КОРПУСНЫЙ ПАРАМЕТРИЧЕСКИЙ АНАЛИЗ ЛЕКСИКИ

А.А. Кретов, О.В. Донина, К.М. Шилихина

В статье рассматривается способ определения параметрических весов слова в лингвистическом корпусе, состоящем из N текстов. Деривационный параметр слова может быть оценен по количеству производных от него слов; функциональный параметр - по средней длине словоформ данного слова в корпусе (в буквах или звуках); синтагматический вес - как среднее количество разных лексем, приходящихся на одну синтагматическую связь данного слова. Парадигматический вес слова предлагается определять с помощью предобученных векторных семантических моделей типа RusVectōrēs, word2vec, fastText, ELMO, BERT.

КЛЮЧЕВЫЕ СЛОВА: корпусная лингвистика, компьютерная лингвистика, параметрический анализ лексики, лингвистический корпус

КРЕТОВ Алексей Александрович – доктор филологических наук, профессор, профессор кафедры теоретической и прикладной лингвистики Воронежского государственного университета. kretov@rgph.vsu.ru

ДОНИНА Ольга Валерьевна – кандидат филологических наук, доцент кафедры теоретической и прикладной лингвистики Воронежского государственного университета. olga-donina@mail.ru

ШИЛИХИНА Ксения Михайловна – доктор филологических наук, доцент, заведующая кафедрой теоретической и прикладной лингвистики Воронежского государственного университета. shilikhina@gmail.com

Цитирование: Кретов А.А., Донина О.В., Шилихина К.М. Корпусный параметрический анализ лексики [Электронный ресурс] // Мир лингвистики и коммуникации: электронный научный журнал. – 2022, № 2. – С. 1–24. Режим доступа: www.tverlingua.ru

CORPUS PARAMETRIC VOCABULARY ANALYSIS

Alexey A. Kretov, Olga V. Donina, Ksenia M. Shilikhina

The article discusses a method for determining the parametric weights of a word in a linguistic corpus consisting of N texts. The derivational parameter of a

word can be estimated by the number of words derived from it; the functional parameter - by the average length of the word forms of a given word in the corpus (in letters or sounds); syntagmatic weight - as the average number of different lexemes per syntagmatic connection of a given word. The paradigmatic weight of a word is proposed to be determined using pre-trained vector semantic models such as Rusvectōrēs, word2vec, FastText, ELMO, BERT.

KEY WORDS: corpus linguistics, computational linguistics, parametric analysis of vocabulary, linguistic corpus

KRETOV Alexey A. – DSc in Philology, Professor, Professor of the Department of Theoretical and Applied Linguistics of Voronezh State University. kretov@rgph.vsu.ru

DONINA Olga V. – PhD in Philology, Docent of the Department of Theoretical and Applied Linguistics of Voronezh State University. olga-donina@mail.ru

SHILIKHINA Ksenia M. – DSc in Philology, Docent, Head of the Department of Theoretical and Applied Linguistics of Voronezh State University. shilikhina@gmail.com

Citation: Kretov A.A., Donina O.V., Shilikhina K.M. Corpus parametric vocabulary analysis [Electronic resource] // World of linguistics and communication: electronic scientific journal. – 2022, № 2. – P. 1–24. Access mode: www.tverlingua.ru

Грант «Компьютерно-лингвистическая платформа нового поколения для цифровой документации русского языка: инфраструктура, ресурсы, научные исследования». Номер соглашения по гранту Министерства науки и высшего образования № 075-15-2020-793.

Параметрический анализ лексики (ПАЛ) активно разрабатывается Воронежской филологической школой. Так, за последние пять лет был проведен параметрический анализ славянских языков, карачаево-балкарского, кабардино-черкесского, абазинского, чеченского, киргизского,

турецкого, ненецкого, аварского, малагасийского, санскритского языков. В целом с момента появления данного подхода на сегодняшний день количество языков, обследованных с применением ПАЛа, перешагнуло рубеж в 60 языков и открыло седьмой десяток. Но все исследования проводились исключительно на базе словарей. Сегодня бурное развитие новых информационных технологий и корпусной лингвистики требует применения идеологии параметрического анализа лексики не только к словарям, но и к корпусам текстов.

В связи с этим попробуем наметить некоторые ориентиры на пути к достижению этой цели как новому этапу в развитии параметрического анализа лексики.

ПРОБЛЕМА КОРПУСНОГО ИССЛЕДОВАНИЯ СИСТЕМООБРАЗУЮЩИХ ПАРАМЕТРОВ ЛЕКСИКИ

Корпусная лингвистика, явившаяся результатом бурного развития информационных технологий на рубеже второго и третьего тысячелетий, с точки зрения истории существует еще ничтожно малый период времени. Тем не менее, возможности лингвистов, масштабы их деятельности, ее характер, а также существо вопросов, которые языковеды ставят перед собой, преобразовались настолько кардинально, что дают нам право уже сегодня, продолжая мысль В.А. Плунгяна, говорить о докорпусной и корпусной эпохах в лингвистике (Плунгян, 2010). Из вышесказанного вытекает закономерный вопрос: какие направления и методы, существующие в лингвистике сегодня, окажутся самыми продуктивными, а какие будут «подвинуты» и должны будут пройти серьезную адаптацию в новых условиях?

