УДК 81

КОМПЬЮТЕРНЫЕ ПЕРСПЕКТИВЫ ЛЕКСИКО-ТИПОЛОГИЧЕСКИХ ИССЛЕДОВАНИЙ

Б. В. Орехов, Т. И. Резникова

Научно-исследовательский университет «Высшая школа экономики»

Поступила в редакцию 18 мая 2015 г.

Аннотация: *в статье обсуждается методика, позволяющая частично автоматизировать анализ материала для лексико-типологического исследования. Одной из основных задач при сопоставительном изучении семантического поля является обнаружение таких значений, для которых наблюдаются разные стратегии кодирования в различных языках (ср. значение 'характеризующийся повышенной влажностью': в русском его выражение зависит от описываемых температурных условий – влажный vs. сырой, в немецком в обеих ситуациях используется прилагательное feucht). Такие расхождения можно проследить, сопоставляя сочетаемость слов исследуемого поля. Для прилагательных сочетаемость, как правило, задается определяемым существительным, которое обычно непосредственно примыкает к атрибуту. Это означает, что прилагательные разных языков можно сравнить на материале двусловных сочетаний (биграмм). Такое сравнение предлагается проводить автоматически, используя биграммы, предоставляемые корпорацией Google, и электронные словари. Реализация метода обсуждается на примере семантического поля 'густой' в русском, английском и немецком языках.*
Ключевые слова: *лексическая типология, семантика прилагательных, лексическая сочетаемость, фрейм, компьютерная лингвистика.*

Abstract: *the article develops a technique for partial automation of data analysis in lexical typological studies. One of the main tasks in comparative semantic research is to reveal those meanings that may exhibit different coding strategies in various languages (cf. the meaning 'having high humidity', which opposes different temperature conditions in Russian, i.e. 'vlazhny' for warm environment vs. 'syroj' for cold environment. In contrast, German expresses both situations with a single term 'feucht'). Distinctions of this kind can be detected by comparing the combinability of words under study. The combinatorial properties of adjectives are mostly determined by the adjacent nouns. This means that adjectives from different languages can be compared based on the bigrams analysis. We suggest automation of this analysis by using Google bigrams and machine-readable dictionaries. The technique is illustrated through the adjectives of the semantic domain 'dense' in Russian, English, and German.*
Key words: *lexical typology, semantics of adjectives, lexical combinability, frame, computer linguistics.*

1. Лексическая типология: расширение инструментария

С развитием корпусных технологий перед лингвистикой в целом и лексической типологией, в частности, открылись новые перспективы, связанные с расширением эмпирической и методологической базы исследований. В докорпусную эпоху межъязыковое сравнение вынужденно ограничивалось семантическими зонами, которые можно было описать, предъявляя информантам конкретные предметы или их изображения (отсюда особая популярность таких наглядных и имеющих видимость четкой структуры зон, как части тела или цвета, ср. [1–3]).

Эта методика, опирающаяся на экстралингвистические стимулы, имеет, с одной стороны, очевидные преимущества. Дело в том, что большинство лексико-типологических исследований имеют ономасиологическую направленность: предметом анализа является то, как смыслы «упаковываются» в слова в каждом конкретном языке. Тем самым, чтобы сравнение получилось корректным, исследователь в разных языках должен работать с одними и теми же смыслами. Между тем означаемое, в отличие от означающего, – ненаблюдаемая сущность, так что идентичность сравниваемых смыслов требует надежных подтверждений. Таким подтверждением как раз и являются экстралингвистические стимулы, предъявляемые информанту в ходе эксперимента. Если носители разных языков описывают одни и те же предметы или изображения, то можно считать, что они тем самым выражают одни и те же значения. Благодаря своей методологической прозрачности этот подход продолжает оставаться востребованным и на нынешнем этапе развития лексической типологии, когда уже появились альтернативные способы сравнительного изучения значений. В частности, он широко применяется в исследованиях Института имени Макса Планка в Неймегене – на сегодняшний день одного

из крупнейших центров, занимающихся межъязыковым анализом лексики (ср., например: [4–6]).

