Тимофеева М. К. Возможности использования сервиса RusVectōrēs для выявления семантических ассоциатов глаголов русского языка / М. К. Тимофеева // Научный диалог. — 2018. — № 9. — С. 117—131. — DOI: 10.24224/2227-1295-2018-9-117-131.

Timofeeva, M. K. (2018). Possibility of Extracting Semantic Associates of Russian Verbs by the Instrument RusVectōrēs. *Nauchnyy dialog, 9:* 117-131. DOI: 10.24224/2227-1295-2018-9-117-131. (In Russ.).

УДК 81’33+811.93
DOI: 10.24224/2227-1295-2018-9-117-131

Возможности использования сервиса RusVectōrēs для выявления семантических ассоциатов глаголов русского языка

© **Тимофеева Мария Кирилловна (2018),** orcid.org/0000-0001-8999-2330, Scopus Author ID 57169993900, доктор филологических наук, старший научный сотрудник, Институт математики им. С. Л. Соболева СО РАН; заведующая кафедрой фундаментальной и прикладной лингвистики, Новосибирский государственный университет (Новосибирск, Россия), m.timofeeva@g.nsu.ru.

Представлены результаты анализа возможностей дистрибутивной семантики в её реализации в виде Web-сервиса RusVectōrēs. Рассматривается вопрос об использовании данного сервиса для изучения семантики глаголов русского языка. Актуальность исследования обусловлена недостаточностью информации о лингвистических возможностях дистрибутивной семантики в целом. Новизна состоит в том, что вопрос об использования инструмента RusVectōrēs для изучения глагольной семантики поставлен впервые. Предлагается классификация семантических отношений, полученных в результате выявления семантических ассоциатов глаголов. Материалом послужили две выборки по 30 слов, для каждой из них посредством сервиса RusVectōrēs было построено множество ассоциатов, общее количество которых составило 468 лексем. Особое внимание уделяется семантическим отношениям, описываемым в терминах лексических функций, поскольку такие ассоциаты оказались наиболее частотными для глаголов; общее количество выявленных лексических функций равно 28. Показано, что чаще всего встречаются функции, соответствующие видовым вариантам глаголов, отношениям синонимии и конверсии. Установлено, что среди семантических отношений, отличных от лексических функций, наиболее частотными являются отношения гипонимии и согипонимии, реже выявляются ситуационные связи и актанты рассматриваемых глаголов. Уделяется внимание ряду деталей, существенных для использования инструментария лексических функций применительно к семантически отношениям, выявляемым посредством сервиса RusVectōrēs.

Ключевые слова: лексическая семантика; дистрибутивная семантика; лексические функции, машинное обучение; глаголы русского языка; RusVectōrēs.

1. Введение

Развитие компьютерных средств анализа очень больших объёмов языковых данных открыло возможность разработки новых методов лингвистического исследования, в которых язык предстаёт как изменчивая реальность (Energeia), обладающая, вместе с тем, определённым постоянством [Гумбольдт, 2000, с. 70]. Возможность обработки больших объёмов данных позволяет изучать оба названных аспекта: изменчивость языка при его использовании в разных сферах и наличие инвариантных свойств, характерных для любой сферы.

Примером успешного использования в лингвистике методов работы с большими объёмами языковых данных являются корпусные исследования. В настоящее время существует ряд общедоступных ресурсов с удобным интерфейсом, предоставляющих широкие возможности применения в области корпусной лингвистики. Варьируя состав рассматриваемых корпусов и / или способ их разметки, можно изучать особенности речевой деятельности в разных её реализациях, либо исследовать её инвариантные свойства.

Инвариантные свойства могут быть, в частности, операциональными, то есть представлять собой определённые последовательности операций, производимых с языковыми данными и приводящих к лингвистически значимым результатам.