Параметрический анализ лексики зародился и развивается в русле системной лингвистики как комплекс приемов системного описания лексики языка. По оценке Г.П. Мельникова, системная лингвистика воплощает греческий стиль научного мышления (всесторонний, комплексный, синтезирующий), наиболее ярко выразившийся в трудах отечественных

ученых-космистов (Мельников, 2003). Сам Г.П. Мельников отмечал, что подход к языку с точки зрения системологии представляет собой преодоление сугубо формально-математического моделирования за счет ориентации на субстрат системы и шире – на ее содержательные аспекты (Мельников, 1978). Внимание непосредственно к элементам системы (при этом не перерастающее в атомистический взгляд), к их содержательной нагруженности помогает осмыслить «вес» того или иного элемента в этой системе. Вместо анализа «чистых отношений» между элементами – парадигматических, синтагматических, эпидигматических – исследователь с опорой на эти отношения обращается к природе и смысловой нагруженности самого элемента. Подобный подход (как, впрочем, и некоторые элементы космизма) гармонично вписывается в новейшие естественнонаучные модели и соответствующую им философию – ср. идеи К.Э. Циолковского о «мыслящем атоме» и некое подобие психичности фундаментальных микрочастиц, проявляющееся в нарушении неравенств Белла (см. о последнем: (Копейкин, 2016)); идею обращения к содержательной стороне систем в системологии и проблему интерпретации квантовой механики, см.: (Русский космизм, 1993; Гинзбург, 2004). Основа вырисовывающейся при этом научной парадигмы видится К.В. Копейкиным в общей задаче современной науки наполнить смыслом формальные математические структуры (Копейкин, 2016).

Идея системного лингвистического описания восходит к В. фон Гумбольдту, развивавшему точку зрения на язык сквозь призму антиномий (ср., например, антиномный подход В. фон Гумбольдта и основные постулаты лингвистики, сформулированные уже во второй половине двадцатого века Ю.С. Степановым (Гумбольдт, 1984; Степанов, 1975)) и при этом развивает диалектическую точку зрения на язык сквозь призму триад (ср.: (Гегель, 1970) и (Мельников, 2003). Поскольку теоретические воззрения В. фон Гумбольдта стали отправной точкой развития общего языкознания в целом, то можно сказать, что системная лингвистика опирается на базовые

принципы лингвистического теоретизирования. Так, рассматриваемые системной лингвистикой системообразующие параметры лексики являются одновременно и основными лингвистическими понятиями – речь идет о синтагматике, парадигматике и эпидигматике. В системной лингвистике стоящие за этими понятиями языковые явления изучаются в единстве и взаимосвязи. Таким образом, системная лингвистика представляет собой Лингвистику во всём богатстве и единстве накопленных ею к настоящему времени методов. Именно так и определял центральное направление отечественной лингвистики Н.Д. Андреев (Андреев, 1965).

Сам по себе ПАЛ является одним из таких обогащающих методов, поскольку в нем три вышеназванных понятия традиционной лингвистики получили строгую системную интерпретацию. Однако до сих пор данный метод был ориентирован на лексикографические источники. Для того, чтобы понять, как этот метод будет работать на корпусном материале, необходимо определить отличия этого материала от материала традиционного.

Для начала определимся с самой природой языкового материала. По Л. В. Щербе, язык-материал – это один из аспектов, в которых существуют языковые явления, наряду с языком-системой и речевой деятельностью (Щерба, 1974). На основе материала – текстов – исследователь путем наблюдений и логических умозаключений восстанавливает принципы устройства и закономерности функционирования языковой системы, фиксируемые в словарях и грамматиках. Таким образом, словарь – это в строгом смысле уже вторичный источник, созданный на базе некоторого корпуса – первичного источника. Поэтому вопрос о степени и характере применения ПАЛа в анализе корпусов является одновременно и вопросом его применения к первичному речевому материалу. При этом словарь благодаря работе лексикографов является источником, в неизмеримо большей степени подготовленным для извлечения из него ядра лексики, в то время как богатство корпуса – это прежде всего – богатство данных о периферии (Кретов 2021).

Если далее сравнивать словарь и корпус, то оба этих источника характеризуются конечным объемом. Каких бы циклопических размеров ни достигал электронный словарь или корпус, это всегда будет какое-либо натуральное число, выражающее количество словарных статей или корпусных вхождений. Другое дело, что, по сравнению со стабильно существующим словарем, корпус – это материал перманентно пополняемый (по крайней мере, постоянный рост объема корпуса воспринимается как норма в современной практике). Еще одно фундаментально различие – корпус, как и всякий традиционный текст, не дает готовых сведений о синонимии, сочетаемости, количестве значений того или иного слова (что даёт словарь). Более того, сама идентификация того или иного ЛСВ слова остается на откуп исследователю корпусных данных. Поэтому корпусное исследование лексики по природе своей будет *синтаксичным* – ведь вся лексика будет представлена только в виде употреблений словоформ в составе синтаксического ряда (см.: Ломов, 1977).

При этом важнейшим в контексте лексико-семантической макротипологии будет еще одна черта корпусного представления данных: по запросу исследователь будет получать «много (по меньшей мере, в относительном измерении) и сразу» (Sinclair 1991), что адекватно задачам *макро*типологии, имеющей место с большими объемами данных о языковой системе в целом.

Таким образом, корпус исловарь как источники материала для исследователя характеризуются отношением «первичное» – «вторичное», работа с каждым из них будет отмечена своими особенностями, но в целом она будет носить взаимодополнительный характер, причем словарь даже в корпусную эпоху остается единственным в своем роде надежным источником лингвистических данных (особенно о ядерных явлениях языка, см.: (Кретов, 2021)).

Теперь обратимся к особенностям корпусного варианта ПАЛа. Содержательной интерпретации системных параметров в ходе исследования

предшествует формализация данных о языковой системе, служащая посредником между метаязыком корпуса – в конечном счете, машинным кодом – и живым мыслящим субъектом – исследователем. Рассмотрим теперь основные линии, по которым может быть проведена формализация каждого системного параметра.

