УДК 81’37

ПОСТРОЕНИЕ ЛЕКСИКО-ТИПОЛОГИЧЕСКОЙ АНКЕТЫ С ПОМОЩЬЮ МОДЕЛЕЙ ДИСТРИБУТИВНОЙ СЕМАНТИКИ

Д. А. Рыжова

Национальный исследовательский университет «Высшая школа экономики»

Поступила в редакцию 20 апреля 2015 г.

Аннотация: *статья посвящена разработке алгоритма автоматического построения анкеты для типологического исследования признаковой лексики. Предлагаемый метод включает несколько этапов: сбор списка существительных, с которыми сочетается изучаемое прилагательное; создание векторов сочетаемости для каждого полученного словосочетания; кластеризация векторного пространства; извлечение из каждого кластера трех ядерных элементов. Достаточно высокое качество работы алгоритма показывает, что задача автоматизации такого рода исследовательских процессов выполнима, и описанный в статье метод – одно из возможных ее решений.*
Ключевые слова: *семантика, лексическая типология, анкета, дистрибутивные модели, кластеризация.*

Abstract: *this article describes the algorithm of automatic construction of a questionnaire for a lexical typological research of adjectives. The method includes four steps: collecting a list of nouns combining with the adjective in question; computing a co-occurrence vector for every noun phrase from the list; clustering the vector space; extracting three core elements from every cluster. A rather high quality of the algorithm’s performance demonstrates that the task can be completed, and the method presented in the article is one of the options.*
Key words: *semantics, lexical typology, questionnaire, distributional models, clustering.*

Исследования в области лексической типологии всегда трудоемки. Даже если мы оставим в стороне долго господствовавшие в лингвистике представления о том, что лексика – это хаос, не подлежащий систематизации и теоретическому осмыслению, мы вынуждены будем согласиться по крайней мере с тем, что процесс изучения этого уровня языка связан с целым рядом затруднений. Среди них самая серьезная проблема – это задача сбора данных. Информацию о значениях слов нельзя почерпнуть из грамматик, а словари в большинстве случаев неполны и недостаточно точны (особенно это касается «экзотических» языков, которые, конечно, представляют большой интерес для типологов). Еще меньше можно рассчитывать на корпуса: несмотря на бурное развитие корпусной лингвистики в последние годы, коллекции размеченных текстов достаточного объема (а для анализа лексики нужны значительно более крупные корпуса, чем для грамматических исследований) существуют лишь для очень ограниченного количества языков, преимущественно крупных европейских.

Наиболее эффективный метод сбора информации в такой ситуации – анкетирование носителей. Именно на этот метод опирается группа Института имени Макса Планка в Неймегене, чей подход к лексической типологии в настоящее время является наиболее популярным. Свои анкеты приверженцы этого подхода строят на основе экстралингвистических стимулов: глаголы разрушения объекта изучаются с помощью видеоклипов [1], прилагательные вкуса – с помощью растворов, которые информанту дают попробовать [2], лексемы, описывающие запахи, – на основе флаконов с разными химическими смесями, которые информанту предлагают понюхать [3]. Такие анкеты, с одной стороны, служат единой базой для сравнения данных различных языков, а с другой стороны, позволяют относительно быстро и легко собирать материал: их не надо переводить с языка на язык и адаптировать к реалиям конкретных культур.

Однако этот метод связан и с рядом ограничений. Во-первых, не всякую лексическую зону можно так изучить: например, для оценочных прилагательных (ср. *хороший, плохой, кошмарный, фантастический*) или для предикатов ментальных процессов (*думать, размышлять, соображать*) составить анкету на основе экстралингвистических стимулов будет крайне затруднительно. Во-вторых, такой подход не позволяет целенаправленно исследовать переносные значения слов. Наконец, в-третьих, этот метод психолингвистический по своей природе, и его разработчики ставят перед собой особые задачи: их интересуют в первую очередь особенности человеческого воспри-

ятия и их взаимосвязь с языком как одной из когнитивных способностей, а не устройство естественного человеческого языка как такового.