Одним из примеров операциональных инвариантных свойств можно считать исследовательские модели, имитирующие деятельность лингвиста. Такие модели воспроизводят процедуры, используемые лингвистами для обнаружения языковых явлений [Апресян, 1966, с. 99—100]. Исходным материалом анализа является текст, возможно, сегментированный на правильно построенные фразы или размеченный определённым образом. Применение исследовательских процедур к таким входным данным должно приводить к обнаружению (изначально неизвестных) грамматических и лексических структурных составляющих рассматриваемого языка. Впервые целенаправленной разработкой такой модели начали заниматься дескриптивные лингвисты. Они предложили исследовательские процедуры, основанные на анализе дистрибуции, и стремились сделать эти процедуры максимально формализованными, как можно более независимыми от субъективности исследователя и от смысла рассматриваемых текстов. Понятно, что максимальная независимость такого рода достигается тог-

да, когда применение исследовательских операций осуществляет компьютер. Поэтому в случае компьютерной реализации методов дескриптивной лингвистики её можно было бы отнести к ранним этапам развития машинного обучения. Сами дескриптивные лингвисты, конечно, в таких терминах свой подход не обсуждали.

Дистрибутивная семантика, основывающаяся на анализе совместной встречаемости слов в текстах и родственная идеям дескриптивной школы американской лингвистики, относится к числу методов машинного обучения. Возможности дистрибутивной семантики в настоящее время активно исследуют в связи с разработкой систем информационного поиска и другими прикладными задачами общего назначения, для решения которых требуется автоматическая обработка текстов естественного языка.

Методы машинного обучения могут со временем стать столь же полезным инструментом исследования языка, как и корпусы, однако пока их потенциал для собственно лингвистических исследований остаётся неясным. Тем не менее можно привести ряд примеров использования методов машинного обучения для решения задач, являющихся по своей сути лингвистическими.

Такие методы, например, нашли применение и демонстрируют неплохие результаты в системах автоматического перевода Google и Яндекс [Калинин, 2017]. Машинное обучение применяется для автоматического выявления актантов [Кузнецов, 2016], обучение здесь происходит на основе корпуса примеров из FrameBank [Ляшевская, 2009].

Пример использования дистрибутивной семантики для решения по сути лингвистической задачи описан в [The emergence …, 2010; A quantitative philology …, 2012]. Здесь применён один из вариантов дистрибутивной семантики — латентный семантический анализ, или LSA (Latent Semantic Analysis), — базирующийся на гипотезе о том, что семантически связанные слова должны совместно встречаться в текстах сходной тематики. Данный метод использован для изучения динамики употребительности слов и словосочетаний, указывающих на интроспекцию, в текстах разных эпох, начиная с периода 800—200 годов до нашей эры и заканчивая XX веком нашей эры. Как показало исследование, обращения к интроспекции с течением времени растут, авторы связывают это с появлением и развитием письменности (и соответственно необходимости запоминания и рационального осмысления текстов), культурным и ментальным развитием, анатомическими мозговыми изменениями.

Изучение лингвистического потенциала дистрибутивной семантики требует времени, так как реализация данного метода может происходить

в разных режимах, на основе разных алгоритмов, разных принципов сбора и разметки текстов. Настоящая статья посвящена исследованию лингвистических возможностей общедоступного Web-сервиса RusVectōrēs [Kutuzov et al, 2017], реализующего алгоритмы дистрибутивной семантики.

Цель состоит в анализе и классификации семантических ассоциатов, выявляемых для глаголов русского языка посредством RusVectōrēs. Ранее были рассмотрены типы семантических ассоциатов, выявляемых посредством указанного сервиса для существительных [Тимофеева, 2018]. Для глаголов аналогичная задача не ставилась. Изучение данного вопроса поможет понять возможности использования сервиса в области глагольной семантики.

2. Базовые понятия, материалы и методы

При использовании методов машинного обучения применительно к материалу естественного языка компьютер «обучается» выполнению определённых действий на основе анализа очень больших объёмов текстов. На том же принципе работы основаны методы дистрибутивной семантики. Семантика слова здесь характеризуется вектором в многомерном пространстве текстов, позиции в составе таких векторов отражают совместную встречаемость слов в контекстах заданной длины. Реализация общих идей дистрибутивной семантики имеет варианты, можно выбирать тип и размер рассматриваемого контекста, используемые алгоритмы, корпусы, по которым производится обучение. Работа сервиса RusVectōrēs основана на использовании моделей, построенных по текстам с морфологической разметкой. В результате морфологической разметки словоформы одной лексемы объединяются и представлены в тексте начальной формой лексемы.