Однако описываемый экспериментальный подход имеет и существенные ограничения. Наиболее значимым из них является то, что он не применим к семантическим полям, которые не поддаются непосредственной перцептивной верификации, как, например, зона боли или эмоций [7]. Более того, зачастую даже те концептуальные области, где работа с экстралингвистическими стимулами в принципе возможна, оказываются недоступны для экспериментального изучения – в силу субъективности восприятия соответствующих стимулов. Так, несложно представить себе, как можно организовать исследование для семантического поля ‘мягкость’ / ‘твердость’: нужно предъявлять носителям объекты разной степени жесткости и просить их описать воспринимаемый признак словами своего языка. Но сложность здесь будет состоять в том, что предметы, кажущиеся мягкими одному информанту, другим могут восприниматься как твердые (что связано с градуальностью проявления признака), так что результаты, полученные из разных языков, едва ли будут сопоставимы друг с другом. Тем самым значительный пласт лексики оказывается «скрыт» от экспериментальной методики.

С появлением больших массивов текстовых данных ограничений на языковой материал больше нет: объектом типологического анализа могут становиться любые семантические поля (ср. исследование лексики боли [8] или субъективно воспринимаемых свойств объектов [9; 10]). Дело в том, что методика в этом случае опирается на употребление слов в контексте, т.е. на их сочетаемость, соответственно, если лексемы определенного поля встречаются в корпусе достаточное количество раз (обычно для анализа требуется не менее сотни употреблений), то межъязыковое сопоставление уже становится возможным, вне зависимости от степени абстрактности изучаемых значений. Правда, при корпусном подходе исследователь не располагает изначально заданной основой для сравнения, как это имеет место в случае использования экстралингвистических стимулов, – эту основу ему только предстоит определить.

В качестве такой основы в методологии Московской лексико-типологической группы используются фреймы – минимальные ситуации, которые могут лексически противопоставляться в естественном языке [11]. Так, к фреймам семантического поля ‘мокрый’ относятся, например, ‘ткани: мокрые от контакта с водой’, ‘ткани: мокрые от контакта с другой жидкостью’, ‘ткани: не до конца высохшие’, ‘воздух с повышенной влажностью’, ‘воздух в закрытом помещении с повышенной влажностью’, ‘предметы, хранившиеся в помещении с повышенной влажностью’ и др. Выделение этих фреймов основано на лексических противопоставлениях в том или ином языке, например: ‘ткани: мокрые от контакта с водой’ и ‘ткани: мокрые от контакта с другой жидкостью’ описываются разными прилагательными в венгерском (соответственно, *vizes* vs. *nezves*), ‘воздух с повышенной влажностью’ и ‘воздух в закрытом помещении с повышенной влажностью’ противопоставляются в русском (*влажный* vs. *сырой*) и т.д.

Чтобы обнаружить подобные лексические противопоставления в пределах семантического поля (т.е., по сути, найти такие ситуации, которые описываются в некотором языке одним словом, а в другом языке – разными), необходимо провести подробный сравнительный анализ корпусных данных. Заметим, что этот этап работы сопряжен с серьезными временными затратами, поскольку предполагает поиск тонких семантических различий как в пределах одного языка, так и в типологической перспективе. Однако после того, как список фреймов сформирован, наиболее трудоемкий этап лексико-типологического проекта можно считать завершенным: дальнейшая работа связана с экстенсивным накоплением материала – готовые фреймы предъявляются информантам и заполняются лексическими данными.

Тем самым, если можно было бы ускорить начальный процесс выделения фреймов, то вся процедура исследования занимала бы существенно меньшее время, и, соответственно, развитие лексической типологии, пока еще остающейся на периферии общей лингвистической типологии, могло бы выйти на качественно новый уровень.

Ускорить этап выявления фреймов, как представляется, можно при помощи методов компьютерной лингвистики. Действительно, текущее состояние источников данных позволяет до некоторой степени автоматизировать этот процесс, привлекая существующие электронные ресурсы и программы обработки текстовой информации. В настоящей статье мы предлагаем методику, которая призвана упростить типологическое изучение признаковой лексики.