Синтагматика слова наглядно формализуется при древесном представлении синтаксических связей в предложении. Здесь исследователю могут оказаться полезными идеи грамматики зависимостей (см.: Теньер, 1988; Тестелец, 2001). Действительно, слово в составе высказывания выделяется путем абстрагирования от конкретной словоформы в составе замкнутого синтаксического ряда, которая, в свою очередь, обладает свойствами синтаксической ориентации и/или синтаксической проекции (см. подробнее в (Распопов, 1973)). Синтаксические связи между словоформами и складываются в позиционную схему высказывания, которую, кроме прочего, можно формально представить в виде дерева зависимостей (см.: (Попова, 2004)). Еще одним вариантом формализации синтаксических связей может служить модель управления, отображающая информацию не только о синтаксической проекции слова, но и о семантических актантах, выражаемых синтаксическими позициями (Апресян, 2010).

Ключ к изучению эпидигматики слова в корпусе дает антиномия внутреннего и внешнего, их взаимосвязь, раскрываемая через соотношение количества значений слова и его дериватов. Системное рассмотрение оппозиции предполагает обнаружение ее нейтрализации (Руделев, 1980). На наш взгляд, оппозиция внешнего и внутреннего в плане эпидигматики также носит системный характер и находит свою нейтрализацию в таком свойстве разных ЛСВ одной и той же лексемы, как различная их дистрибуция. Действительно, разные ЛСВ могут обладать различными проективными свойствами. Приведем для наглядности примеры Ю. Д. Апресяна: *дрожать над каждой копейкой* (беречь, жалеть) и *дрожать перед самодуром*

(бояться) (Апресян, 1995, т. II: 538). Здесь внутреннему различию (семантика) соответствует внешнее различие (управление).

Наиболее сложно формализуемым параметром является **парадигматика**. Одна из причин этой сложности заключается в том, что непосредственно из текста путем абстрагирующей деятельности мы извлекаем именно синтагмы, а парадигмы являются результатом абстракции уже следующего уровня, когда полученные синтагмы сопоставляются. Поэтому принцип изучения парадигматики через синтагматику (Мельников, 2003; Максименко, 2019); см. также в этой связи (Шайкевич, 2016)) оказывается в контексте корпусного исследования крайне актуальным.

В условиях большого объема анализируемых данных и слабой формализуемости изучаемого явления нам видится логичным выводить парадигмы на основе однотипной дистрибуции слов. В 1960-х гг. подобную методику применил Ю.Д. Апресян для анализа семантики русского глагола, см.: (Апресян, 1967; 1995). Исследователь убедительно показал существование зависимости глагольной семантики от дистрибутивно-трансформационных свойств слова на основе синтаксической обусловленности лексических значений. В качестве базовых критериев идентичности значения слова Ю.Д. Апресян предлагает синтаксические отношения совместимости и трансформируемости, отмечая при этом, что синтаксическая обусловленность распространяется и на семантические классы слов (Апресян, 1995). При всей (признаваемой самим автором) ограниченности формально-синтаксического подхода к описанию лексической семантики, последний остается единственным способом получить знания о парадигматике слов в таком большом объеме данных, как корпус. При этом, естественно, мы отдаем себе отчет, что высокой степенью семантической близости придется пожертвовать: в одной и той же позиции могут оказаться, например, эквонимы (см.: (Никитин 1996)) и т.д.

Нетривиальный, но идейно перекликающийся с приведенным выше подход к определению семантики слов в церковнославянском языке был

предложен Е.М. Верещагиным в работе (Верещагин, 1997). Данный подход также базируется на анализе дистрибутивных свойств единиц, в частности, явлении параллелизма, обнаруженного автором в тексте Псалтири. Анализ семантики лексем, оказывавшихся в аналогичных позициях параллельной синтаксической структуры, давал исследователю информацию о семантических характеристиках церковнославянских слов. Таким образом, формально-синтаксический анализ лексических значений применялся в истории лингвистики к разному материалу и доказал свою эффективность (при учете ограничивающих факторов).

В целом мы можем сделать промежуточный вывод, что системный анализ лексики принципиально возможен в лингвистическом корпусе. В частности, существуют подходы, в рамках которых основные положения ПАЛа оказываются приложимыми к корпусному анализу. Последнее позволяет предположить, что ПАЛ может занять достойное место в будущем корпусной лингвистики. Ниже мы рассмотрим конкретные пути изучения параметров лексики в Национальном корпусе русского языка.

Напомним, что ПАЛ охватывает три системообразующих параметра лексики: синтагматический, парадигматический и эпидигматический и один функциональный.

Начнем обсуждение проблем в порядке возрастания их сложности.

ФУНКЦИОНАЛЬНЫЙ ПАРАМЕТР В КОРПУСЕ

В словарях он косвенно оценивался по длине леммы (словарной формы, представляющей слово в двуязычном иноязычно-русском словаре), измеренной в звуках. В корпусе же представлены словоформы, каждую из которых следует лемматизировать – т.е. поставить в соответствие лемме.

На сегодняшний день для русского языка существует немало лемматизаторов. Один из них применяется в НКРЯ. Ни один из автоматических лемматизаторов не совершенен, но погрешность лемматизации при обработке больших корпусов уже может быть признана несущественной.

Если мы можем лемматизировать слово, значит, можем получить все его словоформы в данном тексте с их частотой. А если так, то мы получаем возможность заменить Ф-вес с дискретного (в целых) на недискретный, выбрав в качестве Ф-вес среднюю длину словоформы данного слова в данном тексте.