Поскольку задача классификации семантических отношений, выявляемых для глаголов русского языка посредством сервиса RusVectōrēs, ранее не ставилась и результат заранее был неясен, сначала рассматривалась пробная выборка. Задачи анализа пробной выборки состояли в следующем:

1) исходя из того, что отношение синонимии является одним из базовых семантических отношений в области лексики, выяснить, насколько эффективен сервис для выявления синонимов глаголов;

2) выделить класс наиболее часто выявляемых сервисом семантических отношений и исследовать возможности их выявления более детально на дополнительной выборке.

Для составления пробной выборки были взяты 30 первых по порядку опорных слов-глаголов из «Словаря синонимов русского языка» [СС, 1970]. Каждому из них сервис поставил в соответствие по 10 семантически близких слов (что составляет столько же семантических отноше-

ний). Из числа набранных 300 слов были исключены ошибочные слова (обусловленные ошибками в текстах или недостатками работы алгоритмов морфологического анализа), дубликаты и словоформы, относящиеся к одной лексеме. В итоге объём пробной выборки составил 228 лексем и столько же семантических отношений.

Далее на основе пробной выборки были определены принципы составления и обработки дополнительной выборки. Вторая выборка также состояла из 30 первоначально отобранных слов (в число которых входило 11 слов из «Толково-комбинаторного словаря русского языка» (ТКС) [ТКС, 2016]), для каждого слова сервис выявил по 10 семантически близких слов, связанных семантическими отношениями с данным словом. В сумме это составило 300 слов. После исключения ошибочных слов, дубликатов и лишних словоформ в данной выборке осталось 240 слов и столько же семантических отношений. Таким образом, суммарное количество рассмотренных семантических отношений составило 468.

При классификации выявленных для глаголов семантических отношений использовались сведения из «Словаря синонимов русского языка» [СС, 1970], электронного тезауруса RuWordNet [Лукашевич, 2011] и «Толково-комбинаторного словаря русского языка» [ТКС, 2016].

Анализ материала первой выборки показал, что среди выявленных семантических отношений многие подходят под понятие лексической функции (61,84 %). Поэтому вторая выборка составлялась с целью более детального изучения возможностей выявления лексических функций.

Понятие лексической функции было введено в модели автоматического перевода «СМЫСЛ ⇔ ТЕКСТ» [Мельчук, 1974], послужившей основой для разработки «Толково-комбинаторного словаря русского языка» (ТКС) [ТКС, 2016]. Каждая лексическая функция соответствует определённому регулярному семантическому отношению между словами / словосочетаниями и отражает идиоматику рассматриваемого языка. Относительно общего количества функций точной оценки нет, в ТКС использованы 62 элементарные функции, на основе которых могут строиться сложные функции. Сложная функция определяется как «такая комбинация синтаксически связанных простых лексических функций, которые имеют единое лексическое выражение, покрывающее смысл всей комбинации целиком» [ТКС, 2016, с. 107].

Лексические функции описываются в виде формул вида $F(w) = \{w_1,\ldots,w_n\}$, где $w_1,\ldots, w_n$ — слова или словосочетания, являющиеся возможными результатами применения лексической функции F к слову или словосочетанию w. Если возможный результат применения функции единственен, то фигурные скобки можно не использовать. К числу хорошо известных

элементарных лексических функций относятся функции Syn и Anti, то есть отношения синонимии и антонимии соответственно:

Syn (*языкознание*) = {*лингвистика*, *языковедение*};

Anti (*надеяться*) = *сомневаться*.

Хотя, строго говоря, значением лексической функции является множество, этот же термин применяется и к элементам данного множества [Мельчук, 1974, с. 101], например, допустимо говорить, что *лингвистика* — значение функции Syn (*языкознание*).

Ниже кратко описаны только те элементарные лексические функции, которые встретились при анализе семантических отношений, выявленных посредством RusVectōrēs. Более полный список стандартных лексических функций описан в [ТКС, 2016, с. 98—109] и [Мельчук, 1974, с. 78—133]. Сервис RusVectōrēs работает на уровне слов, а не словосочетаний. Соответственно в роли аргумента w лексической функции в нашем случае может выступать только лексема. Обозначим эту лексему символом L.