Более логичным с собственно языковой точки зрения представляется решение проводить типологические исследования с помощью анкетирования, но в основу анкет закладывать лингвистические принципы. Фреймовый подход к лексической типологии [4], на который мы будем опираться в настоящей статье, в качестве такого принципа выбирает дистрибутивную гипотезу, согласно которой значение слова можно определить путем анализа его сочетаемости, без непосредственного обращения к экстралингвистической действительности. Однако и этот метод сопряжен со своего рода осложнениями: в частности, составление анкеты такого типа – очень непростая задача. В настоящей статье мы предложим один из возможных способов решения этой проблемы: представим алгоритм автоматического построения анкеты для лексико-типологического исследования.

Фреймовый подход к лексической типологии

В основе подхода, в рамках которого мы работаем, лежит представление о значении лексемы как о наборе контекстов, в которых она может выступать. Эта идея интуитивно понятна: всякий, кто когда-либо изучал иностранный язык, обращал внимание, что нередко значения незнакомых слов ему удается просто вывести из контекста. В последние десятилетия дистрибутивная гипотеза неоднократно оказывалась в центре тех или иных лингвистических теорий: на ней базируется Грамматика Конструкций [5], она же еще в 70-х гг. XX в. высказывалась Ю. Д. Апресяном ([6]) и была подхвачена адептами Московской семантической школы (МСШ) (см., например, [7]).

Наш подход к лексической типологии восходит к традициям МСШ. Он был разработан Московской лексико-типологической группой под руководством Е. В. Рахилиной и получил название фреймового. Один из ключевых его постулатов гласит: значение лексемы удобнее представлять не в виде традиционных толкований, набора семантических компонентов или возможных денотатов, а в виде правил сочетаемости. Проиллюстрируем это положение на примере прилагательного *тонкий*.

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

– *тонкий* + названия длинных вытянутых предметов = малый диаметр поперечного среза;

– *тонкий* + обозначения звука = слабая громкость и высокий частотный диапазон и т.п.[1]

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

– ‘тонкий’ + название длинного вытянутого предмета => *xì* (*xì gùnzi* – ‘тонкая палка’);

– ‘тонкий’ + название плоского предмета => *báo* (*báo zhǐ* – ‘тонкая бумага’) и т.д. (подробнее см. [9]).

Наконец, важно отметить, что для каждого семантического поля этот набор правил ограничен: выделяется несколько минимальных групп ситуаций, которые каждый язык по-своему компонует, подчеркивая одни границы и стирая другие. Именно эти группы ситуаций (они же группы контекстов) мы называем фреймами, и задача лексического типолога сводится к выявлению набора фреймов, релевантных для изучаемого семантического поля, и описанию стратегий их объединения в рамках лексемы.

Такой метод изучения семантики слов был опробован на обширном лексическом материале [10–13]. В ходе проведенных исследований доказано, что этот подход действительно позволяет аккуратно анализировать лексемы как в рамках одного языка, так и на типологическом уровне, составляя простые правила соответствия переводных эквивалентов друг другу на основе их сочетаемостных свойств.

Так же, как и для исследователей, работающих в рамках психолингвистической парадигмы Института имени Макса Планка, основным инструментом сбора языковых данных для приверженцев фреймового подхода является анкета-опросник. Разница заключается в том, что эта анкета состоит не из экстралингвистических стимулов, а из контекстов, в которых

[1] Ср., например, попытку использовать такого рода правила для снятия семантической омонимии в Национальном корпусе русского языка (см. [8]).

могут быть употреблены лексемы изучаемого поля. Таким образом, фреймовый подход также позволяет решить проблему отсутствия для большинства языков необходимых информационных ресурсов, но при этом он не предполагает выхода за рамки собственно лингвистических данных и, по-видимому, не устанавливает ограничений на предмет исследования.