(1) Например, словоформы глагола *дать* (лемма имеет длину 3 звука) в 6-томном Собрании сочинений А.Т. Твардовского представлены со следующей частотой (см. Табл. 1). (Ср. длина СлФормы = Σч*д :Σч, где «ч» - частота словоформы, а «д» - длина словоформы в корпусе).

Таблица 1. Вычисление средней длины словоформы глагола *дать* в текстах А.Т. Твардовского

СлФорма	**Частота**	**Длина**	**Ч*Д**
ДАВ	2	3	6
ДАДИМ	1	5	5
ДАДУТ	4	5	20
ДАЙ	38	3	114
ДАЙКА	6	5	30
ДАЙТЕ	14	5	70
ДАЛ	11	3	33
ДАЛА	9	4	36
ДАЛИ	9	4	36
ДАЛО	1	4	4
ДАМ	14	3	42
ДАН	3	3	9
ДАНА	6	4	24
ДАНО	10	4	40
ДАНЫ	1	4	4
ДАСТ	8	4	32
ДАТь	6	**3**	18
ДАШь	10	3	30
ВСЕГО:	**153**		**553**
Ср.ДлинаСлФ			**3,61437908496732**

При этом мы можем округлять среднюю величину до десятых, сотых, тысячных и т.д. Но даже если мы округлим её до целых, то получим не 3 звука, как в лемме, а 4, как в обследованных текстах.

Иными словами, мы получаем возможность регулировать масштаб описания объекта, выбирая наиболее целесообразный, и получая более достоверные данные о длине слова в виде средней длины словоформ, представляющих это слово в тексте.

СИНТАГМАТИЧЕСКИЙ ПАРАМЕТР В КОРПУСЕ

В словаре синтагматически вес слова оценивается по количеству словосочетаний (фразеологизмов, речений), представленных в словарной статье. Как правило, словари (кроме словарей сочетаемости) скупы на информацию о синтагматике слова.

В корпусе ситуация потенциально значительно более благоприятная. Синтагматическими связями слов являются синтаксические связи. В настоящее время уже существует немало парсеров, автоматически устанавливающих синтаксические связи между словами. При этом если лемматизатор ориентируется преимущественно на морфологию в варьирование флексий, то парсер предполагает предварительное выполнение морфологического анализа словоформ.

Синтагматический вес слова предлагается измерять средним количеством разных слов, приходящихся на одну синтаксическую связь слова.

Зачем это нужно?

Возьмём фразу ***Белая*** *берёза под моим окном* ***принакрылась*** *снегом, точно серебром* (С.Есенин).

У прилагательного *белая* всего одна синтаксическая связь – со словом *берёза*.

У существительного *берёза* уже 2 синтаксических связи: с прилагательным *белая* и глаголом *принакрылась*.

У глагола *принакрылась* 3 синтаксических связи: с существительным *берёза* и существительным *снегом,* а также с предложно-падежной конструкцией *под окном (моим)*, содержащей форму существительного *окно*.

Если просто считать слова, имеющие синтагматические связи с данным, слова разных частей речи окажутся в неравных условиях: у прилагательного 1 слово (*берёза*), у существительного – 2 (*белая, принакрылась*) а у глагола – 3 (*берёза, снегом, под окном*). Но если мы посчитаем среднее количество слов, приходящихся на одну синтаксическую связь, то получится, что прилагательное, существительное и глагол – в равном положении: у каждого по одному слову на синтаксическую связь.

Именно средним количеством разных слов (не словоформ!) и предлагается измерять синтагматический вес слова в корпусе ($ВЕС_{синтагм.}$ = Σлемм :Σсинт.связ.).

ДЕРИВАЦИОННЫЙ ПАРАМЕТР В КОРПУСЕ

№	Лемма	Этимологическая транскрипция и морфемное членение	Правило 1	П2	П3	П4	П5
51586	КОРКОВИДНЫЙ	кор=ък_овИд=ьн_ый	68	702	53	55	52
51587	КОРКОВЫЙ	кОр=ък=ов_ый	269	53	52		
51588	КОРКОРЕЗКА	кор=ък_орЁз=ък_а	702	70	132	53	55
51589	КОРКОРЕЗНЫЙ	кор=ък_орЁз=ьн_ый	68	702	132	53	55
51590	КОРКУ	кО"рку _#	55	48	52	366	
51591	КОРМ	кЪр=м_ъ	118	50	53	52	
51592	КОРМА	кър=м_А	175	53	52	2723	2775
51593	КОРМАНЬСКИЙ	кО"рман'=ьск_ий	2630	4339	4363	52	149
51594	КОРМАЧ	кър=м=Акй_ъ	50	175	53	52	2723
51595	КОРМАЧКА	кър=м=Акй=ък_а	2157	2697	2839	77	447
51596	КОРМЁЖКА	кър=м=Ег=йък_а	2906	4226	141	2747	175

51597	КОРМЕЖНЫЙ	кър=м=Ег=ьн_ый	2906	141	2747	68	175

В словаре деривационный (он же эпидигматический) параметр оценивается по количеству значений или по количеству дефиниций, содержащихся в словарной статье данного слова.

Ещё в работе (Кретов, 1987) было установлено, что внутренняя деривационная активность слов (многозначность) коррелирует с их внешней деривационной (т.е. словообразовательной активностью) или продуктивностью. **Деривационную продуктивность каждого слова можно измерять количеством слов, производных от него.**

Можно ли формализовать нахождение этой величины в корпусе? Да, можно. Для русского языка тут поможет БРУМС – Большой Русский Морфемный Словарь (Кретов, 2003; Кретов, 2018), в котором представлено морфемное членение более чем 165.000 слов в виде базы данных в форматах Эксель, Аксесс и др. и отражает практически всё лексическое богатство русского литературного языка по состоянию на 1991-ый год.