2. Компьютерные ресурсы в лексической типологии

Наиболее естественно в качестве материала для лексико-типологического исследования было бы использовать параллельные корпуса – электронные коллекции текстов, в которых эквивалентные отрезки на разных языках сопоставлены друг другу. Такие корпуса должны были бы выявлять особенности сочетаемости лексических единиц, отражать различия в употреблении переводных эквивалентов и, таким образом, помогать в определении границ фреймов для изучаемого поля.

И действительно, сравнительный анализ лексики разных языков иногда проводится на материале параллельных корпусов. Прежде всего следует назвать

исследование, посвященное глаголам движения в переводах Евангелия от Марка на 100 языков [12]. Объектом сопоставления в этой работе стали 360 текстовых фрагментов, описывающих различные ситуации движения. Параллельный корпус показывал расхождения в употреблении глагольных единиц: в частности, выявлял такие случаи, когда один язык лексически противопоставляет две ситуации, которые в другом языке задаются одной и той же лексемой.

Однако данное исследование в определенном отношении является исключительным. Дело в том, что базовые глаголы движения относятся к частотной лексике, так что даже на небольшом по объему корпусе этот лексический материал может послужить основой для построения некоторых типологических обобщений. Вместе с тем преобладающая часть лексического состава языка встречается в текстах существенно реже, чем глаголы движения, и для ее изучения объемы существующих на сегодняшний день параллельных корпусов явно недостаточны. Так, русско-английский параллельный корпус Национального корпуса русского языка (далее – НКРЯ) содержит чуть более 20 млн словоупотреблений. Ориентированный на славянские языки ParaSol на март 2014 г. насчитывал 27 млн токенов для 31 языка. Немецко-английский European Parliament Proceedings Parallel Corpus чуть больше по объему – в нем свыше 47 млн слов, но в то же время образующие его тексты весьма специфичны – это законодательные акты: они отличаются своеобразным лексическим наполнением и, соответственно, плохо подходят для изучения общеупотребительной лексики. Заметим к тому же, что и при более стандартном лексическом составе текстов объем в 47 млн словоупотреблений для семантического анализа все равно был бы недостаточен (показательно, что в случае внутриязыковых исследований специалисты по русской семантике иногда не удовлетворяются даже основным корпусом НКРЯ, содержащим около 230 млн токенов, предпочитая пользоваться ресурсом RuTenTen, объем которого превышает 18 млрд слов). Таким образом, установление границ фреймов на типологическом материале нескольких языков при помощи параллельных корпусов можно считать только делом будущего.

Однако, как нам представляется, и в отсутствие больших параллельных корпусов частичная автоматизация процесса выявления фреймов все же возможна. Для решения этой задачи в отношении признаковой лексики мы предлагаем использовать наборы биграмм, составленные корпорацией Google для 8 языков (английский, русский, немецкий, французский, итальянский, испанский, китайский, иврит) на основе текстов книг, оцифрованных в рамках проекта Google Books. Биграммы размещены в интернете для свободного использования (Google предоставляет доступ и к наборам n-грамм, где n больше двух, но для нашего исследования мы пользовались только данными по биграммам).

Наборы биграмм представляют собой частотные списки пар словоформ, следующих в текстах одна за другой (отметим, что в коллекции отражены только те биграммы, которые встретились в корпусе не менее 40 раз). Этот ресурс, соответственно, оказывается чрезвычайно ценным источником материала для исследования семантики прилагательных. Дело в том, что особенности значения слова, как известно, проявляются в его сочетаемости. Для прилагательных сочетаемость в общем случае полностью задается существительным, которое выступает при нем в качестве определяемого (об атрибутивной позиции как более показательной для семантики прилагательного, чем предикативная, см. [13–15]). Между тем в атрибутивной группе, как правило, определение непосредственно примыкает к определяемому, тем самым извлекая из общего массива биграмм те пары, в которых на первом (в случае языков с препозитивным определением) или на втором месте (для языков с постпозицией атрибутива) стоит искомое прилагательное, мы с большой вероятностью получим как раз его сочетания с определяемым существительным. Возможности применения биграмм, извлеченных таким образом, в лексической типологии мы обсудим на материале прилагательных семантического поля ГУСТОЙ в русском, немецком и английском языках.