Однако, как уже было сказано выше, и эта теория не лишена недостатков. Помимо того, что анкету необходимо каждый раз переводить на анализируемый язык, само по себе составление исходного списка релевантных для рассматриваемого поля контекстов – большая работа, требующая много времени и внимания. Обычно она выполняется вручную на материале русского языка. Задача лексического типолога на этапе составления предварительного варианта анкеты – 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 этого поля. 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.

Таким образом, задача составления первого варианта анкеты сводится в двум подзадачам:

1) составить список контекстов употребления лексем рассматриваемого поля;

2) разделить эти контексты на смысловые группы (= фреймы), которые затем будут положены в основу правил сочетаемости и станут базой для сравнения языков.

Именно эти процессы должен уметь автоматизировать алгоритм построения лексико-типологической анкеты.

Алгоритм построения лексико-типологической анкеты

Составление списка контекстов

Алгоритм, который мы предлагаем в настоящей работе, разрабатывался и тестировался на материале признаковой лексики. Семантические поля именно этого типа были выбраны по нескольким причинам. Во-первых, несколько зон качественных признаков уже было обследовано нами вручную, 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 очень часто расположено контактно по отношению к определению, так что задача составления списка возможных контекстов фактически сводится к извлечению 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[2]. 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 (НКРЯ)[3]. Эта коллекция демонстрирует обратное соотношение полноты и точности: ресурс позволяет извлекать необходимые словосочетания, но, как правило, в очень ограниченных количествах.

В рамках настоящей работы мы принимаем компромиссное решение: собираем пары слов, состоящие из интересующих нас прилагательных и стоящих справа от них существительных, по основному подкорпусу НКРЯ (т.е. фактически опираемся на более аккуратные биграммы НКРЯ), но расширяем полученные данные 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 русской лексемы *густой*, занимающей позицию 2464 по частотному словарю О. Н. Ляшевской и С. А. Шарова[4], но для изучения прилагательного *тугой* (ранг 7283) предоставляемых им данных уже явно недостаточно.

Выделение фреймов

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 основания

[2] Именно этим ресурсом (http://storage.googleapis.com/books/ngrams/books/datasetsv2.html) пользуются при составлении анкеты Т. И. Резникова и Б. В. Орехов [14].

[3] URL: http://ruscorpora.ru/search-ngrams_2.html

[4] URL: http://dict.ruslang.ru/freq.php

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 (см. [15]).

Теория моделей дистрибутивной семантики, как и фреймовый подход к лексической типологии, зиждется на дистрибутивной гипотезе. В рамках этой теории значением лексической единицы считается сумма контекстов, в которых она употребляется в некотором обучающем корпусе. Сумма контекстов представляется в виде вектора, измерения которого, как правило, – слова. Значение каждого измерения – количество употреблений лексемы, для которой создается вектор, в контексте данного слова. При этом контекстом считаются все слова, попадающие в окно фиксированного размера, т.е. находящиеся на определенном расстоянии слева или справа от основной лексической единицы.

С целью кластеризации списка словосочетаний, полученного на первом этапе работы алгоритма, мы построили для каждого элемента этого множества векторное представление. Вектора создавались по следующим параметрам: 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. С другой стороны, вектор словосочетания можно строить путем композиции векторов его элементов, т.е. сначала собирать отдельные вектора для прилагательного и для существительного, а затем их объединять. Существует несколько стандартных моделей вычисления результирующего векторного представления словосочетания на основе векторов его частей (см. [16])[5]. В нашем алгоритме используется одна из самых простых моделей композиции – аддитивная взвешенная, поскольку в наших предыдущих исследованиях она стабильно демонстрировала хорошие результаты [18][6]. Эта схема композиции подразумевает сложение векторов прилагательного 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.