Таблица 2. Фрагмент БРУМСа

Правда, морфемному членению подвергнуты леммы (т.е. словарные формы), а в корпусе встречаются не только они. Но поскольку изначально анализ корпусов немыслим без лемматизации и лемматизатора, эта проблема снимается: ведь каждой словоформе лемматизатор ставит в соответствие его лемму. Облегчает задачу и то, что морфемной членение в БРУМСЕ дано не в орфографической записи (что для русского языка как фузионного просто исключено), а в этимологической (=фонематической) транскрипции.

Понятие производности также формализуемо. **Производным является слово, отличающееся от производящего не менее чем на одну терминальную морфему**. Терминальной морфемой может быть как приставка: *писать*>***на-****писать*, так и суффикс: *писать* >*писа*=***тель.***

Продуктивность слова *вид* будет больше, чем продуктивность образованного от него глагола *видеть*, продуктивность глагола *видеть* будет

больше, чем продуктивность производного от него глагола *свидеться,* продуктивность слова *свидеть(ся)* будет больше, чем продуктивность его производного – существительного *свидетель* и т.д.

БРУМС существует в виде базы данных, что позволяет автоматически осуществлять морфемное членение словоформ через их соотнесение с леммой и собирать все однокоренные слова. Нулевой деривационный вес будет у слов, не имеющих производных.

ПАРАДИГМАТИЧЕСКИЙ ВЕС В КОРПУСЕ

Даже в словаре определение парадигматического веса является самой сложной, трудоёмкой и затратной процедурой: ведь парадигматический вес измеряется количеством синонимов у данного слова.

Поскольку многозначные слова составляют около половины словаря, парадигматический вес слова определяется по тому значению, которое объединяет максимальное число синонимов. Часто (хотя и не всегда) это бывает первое значение слова. В параметрическом анализе синонимами в узком смысле принято считать слова с тождественными дефинициями, эквивалентами или же слова различающиеся лишь порядком следования метаслов. В широком смысле синонимами признаются слова, в толкованиях которых совпадает не менее 50% метаслов.

Если в словарях представлено деление слов на значение и толкования этих значений с помощью метаслов, то в корпусах такая информация отсутствует, а не имея значений и дефиниций мы лишаемся возможности выявлять синонимические ряды и получать дефиниции слов – пусть даже приблизительные, т.е. с минимумом дифференциальных сем.

Для параметрического анализа и вычисления парадигматического веса слова в корпусе можно применить дистрибутивно-статистический анализ А.Я. Шайкевича (Шайкевич, 2016) или вычисление «индекса значимости» для идиоглосс по методу Е.Л. Гинзбурга (Гинзбург, 1998).

Правда, первый метод громоздок, и эффективность его не очевидна, а второй предполагает первоначальное выделение множества слов, для

которых будет определяться парадигматический вес. Задачу выделения слов –кандидатов в идиоглоссыможно решить, выделив слова, вошедшие в 3 или 2 другие частно-параметрических ядра. Однако точность и надёжность получаемых этим методов результатов не очевидна.

Решение видится в двух направлениях: 1) формальном, восходящем к идее А.А. Холодовича и раннего Ю.Д. Апресяна о выделении семантических классов по сходству дистрибуции и 2) содержательном (метаязыковом) – в том направлении, в каком разрабатывает метаязык для семантической классификации полнозначных слов Е.Н. Подтележникова (Подтележникова, 2015).

При формальном подходе семантически близкими словами могут считаться слова со сходной дистрибуцией. При этом чем больше сходство дистрибуции двух слов, тем ближе, по идее, они должны быть семантически.

Проблема состоит в том, что сходной дистрибуцией могут обладать и различные по семантике слова. Можно, конечно, пробовать в качестве подстраховки использовать словари синонимов с готовыми синонимическими рядами, но эти ряды могут быть не полны.

Думается, результатом парадигматического анализа текста может быть и объединение более крупное, чем синонимический ряд: например, эквонимы (согипонимы), имеющие общий гипероним. Как показывает опыт исследования прилагательных, сочетаемость слов *белый, черный* и *серый* во многом совпадает (Кретов, 2016). Но общее у них – только визуальная архисема ‘цвет’.

Ещё одно направление также не сулит быстрых результатов: это обучение нейронных сетей по размеченным или неразмеченным текстам, содержащим синонимы.

На данный момент наиболее перспективным методом для вычисления парадигматического веса в корпусе кажется возможность использования векторных семантических моделей. Векторное представление токенов, основанное на дистрибутивной гипотезе, подразумевает, что «геометрические

отношения между векторами слов должны отражать семантические связи между соответствующими им словами. Как предполагается, векторные представления слов должны отображать человеческий язык в геометрическом пространстве» (Шолле, 2018: 217).

Выделяют два способа получения векторного представления.

Первый – конструирование векторных представлений слов с помощью слоя Embedding. В этом случае изначально пространство векторов не имеет структуры; в рамках решения основной задачи создают случайные векторы, которые, подобно весам нейронных сетей, постепенно обучаются. При этом рекомендуется под каждую конкретную задачу проводить новое обучение векторного пространства. Для решения этой задачи может использоваться прием обратного распространения ошибки (при применении которого в ходе обучения происходит постепенная корректировка векторов и образование структурированного пространства); в данном случае это может быть обучение весов слоя Embedding: «слой Embedding лучше всего воспринимать как словарь, отображающий целочисленные индексы (обозначающие конкретные слова) в плотные векторы. Он принимает целые числа на входе, отыскивает их во внутреннем словаре и возвращает соответствующие векторы» (Шолле, 2018: 218).