Conv (L) — конверсив, то есть лексема, по смыслу совпадающая с лексемой L, но предполагающая другой порядок актантов. Изменение порядка актантов указывается в виде числового индекса, например, $Conv_{21}$ (*восхищаться*) = *восхищать*. Таким образом предполагается, что для данной лексемы стандартный порядок следования актантов таков: на первом месте субъект действия, представленного глаголом, на втором месте — объект этого действия. В результате конверсии субъект и объект меняются местами. Для глаголов с большим числом актантов возможно несколько вариантов перестановок.

S_0 (L), A_0 (L), V_0 (L) — синтаксические дериваты лексемы L, совпадающие с ней по смыслу и являющиеся соответственно существительным, прилагательным, глаголом. Например, S_0 (*просить*) = *просьба*, A_0 (*восторгаться*) = *восторженный*, V_0 (*учиться*) = *обучаться*. Значением функции S_0 не обязательно должен быть словообразовательный вариант лексемы L (например, в [ТКС, 2016, с. 415] приведено значение S_0 (*светать*) = *заря*), однако такие случаи в рассмотренном материале не встретились.

Значением функции S_i (L) является типовое название i-го глубинно-синтаксического актанта лексемы L, например, S_1 (*учиться*) = *ученик*, S_3 (*учиться*) = *учитель*.

Значением функции A_i (L) является типовое свойство i-го глубинно-синтаксического актанта лексемы L, например, A_2 (*понимать*) = *понятный*.

Функция Magn (L) служит для выражения интенсивности смысла s, обозначенного лексемой L (s в высшей степени проявления), например, Magn (*обещать*) = *твёрдо*.

Если L — лексема со сложным смыслом, предполагающим несколько шкал интенсивности, то возможно несколько вариантов функции Magn, которые обычно различаются посредством индексов. Например, $Magn^{temp}$ (*дождь*) = *затяжной*, $Magn^{капли}$ (*дождь*) = *крупный*. Мы не будем рассматривать подобные нюансы, так как на имеющемся материале это приводит к слишком дробной классификации. По той же причине не будут учитываться индексы функции Conv.

Функции Incep (L) и Fin (L) служат для обозначения соответственно смыслов «начало процесса, обозначенного лексемой L» и «завершение процесса, обозначенного лексемой L», например, Incep (*спать*) = *засыпать*, Fin (*гневаться*) = *остывать*.

Функции Imperf (L) и Perf (L) соответствуют незавершённости и завершённости действия, например, Perf (*понимать*) = *понять*, Imperf (*прыгнуть*) = *прыгать*.

Функция Gener (L) служит для обозначения родового по отношению к L понятия. Например, Gener (*сомневаться*) = *чувство*. Формальные критерии установления этой функции описаны в [Мельчук, 1974, с. 84—85].

Функции Real (L) и Result (L)обозначают соответственно реализацию и результат процесса L, например, Real (*учиться*) = *научиться*, Result (*учиться*) = {*знать*, *уметь*}.

Примеры сложных лексических функций: ConvPerf (*удивлять*) = *удивиться*, SynConv (*учиться*) = *преподавать*.

Кроме лексических функций, при построении классификации семантических отношений были использованы отношения гипонимии, согипонимии, ситуационной связанности понятий, отношения с потенциальными глубинно-синтаксическими актантами глагола. Понятие X считалось гипонимом понятия Y, если X представляет собой частную форму реализации действия или процесса Y, например, *спрыгнуть* — гипоним по отношению к *прыгнуть*. Понятие X считалось согипонимом понятия Y, если оба эти понятия подпадают под одно и то же родовое понятие, например, *интересовать* и *волновать*. Отношение ситуационной связанности имеется, например, между понятиями, обозначенными словами *обещать* и *просить*. Семантическая связь между глаголом и его потенциальными глубинно-синтаксическими актантами была выявлена, например, для глагола *плясать*, которому сервис поставил в соответствие потенциальные актанты *цыганочка* и *джига*.