Алгоритмы кластеризации

Для решения нашей задачи было бы удобно воспользоваться алгоритмом, который определял бы итоговое число кластеров автоматически: предполагается, что исследователь изначально не знает, сколько фреймов будет в его анкете. Однако ряд экспериментов с алгоритмами кластеризации такого типа дал неудовлетворительные результаты. В связи с этим было принято решение провести серию экспериментов с алгоритмами, требующими изначального указания числа кластеров. Количество кластеров в каждом эксперименте определялось следующим образом: вычислялась сумма числа значений всех прилагательных, относящихся к рассматриваемому семантическому полю (по Малому академическому словарю[7]), которая затем умножалась на два. Удваивание суммы значений делает общее количество кластеров более независимым от одного конкретного словаря: во-первых, наш опыт показывает, что фреймы часто оказываются более дробными, чем словарные значения; а во-вторых, надежнее получить заведомо большее количество кластеров и удалить лишнее при последующей обработке.

Для кластеризации с помощью алгоритмов без автоматического определения количества кластеров мы использовали пакет программ Cluto[8]. Этот тулкит предлагает несколько методов кластерного анализа (метод ‘rb’ – repeated bisections; ‘graph’, при котором из исходного пространства сначала строится граф, и др.).

В ходе наших экспериментов мы провели кластеризацию тестовых данных с помощью всех доступных методов. Полученные результаты были неоднозначны. Так, наравне с достаточно однородными группами словосочетаний (ср., например, один из кластеров для прилагательного *жаркий*: *жаркий бой, жаркая схватка, жаркое сражение, жаркая пере-*

[5] Все они представлены в тулките DISSECT (см. [17]).

[6] Итоговое векторное пространство подвергается дополнительной обработке: взвешиванию (по схеме ppmi – Positive Point-wise Mutual Information) и уменьшению размерности. 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. [18]. Все дополнительные операции над векторами также проводились с помощью тулкита DISSECT.

[7] URL: http://feb-web.ru/feb/mas/mas-abc/default.asp

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

Таблица

Фрагмент кластеризации контекстов лексемы «прямой»

Кластер 1	Кластер 2	Кластер 3	Кластер 4
Прямой столб	Прямой репортаж	Прямой потомок	Прямой умысел
Прямая дорожка	Прямая трансляция	Прямой предшественник	Прямая измена
Прямая аллея	Прямой номер	Прямое наследие	Прямое предательство

стрелка) выделялись и бессмысленные классы из коллокаций, иллюстрирующих множество различных значений прилагательного (ср. один из кластеров для лексемы *прямой*: *прямая линия*, *прямая наводка*, *прямой путь*, *прямой эфир*, *прямой удар*, *прямой участок*, *прямое направление*, *прямая кишка*).

С целью повышения степени однородности полученных групп мы удалили из каждого кластера периферийные элементы, оставив только по три словосочетания, максимально близких к ядру класса. Эта модификация дала желаемые результаты (ср. фрагмент одного из вариантов кластеризации признака *прямой* в таблице).

На последнем этапе исследования мы провели оценку качества работы алгоритма, сравнив полученные варианты кластеризации с экспертной разметкой тех же словосочетаний. Лучшее значение F-меры для лексемы *прямой*, которое нам удалось получить, равнялось 0,89 (алгоритм кластеризации ‘rbr’); для группы прилагательных семантического поля ‘острый’ – 0,943 (алгоритм ‘rb’).

Таким образом, алгоритм автоматического построения анкеты для типологического исследования признаковой лексики, который мы предлагаем, состоит из нескольких этапов.

1. Составление списка существительных, с которыми могут сочетаться рассматриваемые прилагательные (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.

Этот алгоритм демонстрирует достаточно высокую степень полноты и точности: в разных экспериментах значение F-меры достигает 0,9. Таким образом, проведенное нами исследование показывает, что автоматическое составление анкеты для лексико-типологического исследования возможно, и для решения этой задачи может использоваться разработанный нами метод.

