева М.В., Рыжова Д.А.* Прилагательные ‘острый’ и ‘тупой’ в русском, сербском и французском языках // Проблемы лексической типологии: Сборник научных трудов. Вып. 1. Воронеж, 2011. С. 164–171.

[9] См. *Кюсева М.В.* Лексическая типология семантических сдвигов названий качественных признаков ‘острый’ и ‘тупой’ : Дипломная работа / МГУ, филологический факультет, Отделение теоретической и прикладной лингвистики. М., 2012. А также: *Лучина Е.С.* Пути грамматикализации лексем со значением ‘прямой’ : Дипломная работа / МГУ, филологический факультет, Отделение теоретической и прикладной лингвистики. М., 2014.

[10] Именно этим ресурсом (Коллекция Google N-grams. URL: http://storage.googleapis.com/books/ngrams/books/datasetsv2.html) пользуются при составлении лексико-типологической анкеты Т.И. Резникова и Б.В. Орехов (см. *Орехов Б.В., Резникова Т.И.* Компьютерные перспективы лексико-типологических исследований // Вестник ВГУ. Серия: Лингвистика и межкультурная коммуникация. 2015. Вып. 3. С.17–23).

[11] Биграммы Национального корпуса русского языка. URL: http://ruscorpora.ru/search-ngrams_2.html.

[12] *Ляшевская О.Н., Шаров С.А.* Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М.: Азбуковник. 2009. URL: http://dict.ruslang.ru/freq.php.

[13] См.: *Шапиро М.М.* Глаголы колебательного движения в уральских языках (на материале финского, ненецкого и коми-зырянского языков): семантика и типология // Урало-алтайские исследования. 2015. №1 (16). С.29–52.

[14] См.: *Baroni M., Bernardi R., Zamparelli R.* Frege in Space: A Program for Compositional Distributional Semantics // Linguistic Issues in Language Technologies, Vol. 9. CSLI Publications. 2013. P. 5–110.

[15] См.: *Mitchell J.* Composition in distributional models of semantics // Cognitive science. 2010. 34(8). P. 1388–1429. Для языка программирования Python все модели композиции представлены в библиотеке DISSECT (URL: http://clic.cimec.unitn.it/composes/toolkit/index.html), см. также: *Dinu G., Pham N.T., Baroni M.* DISSECT: DIStributional SEmantics Composition Toolkit // Proceedings of ACL (System Demonstrations). Sofia, Bulgaria. 2013. P. 31–36.

[16] Итоговое векторное пространство подвергается дополнительной обработке: взвешиванию (по схеме PPMI – Positive Point-wise Mutual Information) и уменьшению размерности (SVD). Обоснование выбора этих параметров не связано напрямую с темой настоящей статьи, поэтому в рамках данной работы мы на нем не останавливаемся. Подробнее о подборе параметров моделей для решения задач в области типологии качественных признаков см.: *Ryzhova D., Kyuseva M., Paperno D.* Typology of Adjectives Benchmark for Compositional Distributional Models // Proceedings of the Language Resources and Evaluation Conference. 2016. P. 1253–1257. Все дополнительные операции над векторами также производились с помощью пакета DISSECT.

[17] См., напр.: *Кюсева М.В.* Верификация фреймового подхода к лексической типологии с помощью векторных моделей : Выпускная квалификационная работа / НИУ ВШЭ, факультет филологии. М., 2014. А также: *Ryzhova D., Kyuseva M., Paperno D.* Typology of Adjectives Benchmark for Compositional Distributional Models // Proceedings of the Language Resources and Evaluation Conference, 2016. P. 1253–1257.

[18] См., напр.: *Frey B.J., Dueck D.* Clustering by Passing Messages Between Data Points // Science. Feb. 2007. Vol. 315, Issue 5814. P. 972-976.

[19] См.: *Comaniciu D., Meer P.* Mean Shift: A robust approach toward feature space analysis // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2002. P. 603–619.

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

[21] См., напр.: *Johnson S.C.* Hierarchical clustering schemes // Psychometrika. 1966. № 32(2). P. 241–54.

[22] Подробнее см.: *Рыжова Д.А.* Построение лексико-типологической анкеты с помощью моделей дистрибутивной семантики : Выпускная квалификационная работа / НИУ ВШЭ, факультет филологии. М., 2014.

[23] Алгоритм был реализован с помощью пакета программ SciPy. URL: http://docs.scipy.org/doc/scipy/reference/cluster.hierarchy.html#module-scipy.cluster.hierarchy.