3. Сопоставление биграммных наборов: описание методики

Для тестирования методики, основанной на автоматическом анализе биграмм, нами были выбраны прилагательные *густой* для русского языка, *dicht, dick, dickflьssig, buschig* – для немецкого и *thick, dense, bushy* – для английского.

Последовательность обработки материала была следующей. Исходные данные разбиты программистами корпорации Google на отдельные файлы-порции, в каждой из которых хранятся только те биграммы, у которых совпадают первые два символа (например, в одном файле содержатся все биграммы, начинающиеся на буквосочетание *aa*, в другом – на *ab*, в третьем – на *ac* и т.д.). Таким образом, мы работали с файлами *gu* для русского, *di* и *bu* для немецкого и *bu, de, ti* для английского.

Из этих наборов биграмм были выделены такие, в которых в препозиции стояли интересующие нас прилагательные. Заметим, что поскольку файлы с биграммами достаточно велики по объему, то операция по их обработке и извлечению нужных нам словосочетаний занимает достаточно большое время. Более того, нужно учитывать, что в полученных биграммных наборах вторым элементом после прилагательного совсем не обязательно должно идти какое-то определяемое им существительное, так что

в чистом виде исходные данные еще не свидетельствуют именно о языковой сочетаемости лексем. Все случаи вхождения не существительных на второй позиции в биграмме воспринимались нами как шум и по возможности исключались из исследовательской выборки. Особенно просто было осуществить отсечение нерелевантных частей речи для немецкого, в котором существительное графически маркируется заглавной буквой.

На следующем этапе для существительных, сочетающихся с прилагательным *густой*, были автоматически получены переводные эквиваленты в немецком и английском языках (предварительное обсуждение аналогичной методики см. в [16]). Процесс их поиска был предельно простым и включал в себя обращение к свободно распространяемым в машиночитаемом виде двуязычным словарям. В случае, если русское существительное из списка находилось в словаре, его переводной эквивалент (английский или немецкий) искался среди существительных, извлеченных из биграммных наборов соответствующего языка. Так, для русского слова *население* выдавался немецкий перевод *Bevölkerung*, который, в свою очередь, обнаруживался в немецком списке существительных, сочетающихся с одним из прилагательных изучаемого поля (а именно, с лексемой *dicht*).

Этот этап обработки материала дает наиболее очевидные результаты в тех случаях, когда для русского существительного в словаре, во-первых, находится ровно один переводной эквивалент, который, во-вторых, встречается в соответствующем иноязычном биграммном списке (именно такое соотношение представляет приведенный выше пример *население – Bevölkerung*).

Каждое из этих условий может нарушаться. Обсудим сначала первый круг проблем – «осложнения», возникающие на стадии словарной проверки, а именно: существительное может не обнаружиться в словаре (a) или же для него может встретиться более одного переводного эквивалента (b).

(a) Слова, которые не были найдены в словаре, программа собирает в отдельный файл (not_in_dic.csv): в результате в нем оказались преимущественно лексемы, записанные в старой орфографии (например, *покровъ*, *садъ* и под.)[1], т.е. переводные словари обнаружили довольно хорошее покрытие нужной для анализа лексики.

(b) Гораздо чаще для искомого существительного обнаруживается более одного переводного эквивалента. По нашим данным, эта ситуация может быть обусловлена одной из следующих причин:

1. В русском языке имеет место полисемия или омонимия, которая не характерна для языка перевода (ср. рус. *налет* и его немецкие аналоги *Überfall* ‘налет как нападение’ и *Belag* ‘налет как слой чего-л. на поверхности’). В подобных случаях наличие более одного перевода не создаст дополнительного «шума». Действительно, с прилагательными семантического поля ГУСТОЙ сочетается только слово *Belag*, а «лишняя» лексема *Überfall* будет отброшена просто в силу того, что она не встретится в биграммах, извлеченных нами для немецкого языка.