Безусловно, многие вопросы на данном этапе остались без ответа: непонятно, как должен быть устроен список контекстов для других частей речи; не выявлено, какой из алгоритмов кластеризации лучше всего подходит для наших целей; неясно, как анализировать сочетаемость не очень частотных лексем. Однако первый шаг на пути автоматизации процессов лексико-типологического исследования сделан, и если все последующие шаги будут столь же успешными, работу лексического типолога можно будет существенно облегчить и ускорить, а значит, и увеличить количество научных достижений в этой области лингвистики.

ЛИТЕРАТУРА

1. *Majid A.* Cutting and breaking events : A crosslinguistic perspective / A. Majid and M. Bowerman (eds.) // Cognitive Linguistics [Special Issue] – 2007. – № 18 (2).

2. *Majid A.* Language does provide support for basic tastes [Commentary on A study of the science of taste : On the origins and influence of the core ideas by Robert P. Erickson] / A. Majid and S. C. Levinson // Behavioral and Brain Sciences. – 2008. – № 31. – P. 86–87.

3. *Wnuk E.* Revisiting the limits of language : The odor lexicon of Maniq / E. Wnuk and A. Majid // Cognition. – 2014. – № 131. – P. 125–138.

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

5. *Goldberg A.* Constructions. A Construction Grammar Approach to Argument Structure / A. Goldberg. – Chicago, London : The University of Chicago Press, 1995. – 271 p.

6. *Апресян Ю. Д.* Избранные труды. Т. 1. Лексическая семантика (синонимические средства языка) / Ю. Д. Апресян. – М. : Языки русской культуры, 1995. – 472 с.

7. Языковая картина мира и системная лексикография / Ю. Д. Апресян [и др.]. – М. : Языки славянских культур, 2006. – 912 с.

8. *Шеманаева О. Ю.* Семантические фильтры для разрешения многозначности в Национальном корпусе русского языка : прилагательные / О. Ю. Шеманаева [и др.] // Компьютерная лингвистика и интеллектуальные технологии : труды Междунар. конф. «Диалог–2007» (Бекасово, 30 мая – 3 июня 2007 г.) / под ред. Л. Л. Иомдина [и др.]. – М. : Изд-во РГГУ, 2007. – С. 582–587.

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

10. Концепт БОЛЬ в типологическом освещении / В. М. Брицын [и др.]. – Киев : Видавничий Дім Дмитра Бураго, 2009. – 424 с.

11. *Майсак Т. А.* Глаголы движения в воде : лексическая типология / Т. А. Майсак, Е. В. Рахилина (ред.). – М. : Индрик, 2007. – 752 с.

12. *Круглякова В. А.* Семантика глаголов вращения в типологической перспективе : дис. ... канд. филол. наук / В. А. Круглякова. – М. : РГГУ, 2010.

13. *Кашкин Е. В.* Языковая категоризация фактуры поверхностей (типологическое исследование наименований качественных признаков в уральских языках) : дис. ... канд. филол. наук / Е. В. Кашкин. – М. : МГУ, 2013.

14. *Орехов Б. В.* Компьютерные перспективы лексико-типологических исследований / Б. В. Орехов, Т. И. Резникова // Вестник Воронеж. гос. ун-та. Сер. : Лингвистика и межкультурная коммуникация. 2015. № 3.

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

16. *Mitchell J.* Composition in distributional models of semantics / J. Mitchell, M. Lapata // Cognitive science. – 2010. – № 34 (8). – P. 1388–1429.

17. *Dinu G.* DISSECT : DIStributional SEmantics Composition Toolkit / G. Dinu, N. T. Pham, M. Baroni // Proceedings of ACL (System Demonstrations). – Sofia, Bulgaria, 2013. – P. 31–36.

18. *Ryzhova D.* Typology of Adjectives as a Benchmark for Compositional Distributional Models / D. Ryzhova, M. Kyuseva, D. Paperno // To appear.

Национальный исследовательский университет «Высшая школа экономики»

Рыжова Д. А., аспирант

E-mail: daria.ryzhova@mail.ru

National Research University «Higher School of Economics»

Ryzhova D. A., Post-graduate Student

E-mail: daria.ryzhova@mail.ru