Второй способ получения векторного представления – использовать предварительно обученные векторные представления слов, рекомендуемые к применению в случае небольшого объема обучающих данных. Несмотря на то, что первые работы, в которых описывались плотные малоразмерные пространства векторных представлений слов, обучаемые без учителя, начали появляться с начала 21 века (Bengio, 2006), широкое распространение этот подход получил после 2013 г., когда Томас Миколов разработал алгоритм word2vec (https://code.google.com/archive/p/word2vec/ ((Mikolov, 2013a, 2013b). Еще одно готовое к использованию, предобученное векторное представление слов, появившееся в 2014 г. – Global Vectors for Word Representation (GloVe) (https://nlp.stanford.edu/projects/glove/).

Для вычисления парадигматического веса можно использовать предварительные обученные векторные представления слов. Рассмотрим, как это может работать для русского языка при использовании предобученных векторных семантических моделей RusVectōrēs (https://rusvectores.org/ru/models/). Например, для глагола *бить* были получены следующие семантические ассоциаты, которые могут быть использованы для расчета парадигматического веса (рис.1). Несмотря на довольно неплохие результаты, возникает две основные сложности: 1) необходимо ограничить степень близости, чтобы при расчетах учитывались только наиболее близкие искомому слову ассоциаты, 2) в качестве синонимов мы рассматриваем только слова с разными корнями, т.е. необходимо приведение однокоренных лексико-семантических вариантов к одному лексико-семантическому инварианту. Что касается первого вопроса, необходимо провести исследование для определения порога близости ассоциатов, который мы сможем учитывать при расчете парадигматического веса (по предварительным данным, этот порог не может быть ниже 0,6). Для решения второй сложности мы можем пропускать набор лемм из RusVectōrēs через БРУМС, чтобы отсекать однокоренные слова. Таким образом, парадигматический вес будет рассчитываться как количество семантических ассоциатов с разными корнями.

Рис. 1. Семантические ассоциаты для глагола «бить» в RusVectōrēs

При масштабировании данного подхода можно будет либо получать доступ к данным сайта RusVectōrēs через их API, либо дообучить дистрибутивно-семантические модели, предоставляемые RusVectōrēs, и использовать их в собственных разработках.

В дальнейшем по мере обучения векторных семантических моделей на материале различных языков можно будет сравнить различные модели векторного представления слов (word2vec, fastText, ELMO, BERT и др.) для выбора наиболее подходящего подхода для вычисления парадигматического веса.

Путь к решению задачи определения парадигматического веса слова по корпусу непрост. Но, как говорят на Востоке, дорогу осилит идущий.

Ссылки - References in Russian

Андреев, 1965 – *Андреев Н.Д.* Статистико-комбинаторное моделирование языков. – М.: Наука, 1965. – 502 с.

Апресян, 1967 – *Апресян Ю.Д.* Экспериментальное исследование семантики русского глагола. – М.: Наука, 1967. – 251 с.

Апресян, 1995 – *Апресян Ю.Д.* Избранные труды. –М.: Языки русской культуры, 1995. – Т. I-II.

Апресян, 2010 – *Апресян Ю.Д., Богуславский И.М., Иомдин Л.Л., Санников В.З.* Теоретические проблемы русского синтаксиса: взаимодействие грамматики и словаря. – М.: Языки славянских культур, 2010. – 408 с.

Верещагин, 1997 – *Верещагин Е.М.* История возникновения древнего общеславянского литературного языка: Переводческая деятельность Кирилла и Мефодия и их учеников. – М.: Мартис, 1997. – 314 с.

Гегель, 1970 – *Гегель Г.В.Ф.* Наука логики. – М.: Мысль, 1970. – Т. I. – 501 с.

Гинзбург, 2004 – *Гинзбург В.Л.* О сверхпроводимости и сверхтекучести (что мне удалось, а что не удалось), а также о «физическом минимуме» на начало XXI века // УФН. 2004. – Т. 174. – №11. – С. 1240–1255.

Гинзбург, 1998 – *Гинзбург Е.Л.* Идиоглоссы: проблемы выявления и изучения контекста // Семантика языковых единиц: Доклады VI Международной конференции. – Т.I. – М., 1998. – С. 26–28.

Гумбольдт, 1984 – *Гумбольдт В. фон.* Избранные труды по языкознанию. – М.: Прогресс, 1984. – 400 с.

Копейкин, 2016 – *Копейкин К. В.* Интервью // Портал polit.ru. 28.05.2016. Режим доступа: https://polit.ru/article/2016/05/28/kopeykin_interview/

Кретов, 1987 – *Кретов А.А.* Принципы выделения ядра лексико-семантической системы // Семантика слова и синтаксической конструкции. Межвуз. сб-к научн. Трудов. – Воронеж, 1987. – С. 84–93.

Кретов, 2003 – *Кретов А.А.* "Большой русский морфемный словарь" в обучении иностранцев // Русское слово в мировой культуре. Материалы X Конгресса Международной ассоциации преподавателей русского языка и литературы. Санкт-Петербург, 30 июня – 5 июля 2003 г. – СПб: Политехника, 2003. – С. 273–283.

Кретов, 2016 - *Кретов А.А.* «*Белый дом*» и «*Чёрная дыра*»: алгоритм определения семантической близости слов по их сочетаемости // Вестник ВГУ, Серия: Системный анализ и информационные технологии, 2016, № 4. – С. 174–177.