2. В русском языке имеется моносемичная лексема, но класс объектов, обозначаемых этим словом, в английском или немецком подразделяется на несколько типов, которые лексически разводятся. Пример такого рода – существительное *стена*, которое в немецком имеет два переводных аналога: *Wand* (в первую очередь, стена в здании или помещении) и *Mauer* (прежде всего, городская стена). В подобных случаях различные переводы представляют настолько близкие объекты, что они характеризуются одними и теми же свойствами и, соответственно, описываются одними и теми же прилагательными. Так, по данным биграмм, и *Wand*, и *Mauer* выступают в сочетании с двумя лексемами исследуемого поля – *dicht* и *dick*. Таким образом, и в этом случае наличие более одного переводного эквивалента не «мешает» автоматической процедуре соотнесения разноязычных биграмм.

Второй круг проблем связан со следующей стадией обработки данных – поиском найденных в словаре лексем в биграммных наборах английского и немецкого языков. Перевод, который был получен из словаря, может отсутствовать в этих списках – в таком случае он заносится в отдельный файл (noany.csv). В результате два файла – для английского и немецкого языков – оказались в некоторых отношениях сходными. Основным источником слов, не встретившихся в иноязычных биграммах, являются редкие или культурно-специфичные разновидности тех объектов, которые в принципе частотно выступают при прилагательных семантического поля ГУСТОЙ. Так, ни в английских, ни в немецких биграммах не встретилось сочетаний с переводными эквивалентами для слова *полынь*, хотя существительные более общей семантики со значением ‘трава’ вошли в биграммные наборы обоих языков. Аналогичный пример представляют лексемы *щи* и *борщ*: их переводы, в отличие от перевода для слова *суп*, не содержатся в иноязычных биграммах. Заметим, что примеры такого рода не свидетельствуют о различиях в устройстве семантического поля ГУСТОЙ в исследуемых языках, а только указывают на редкость употребления некоторых существительных в английском или немецком в сопоставлении с русским.

В некоторых случаях, однако, в списке не встретившихся в биграммах лексем оказываются не отдельные представители таксономической категории, а

[1] Заметим, что среди книг, оцифрованных в рамках проекта Google Books, имеются и издания XIX в.

целые классы слов. Так, поиск в немецких биграммах не обнаружил переводных аналогов для существительных *запах*, *аромат*, *вонь*. Тем самым, наш материал позволяет предположить, что метафора интенсивного запаха, характерная для русского прилагательного *густой* и обнаруженная в английском материале (ср. *thick smell* / *stink* / *odour*), не реализуется в немецких прилагательных описываемого поля.

Итак, расхождения в биграммных наборах разных языков могут выявлять важные различия в метафорических употреблениях исследуемых лексических единиц. Явным показателем таких различий является отсутствие в исходных данных определенного таксономического класса слов, тогда как отсутствие отдельных – периферийных – представителей того или иного класса отражает их экзотичность для культуры соответствующего языка.

Описанная до сих пор процедура обработки языкового материала позволяет очертить внешние границы поля и выявить несовпадения в его метафорической «надстройке». В следующем разделе мы обсудим, как предлагаемые нами автоматические методы способствуют анализу внутренней структуры поля.

4. Идентификация фреймов: результаты анализа

Переведенные при помощи словаря существительные, которые затем были найдены в сочетании с одним из интересующих нас прилагательных, заносятся в специальный файл. Каждый выходной файл соответствует одному из прилагательных анализируемого поля, т.е., например, для английского языка программа формирует отдельные файлы для лексем *thick*, *dense* и *bushy* – они выдаются по итогам работы программы наряду с файлами not_in_dic.csv (слова, не встретившиеся в словаре) и notany.csv (слова, не обнаруженные в биграммных наборах, подробнее см. выше).