Основная задача анализа второй выборки состояла в изучении применимости сервиса RusVectōrēs для выявления лексических функций. Поэтому при построении классификации семантических отношений в первую

очередь рассматривалась возможность трактовки каждого отношения как реализации лексической функции. Только в том случае, если это оказывалось невозможным, семантическое отношение включалось в другой класс, то есть могло трактоваться как гипоним, согипоним, ситуационная связь, актантная связь.

3. Анализ результатов

Среди семантических отношений важное место занимает синонимия. Анализ результатов использования сервиса RusVectōrēs для слов из пробной выборки показал, что для 30 опорных слов-глаголов из «Словаря синонимов русского языка» [СС, 1970], входящих в эту выборку, было обнаружено 17,61 % словарных синонимов, то есть эффективность сервиса для выявления таких синонимов для глаголов не очень высока. Для второй выборки аналогичная оценка не проводилась, так как слова выбирались не из синонимического словаря.

Вместе с тем синонимия — одно из наиболее часто выявляемых сервисом семантических отношений.

В построенной классификации множество выявленных синонимов складывается из синонимов-дериватов (результатов применения к исходному глаголу функции V_0) и синонимов, не являющихся дериватами (результатов применения функции Syn).

Например, V_0 (*плясать*) = {*выплясывать*, *отплясывать*}. В тезаурусе RuWordNet указанные три глагола считаются синонимами. Тем самым использование лексических функций разделило отношение синонимии на собственно синонимию, представленную не дериватами, и синонимию дериватов, в сумме это даёт 29 случаев, что составляет 12,08 % от числа всех выявленных при анализе второй выборки семантических отношений или 20,57 % от числа всех семантических отношений, отнесённых к числу лексических функций.

Сервис RusVectōrēs работает на уровне лексики, поэтому разные значения слова неразличимы. Синонимом считалось слово, являющееся таковым хотя бы для одного из значений.

Поскольку доля лексических функций среди выявленных для пробной выборки семантических отношений оказалась довольно высокой, была поставлена задача проследить, какие лексические функции чаще выявляются посредством RusVectōrēs. Для этого была составлена дополнительная выборка. Она состояла из тех семантических групп глаголов, для которых на пробной выборке было выявлено большее количество лексических функций. Это глаголы, обозначающие конкретные действия, осуществляемые человеком,

прежде всего глаголы движения (например, *прыгать, ехать, плясать, смеяться, взять*), говорения (например, *сказать, обещать, спорить*), ментального или эмоционального состояния (например, *понимать, интересовать, досадовать, восхищать, удивлять, надеяться*). Среди рассмотренных в составе дополнительной выборки слов было 11 глаголов, содержащихся в ТКС (*варить, восхищать, досадовать, завтракать, обещать, надеяться, победить, понимать, сниться, удивлять, учиться*).

Результаты анализа дополнительной выборки (табл. 1) показывают, что доля лексических функция среди выявленных семантических отношений близка к той, которая получена для пробной выборки (для пробной выборки 61,84 %, для дополнительной выборки 58,75 %). Имеются функции, выявленные в первой выборке, но отсутствующие во второй (например, Caus, CausFin, AntiBon, Sloc). Некоторое уменьшение доли выявленных лексических функций во второй выборке свидетельствует о том, что необходимо дальнейшее исследование и уточнение семантических типов глаголов, применительно к которым эффективно использовать сервис для выявления лексических функций.

Таблица 1

Типы семантических отношений

Тип семантического отношения	Количество	%
Лексическая функция	141	58,75
Согипонимия	40	16,67
Гипонимия	31	12,92
Ситуационная связь	18	7,50
Актант	10	4,17
Итого	240	

Поскольку использование дополнительной выборки было нацелено на более детальное изучение возможности выявления лексических функций, при построении классификации предпочтение отдавалось именно такой интерпретации отношений. Например, отношение между глаголами *сигать* и *прыгать* в электронном тезаурусе RuWordNet обозначено как синонимия (RuWordNet не опирается на инструментарий лексических функций). При анализе результатов дополнительной выборки данное отношение трактовалось как лексическая функция: Magn (*прыгать*) = *сигать*.