Кретов, 2018 - *Кретов А.А.* БРУМС как электронный словарь и база данных // Тезисы всероссийской конференции «От языковых машинных фондов к лингвистическим корпусам: памяти В.М. Андрющенко». Институт русского языка имени В.В. Виноградова. – Москва, 28-29 сентября 2018 г. http://lcl.srcc.msu.ru/library/abstracts.pdf. – С. 38–42.

Кретов, 2021 – *Кретов А.А., Гасунс М.Ю., Леонченко В.В.* Параметрический анализ санскритско-русского словаря» В. А. Кочергиной //

Проблемы общей и востоковедной лингвистики. Сочетаемость языковых единиц и языковые модели. – М.: ИВ РАН, 2021. – С. 287–302.

Ломов, 1977 - *Ломов А.М.* Очерки по русской аспектологии. – Воронеж: Издательство Воронежского университета, 1977. – 140 с.

Максименко, 2019 - *Максименко О. И.* Автоматизированный дистрибутивно-статистический анализ как системная обработка текста // Вестник РУДН. Серия: Теория языка. Семиотика. Семантика. – 2019. – Т.10. – №1. – С. 92–100.

Мельников, 1978 – *Мельников Г.П.* Системология и языковые аспекты кибернетики. – М.: Сов. радио, 1978. – 368 с.

Мельников, 2003 - *Мельников Г.П.* Системная типология языков: принципы, методы, модели. – М.: Наука, 2003. – 395 с.

Никитин, 1996 - *Никитин М.В.* Курс лингвистической семантики: Учебное пособие. – СПб: Научный центр проблем диалога, 1996. – 760 с.

Плунгян, 2010 – *Плунгян В.А.* Интервью. – Портал bogoslov.ru, 1.02.2010. Режим доступа: https://bogoslov.ru/article/573240

Подтележникова, 2015 – *Подтележникова Е.Н.* Проект метаязыка для компьютерной классификации лексических значений // Вестн. Воронеж. гос. ун-та. Сер. Лингвистика и межкультурная коммуникация. – Воронеж, 2015, № 3. – С. 34–37.

Попова, 2004 – *Попова З.Д., Стернин И.А.* Общее языкознание : учеб. пособие. - Воронеж: Центрально-Черноземное книжное издательство, 2004. – 208 с.

Распопов, 1973 – *Распопов И.П.* Очерки по теории синтаксиса. - Воронеж: Издательство Воронежского университета, 1973. – 220 с.

Руделев, 1980 – *Руделев В.Г.* Теория нейтрализации. Некоторые результаты. Перспективы развития // Теория нейтрализации: сб. науч. статей. Тамбов, 1980. – С. 3–10.

Русский космизм, 1993 - Русский космизм: Антология философской мысли / Сост. С.Г. Семенова, А.Г. Ганева. – М.: Педагогика-Пресс, 1993. – 368 с.

Степанов, 1975 – *Степанов Ю.С.* Основы общего языкознания. Учебное пособие. – М.: Просвещение, 1975. - 271 с.

Теньер, 1988 - *Теньер Л.* Основы структурного синтаксиса. – М.: Прогресс, 1988. – 656 с.

Тестелец, 2001 – *Тестелец Я.Г.* Введение в общий синтаксис. – М.: РГГУ, 2001. – 800 с.

Шайкевич, 2016 – *Шайкевич А. Я., Андрющенко В.В., Ребецкая Н.А.* Дистрибутивно-статистический анализ языка русской прозы 1850–1870-х гг. – М.: Языки славянской культуры, 2016. – Т. I-II. – 848 с.

Шолле, 2018 – *Шолле Ф.* Глубокое обучение на Python. – Спб.: Питер, 2018. – 400 с.

Щерба, 1974 – *Щерба Л.В.* Языковая система и речевая деятельность. –- Л.: Наука, 1974. – 427 с.

References

Andreev, N.D. (1965) *Statistiko-kombinatornoe modelirovanie yazykov*, M., Nauka, 502 p. (In Russian)

Apresyan, YU. D. (1967) *Ehksperimental'noe issledovanie semantiki russkogo glagola*, M., Nauka, 251 p. (In Russian)

Apresyan, YU. D. (1995). *Izbrannye trudy*, M., Yazyki russkoi kultury, T. I-II. (In Russian)

Apresyan, YU. D., Boguslavskii, I.M., Iomdin L.L., Sannikov V.Z. (2010) *Teoreticheskie problemy russkogo sintaksisa: vzaimodeistvie grammatiki i slovarya*, M., Yazyki slavyanskikh kultur, 408 p. (In Russian)

Bengio, Y., Schwenk, H., Senécal, J.-S., Morin, F., Gauvain, J.-L. Neural. (2006) Probabilistic Language Models. *Innovations in Machine Learning*, Springer, pp. 137–186.

Gegel, G.V.F. (1970) *The Science of Logic*, M., Mysl, T. I, 501 p.