Важно, что одно существительное может сочетаться более чем с одним прилагательным из интересующего нас списка. В этом случае оно не только заносится в каждый из релевантных файлов, но и в каждом файле при нем указываются все остальные прилагательные, с которыми оно встречается в биграммах. Фрагмент файла для *thick* приведен в табл. 1.

Таблица 1

Сочетаемость для прилагательного thick (фрагмент)

eyebrows	‘брови’	dense	bushy
branches	‘ветки’	dense	bushy
jungles	‘джунгли’	dense	bushy
soup	‘суп’	dense	
liquid	‘жидкость’	dense	
smog	‘смог’	dense	
porridge	‘каша’		
stink	‘вонь’		
laughter	‘смех’		

Все существительные в первом столбце выступают в качестве определяемого при *thick*, при этом слова *porridge*, *stink* и *laughter* не встречаются в биграммах с другими прилагательными, *soup*, *liquid* и *smog* сочетаются еще и с атрибутом *dense*, а *eyebrow*, *branches* и *jungles* – не только с *dense*, но и с *bushy*. Употребление с несколькими прилагательными возможно, во-первых, собственно за счет их синонимичности в данном контексте (хотя некоторые семантические различия в подобных случаях обычно все же обнаруживаются: так, в сочетании с существительным *soup* ‘суп’ лексема *thick* указывает на густую консистенцию супа-пюре, а *dense* описывает, скорее, неоднородный суп, в котором содержится много ингредиентов и мало жидкости). Во-вторых, множественная сочетаемость иногда обусловлена полисемичностью одного из прилагательных: в контексте того или иного существительного оно может выступать не в том значении, которое относится к интересующему нас полю. Это имеет место, например, в случае слова *thick*: наряду с семантикой густоты оно описывает еще и большой размер объекта в одном из его измерений (‘толстый’). В частности, при существительном *branches* ‘ветки’ атрибут *thick* как раз отсылает к их толщине, а идея густого переплетения, собственно соотносящаяся с нашим полем, выражается только прилагательным *dense*.

Вместе с тем, несмотря на то, что биграммы не всегда принадлежат семантической зоне ГУСТОЙ, сочетаемость с той или иной группой прилагательных задает значимые противопоставления внутри лексической системы. Действительно, по ряду сочетаемости существительные образуют довольно однородные классы, ср. фрагмент таблицы для немецкого прилагательного *buschig* (табл. 2).

Таблица 2

Сочетаемость для прилагательного buschig (фрагмент)

wuchs	‘рост/произрастание’	dicht	
pflanzen	‘растения’	dicht	
schwanz	‘хвост’	dick	
bäume	‘деревья’	dick	dicht
haar	‘волосы’	dick	dicht
augenbrauen	‘брови’	dick	dicht
bart	‘борода’	dick	dicht
schnurrbart	‘усы’	dick	dicht
ufer	‘берег’		
hügel	‘холм’		

Как видно из приведенных данных, слова с одинаковой сочетаемостью обнаруживают общую семантику. Так, в одной группе оказываются существительные, обозначающие покрытые растительностью поверхности (‘берега, холм’) – они выступают только с лексемой *buschig*. Самой широкой сочетаемостью отличаются имена, описывающие растительность, которая состоит

из множества отдельных элементов (‘волосы’, ‘брови’, ‘борода’, ‘усы’ ‘деревья’): каждый из этих элементов может быть охарактеризован по толщине (*dick*), их совокупность воспринимается как густая (*dicht*), при этом, как и другие растущие объекты, они способны присоединять к себе прилагательное *buschig*. Лексемы с собирательным или абстрактным значением ‘растения’ и ‘рост/произрастание’ в целом повторяют сочетаемость предыдущей группы существительных, но, в отличие от последних, не могут мыслиться как состоящие из отдельных элементов, поэтому не определяются атрибутом *dick*. Напротив, ‘хвост’ представляет собой единичный объект, поэтому соответствующее существительное, с одной стороны, может выступать при *dick*, но не сочетается с «множественным» прилагательным *dicht*: густота покрывающей его шерсти описывается лексемой *buschig*.