В результате обращение к лексическим функциям сузило понятия синонимии и гипонимии, а также заменило отношение гиперонимии на функцию Gener. Так, лексемы *интересовать* и *заинтересовывать* от-

несены в электронном тезаурусе RuWordNet к числу синонимов. В построенной классификации они не считаются синонимами, а связаны как аргумент и значение лексической функции: Incep (*интересовать*) = *заинтересовывать*. В тезаурусе *изобразить* считается гиперонимом по отношению к *рисовать*, в построенной классификации использовано соотношение PerfGener (*рисовать*) = *изобразить*. Видовые формы в RuWordNet часто включаются в число синонимов, вместо этого в построенной классификации использованы лексические функции Imperf и Perf, например, Perf (*восхищать*) = *восхитить*), Imperf (*услышать*) = *слышать*.

Перечень лексических функций, выявленных посредством сервиса RusVectōrēs для второй выборки, включает 28 функций (табл. 2). Среди них 17 элементарных функций, что составляет 27 % от числа всех элементарных лексических функций, приведённых в ТКС. Функции Perf, Syn, Conv, V_0 встретились наибольшее число раз. Кроме элементарных лексических функций, выявлено 11 сложных функций.

Надо заметить, что некоторые функции принципиально не могли встретиться на рассмотренном материале. Это прежде всего функции типа Oper, Func, Labor, аргументами которых являются существительные. В силу ограничений сервиса не могли встретиться также функции, значениями которых являются предлоги (Loc, Instr, Propt).

Таблица 2

Типы лексических функций

	Название	Количество	% (от 141)	Название	Количество	% (от 141)
1	Perf	22	15,60	A_0	2	1,42
2	Syn	18	12,77	A_2	2	1,42
3	Conv	17	12,06	Anti	2	1,42
4	V_0	11	7,80	Fin	2	1,42
5	S_0	9	6,38	AntiMagn	2	1,42
6	Gener	9	6,38	ImperfSyn	2	1,42
7	Incep	7	4,96	IncepSyn	2	1,42
8	PerfSyn	7	4,96	ImperfConv	1	0,71
9	Imperf	5	3,55	IncepConv	1	0,71
10	Magn	4	2,84	ConvPerf	1	0,71
11	S_2	3	2,13	PerfGener	1	0,71
12	IncepPerf	3	2,13	Result	1	0,71
13	Real	3	2,13	ResultPerf	1	0,71
14	S_1	2	1,42	SynConv	1	0,71
Всего лексических функций: 141						

Количественное преобладание функции Perf по сравнению с функцией Imperf объясняется тем, что во второй выборке представлены преимущественно глаголы несовершенного вида. Для глаголов совершенного вида сервис выявлял лексическую функцию Imperf. Надо отметить, что в некоторых случаях переход к глаголу другого вида добавляет определённые смысловые элементы, соответственно результат уже трактуется как функция, отличная от Perf или Imperf. Например, для глагола *интересовать* найден ассоциат *заинтересовать*, который трактуется как сложная функция: IncepPerf (*интересовать*) = *заинтересовать*, аналогично Perf (*обещать*) = *пообещать*, PerfSyn (*обещать*) = *посулить*. Сложно разграничивать функции Perf и Real. Так в статье ТКС для слова *учиться* (ТКС, 2016, с. 508—510) найденные сервисом RusVectōrēs ассоциаты *доучиться*, *проучиться*, *выучиться* не трактуются как функции Perf: *выучиться* и *доучиться* описываются посредством функции Real, *проучиться* — как нестандартная лексическая функция. Чтобы избежать излишне дробной классификации семантических отношений, нестандартные лексические функции не рассматривались, поэтому ассоциат *проучиться* был также отнесён к функции Real. Видовые ассоциаты для нескольких глаголов (например, *сказать*, *надеяться*, *плясать*) выявлены не были.