Ginzburg, V.L. (2004) O sverkhprovodimosti I sverkhtekuchesti (chtomneudalos', a chto ne udalos'), a takzhe o «fizicheskom minimumE» nanachalo XXI veka // *UFN*. T. 174, №11, pp.1240–1255. (In Russian)

Ginzburg, E.L. (1998) Idioglossy: problemy vyyavleniya i izucheniya konteksta // *Semantikayazykovykhedinits: DokladyVIMezhdunarodnoikonferentsii.*T.I, pp. 26–28. (In Russian)

Gumboldt, V. fon. (1984) *Izbrannyetrudy po yazykoznaniyu*, M., Progress, 400 p. (In Russian)

Kopeikin, K.V. (2016) Intervyu (Interview) // *Portal polit.ru*. Retrieved from https://polit.ru/article/2016/05/28/kopeykin_interview/

Kretov, A.A. (1987) Printsipy vydeleniya yadra leksiko-semanticheskoi sistemy // *Semantika slovaisintaksicheskoi konstruktsii Mezhvuz.* sb-k nauchn. trudov, Voronezh, pp.84–93. (In Russian)

Kretov, A.A. (2003) "Bol'shoi russkii morfemnyi slovar'" v obuchenii inostrantsev // *Russkoe slovo v mirovoi kul'ture.* Materialy KH Kongressa Mezhdunarodnoi assotsiatsii prepodavatelei russkogo yazyka I literatury. Sankt-Peterburg, 30 iyunya – 5 iyulya 2003 g. SPb: Politekhnika, pp. 273–283. (In Russian)

Kretov, A.A. (2016) «Belyi doM» i «Chernaya dyrA»: algoritm opredeleniya semanticheskoi blizostislov po ikhsochetaemosti // *Vestnik VGU, Seriya: Sistemnyi analiz I informatsionnye tekhnologii*, № 4, pp. 174–177. (In Russian)

Kretov, A.A. (2018) BRUMS kak ehlektronnyi slovar' i baza dannykh // Tezisy vserossiiskoi konferentsii «*Ot yazykovykh mashinnykh fondov k lingvisticheskim korpusam: pamyati V.M. Andryushchenko*». Institut russkogo

yazyka imeni V.V. Vinogradova. – Moskva, 28-29 sentyabrya 2018 g. http://lcl.srcc.msu.ru/library/abstracts.pdf, pp. 38-42. (In Russian)

Kretov, A. A. (2021) Parametricheskii analiz sanskritsko-russkogo slovarya» V. A. Kocherginoi / A. A. Kretov, M. YU. Gasuns, V. V. Leonchenko // *Problemy obshchei I vostokovednoi lingvistiki. Sochetaemost' yazykovykh edinits I yazykovye modeli. Pamyati Z. M. Shalyapinoi (1946-2020)*, M., IVRAN, pp. 287–302. (In Russian)

Lomov, A.M. (1977) *Ocherki po russkoi aspektologii*. Voronezh: Izdatel'stvo Voronezhskogo universiteta, 140 p. (In Russian)

Maksimenko, O.I. (2019) Avtomatizirovannyi distributivno-statisticheskii analiz kak sistemnaya obrabotka teksta // *Vestnik RUDN. Seriya: Teoriyayazyka. Semiotika. Semantika.* T.10, №1, pp. 92–100. (In Russian)

Melnikov, G.P. (1978) *Sistemologiya i yazykovye aspekty kibernetiki*. M.: Sov. radio, 368 p. (In Russian)

Melnikov, G.P. (2003) *Sistemnaya tipologiya yazykov: printsipy, metody, modeli*, M., Nauka, 395 p. (In Russian)

Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013a) Efficient Estimation of Word Representations in Vector Space. *Proceedings of Workshop at ICLR*, 2013a, pp. 1–9.

Mikolov, T., Yih, W., Zweig, G. (2013b) Linguistic Regularities in Continuous Space Word Representations. *Proceedings of NAACL HLT*, 2013b, pp. 746–751.

Nikitin, M.V. (1996) *Kurs lingvisticheskoi semantiki:* Uchebnoeposobie. SPb: Nauchnyitsentr problem dialoga, 760 p. (In Russian)

Plungyan, V.A. (2010) *Interv'yu* // Portal bogoslov.ru, Retrieved from https://bogoslov.ru/article/573240

Popova, Z.D., Sternin, I.A. (2004) *Obshchee yazykoznanie: ucheb. posobie*. Voronezh: Tsentral'no-Chernozemnoe knizhnoe izdatel'stvo, 208 p. (In Russian)

Raspopov I.P. (1973) *Ocherki po teoriisintaksisa.* Voronezh: Izdatel'stvo Voronezhskogo universiteta, 220 p. (In Russian)

Rudelev, V.G. (1980) Teoriya neitralizatsii. Nekotorye rezul'taty. Perspektivy razvitiya // *Teoriyaneitralizatsii: sb. nauch. statei.* Tambov, pp. 3–10. (In Russian)

Russkii kosmizm: Antologiya filosofskoimysli, M., Pedagogika-Press. 1993. 368 p. (In Russian)

Stepanov, YU.S. (1975) *Osnovy obshchego yazykoznaniya: Uchebnoe posobie*. M.: Prosveshchenie, 271 p. (In Russian)

Tener, L. (1988) *Basics of structural syntax*, M., Progress, 656 p.

Testelets, YA. G. (2001) *Vvedenie v obshchii sintaksis*, M., RGGU, 800 p. (In Russian)

Shaikevich, A.YA., Andryushchenko, V.V., Rebetskaya, N.A. (2016) *Distributivno-statisticheskii analiz yazyka russkoi prozy 1850-1870,* M., Yazyki slavyanskoi kultury, T. I-II. (In Russian)

Sholle, F. (2018) *Deep learningin Python*. Spb., Piter, 400 p.

Shcherba, L.V. (1974) *Yazykovaya sistema I rechevaya deyatel'nost,* L., Nauka, 427 p. (In Russian)

Sinclair, J. (1991) *Corpus Concordance Collocation*, Oxford: OUP, 179 p.

Vereshchagin, E.M. (1997) *Istoriya vozniknoveniya drevnego obshcheslavyanskogo literaturnogo yazyka: Perevodcheskaya deyatel'nost' Kirilla i Mefodiyaiikhuchenikov*, M., Martis, 314 p. (In Russian)