Разбиение на гомогенные классы в целом наблюдается и для других исследуемых прилагательных. Таким образом, предложенная методика, представляется, позволяет автоматически выделить те кластеры, которые релевантны для лексических противопоставлений в зоне ГУСТОГО. Иными словами, в результате моделируются семантические классы, которые могут стать прототипами фреймов и послужить основой для анкеты, направленной на анализ данных новых языков.

5. Ограничения методики и перспективы ее развития

Завершая представление нашей методики автоматической идентификации фреймов, необходимо отметить, что, с одной стороны, она действительно способна ускорить процедуру лексико-типологического анализа признакового поля, с другой – ее применение вскрывает некоторые ее ограничения.

Во-первых, формат биграммы не разворачивается в более широкий контекст. Между тем в сочетаниях с существительными высокого уровня абстрактности (ср. ‘форма’, ‘вид’, ‘множество’) специфику употребления прилагательного можно определить только по предложению в целом. Так, в обоих следующих контекстах фигурирует немецкая коллокация *dicht + Form*, при этом атрибут в ее составе имеет разную семантику: *dichte Form der Bebauung* букв. ‘густая форма застройки’ (*dicht* метонимически выражает физическое значение) vs. *Der Kurs vermittelt in dichter Form Grundlagen der Filmwissenschaft* ‘Курс в интенсивной <букв. «густой»> форме знакомит с основами киноведения’ (*dicht* употребляется в метафорическом значении). Тем самым на основании биграмм в некоторых случаях оказывается невозможно определить семантику атрибутивной лексемы. Это ограничение, кажется, нельзя обойти, если продолжать работать именно с предоставляемыми корпорацией Google биграммами.

Во-вторых, существительное, которое следует за прилагательным в биграмме, необязательно синтаксически с ним согласуется. Если эта связь отсутствует, то соответствующее сочетание не представляет собой атрибутивную группу и не является показательным для семантики прилагательного. Возникающий тем самым «шум» затрудняет получение качественных выводов. Отчасти преодолеть это ограничение можно было бы, проверяя все биграммы на наличие формального согласования между входящими в них элементами – и в этом нам видится одно из возможных направлений совершенствования предложенной методики.

В то же время наибольшее количество «шума» обнаружилось в английском материале, а его – ввиду отсутствия формальных показателей атрибутивной связи – нельзя подвергнуть дополнительной обработке. По-видимому, причиной повышенного «уровня шума» в этом случае является существенно больший, чем в случае других языков, объем исходных данных: количество англоязычных книг, оцифрованных в рамках проекта Google Books, значительно превышает число книг на других языках. С учетом этого обстоятельства, как кажется, простым способом, который позволил бы приспособить английские биграммы для наших задач, может стать увеличение порогового значения для частотности биграммы. Те двусловные последовательности, частота которых окажется ниже этого порога, следует отбрасывать и не учитывать в выборке. Тем самым за ее пределами останется целый ряд случайных сочетаний, не образующих атрибутивной группы.

Среди других перспектив совершенствования обсуждаемой методики нам видится автоматическое приписывание существительным семантических помет. Действительно, в ходе представленного здесь пилотного исследования группы с одинаковой сочетаемостью вручную распределялись по семантическим классам – при наличии помет эту процедуру можно будет до некоторой степени автоматизировать.

Кроме того, тестирование методики, описанное в этой статье, осуществлялось в одном направлении: от русских к иноязычным биграммам. Действуя таким образом, разумеется, невозможно выявить такие типы употреблений, которые характерны для прилагательных исследуемого поля в других языках и отсутствуют в русском. Поэтому естественным способом развития методологии должна стать проверка каждой из пар языков L_i и L_j в обоих направлениях: от L_i к L_j, L_j и L_i.

Наконец, типологическая картина оказывается тем полнее, чем больше языков было привлечено в ходе анализа. Как уже упоминалось, помимо рассмотренных здесь русского, английского и немецкого, биграммы Google доступны для 5 других языков, значит, нашим представлениям об устройстве поля еще предстоит обрасти новыми подробностями.

ЛИТЕРАТУРА

1. *Andersen E.* Lexical Universals of Body-Part Terminology / E. Andersen // J. Greenberg (Ed.) Universals of Human Language. – Stanford : Stanford University Press, 1978. – P. 335–368.

2. *Berlin B.* Basic Color Terms : Their Universality and Evolution / B. Berlin, P. Kay. – Berkeley : University of California Press, 1969.

3. *Brown C. H.* General Principles of Human Anatomical Partonomy and Speculations on the Growth of Partonomic Nomenclature / C. H. Brown // American Ethnologist. – 1976. – Vol. 3. – P. 400–424.

4. *Majid A., Bowerman M.* (Eds.) Cutting and Breaking Events : A Crosslinguistic Perspective [Special Issue]. Cognitive Linguistics, 18(2). 2007.

5. *Majid A., Levinson S. C.* (Eds.). The senses in language and culture [Special Issue]. The Senses & Society, 6 (1). 2011.

6. *Kopecka A.* Narasimhan B. (Eds.), Events of Putting and Taking : A Crosslinguistic Perspective (pp. 21–36). – Amsterdam : John Benjamins, 2012.

7. *Reznikova T.* Towards a typology of pain predicates / T. Reznikova, E. Rakhilina, A. Bonch-Osmolovskaya // Linguistics. – 2012. – V. 50. – № 3.

8. *Брицын В. М.* Концепт БОЛЬ в типологическом освещении / В. М. Брицын, Е. В. Рахилина, Т. И. Резникова ; ред. Г. М. Яворская. – Киев, 2009.

9. *Кюсева М. В.* Прилагательные *тяжёлый* и *лёгкий* в типологической перспективе / М. В. Кюсева, Д. А. Рыжова, Л. С. Холкина // Компьютерная лингвистика и интеллектуальные технологии : по материалам ежегодной Междунар. конф. «Диалог 2012» (Бекасово, 30 мая – 3 июня 2012 г.). – Вып. 11 (18). – М. : РГГУ, 2012.

10. *Павлова Е. К.* Лексико-типологическое описание признаков ‘мягкий’ – ‘твёрдый’ в языках разных семей. – М. : МГУ, 2014.

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

12. *Wälchli B., Cysouw M.* Lexical typology through similarity semantics : Toward a semantic map of motion verbs. Linguistics, 50(3), p. 671–710. 2012.

13. *Bhat D.N.S.* The adjectival category : criteria for differentiation and identification. Amsterdam : Benjamins, 1994.

14. *Bolinger D.* Adjectives in English : attribution and predication // Lingua 18, 1–34. 1967.

15. *Рахилина Е. В.* Когнитивный анализ предметных имен : семантика и сочетаемость / Е. В. Рахилина. – М. : Русские словари, 2000.

16. *Кюсева М. В.* Совершенствование одноязычных, двуязычных и мультиязычных словарей : автоматизация процесса сбора материала // М. В. Кюсева, Т. И. Резникова, Д. А. Рыжова. Доклады всероссийской научной конференции АИСТ’2013. – М. : Интуит, 2013.

Научно-исследовательский университет «Высшая школа экономики»

Орехов Б. В., кандидат филологических наук, доцент Школы лингвистики

E-mail: nevmenandr@gmail.com

Тел.: 8-916-526-09-51

Резникова Т. И., кандидат филологических наук, доцент Школы лингвистики

E-mail: tanja.reznikova@gmail.com

Тел.: 8-915-211-46-30

National Research University «Higher School of Economics»

Orekhov B. V., Candidate of Philology, Associate Professor of the Linguistics School

E-mail: nevmenandr@gmail.com

Tel.: 8-916-526-09-51

Reznikova T. I., Candidate of Philology, Associate Professor of the Linguistics School

E-mail: tanja.reznikova@gmail.com

Tel.: 8-915-211-46-30