Количество выявленных лексических функций было бы несколько больше, если бы рассматривались альтернативные варианты трактовок. Поскольку грани между функциями зачастую очень тонкие, в некоторых случаях они имеют одинаковую лексическую реализацию, например, Perf (*обещать*) = Sing (*обещать*) = *пообещать* (ТКС, 2016, с. 301). В таких случаях учитывался только один вариант трактовки, предпочтение отдавалось наиболее частотному из них, в данном случае была выбрана функция Perf.

4. Выводы

1. Основной результат проделанной работы состоит в демонстрации возможности использования сервиса RusVectōrēs для выявления лексических функций глаголов. Анализ дополнительной выборки позволил более детально изучить типологические и количественные характеристики лексических функций глагола, выявляемых посредством рассматриваемого сервиса.

2. Количество найденных лексических функций сравнительно велико для глаголов, обозначающих конкретные действия человека (движение и другие физические действия, ментальные и эмоциональные состояния, коммуникацию). Однако вопрос о семантических типах глаголов, для которых сервис наиболее эффективен в плане выявления лексических функций, нуждается в уточнении.

3. Анализ состава и частот выявленных лексических функций показал, что чаще всего в число ассоциатов попадают функции Perf / Imperf, Syn / V_0, Conv (в сумме 51,77 % от количества всех выявленных лексических функций). Почти все встретившиеся сложные функции также содержат эти элементы: IncepPerf, IncepSyn, IncepConv, SynConv, ImperfConv, PerfGener, ImperfSyn, ResultPerf, ConvPerf. В перспективе, при увеличении объёма текста для машинного обучения, количество выявляемых лексических функций может возрасти как количественно, так и качественно.

4. Сервис выявляет довольно большое количество синонимов, однако сравнение с данными словаря синонимов русского языка [СС, 1970] показывает, что доля обнаруживаемых словарных синонимов не очень велика (17,61 %).

5. Среди семантических отношений, отличных от лексических функций, наиболее часто выявляются гипонимы и согипонимы (в сумме 29,58 %). Эти отношения были наиболее частотными для семантических ассоциатов существительных, выявляемых посредством сервиса RusVectōrēs [Тимофеева, 2018]. Для глаголов доля гипонимов и согипонимов оказалась меньше, так как некоторые из таких семантических отношений попали в число лексических функций.

Проведённый анализ показал, что сервис RusVectōrēs может быть полезен для исследований в области глагольной семантики. Представляется актуальным также изучение возможностей применения дистрибутивной семантики для выявления авторских / стилистических особенностей, формулируемых в терминах лексических функций.

Источники и принятые сокращения

1. *Веб-сервис* дистрибутивно-семантических моделей для русского языка RusVectōrēs [Электронный ресурс]. — Режим доступа : http://rusvectores.org/ru/.

2. СС — *Словарь* синонимов русского языка : в 2 томах / под ред. А. П. Евгеньевой. — Ленинград : Наука, 1970. — Т. 1. — 680 с.

3. *Тезаурус* русского языка RuWordNet [Электронный ресурс]. — Режим доступа : http://www.ruwordnet.ru/ru/.

4. ТКС — *Мельчук И. А.* Толково-комбинаторный словарь русского языка : опыты семантико-синтаксического анализа русской лексики / И. А. Мельчук, А. К. Жолковский. — Москва : Глобал Ком : Языки славянской культуры, 2016. — 544 с.

Литература

1. *Апресян Ю. Д.* Идеи и методы современной структурной лингвистики (краткий очерк) / Ю. Д. Апресян. — Москва : Просвещение, 1966. — 305 с.

2. *Гумбольдт В*. Избранные труды по языкознанию / В. Губольдт. — Москва : Прогресс, 2000. — 400 с.

3. *Калинин С. М*. Обзор современных подходов к улучшению точности нейронного машинного перевода / С. М. Калинин. — Rhema. Рема. — 2017. — № 2. — С. 70—79.

4. *Кузнецов И. О.* Автоматическая разметка семантических ролей в русском языке : диссертация … кандидата филологических наук / И. О. Кузнецов. — Москва, 2016. — 178 с.

5. *Лашевская О. Н*. Семантические роли и сеть конструкций в системе FrameBank / О. Н. Ляшевская, Е. В. Кашкин // Компьютерная лингвистика и интеллектуальные технологии : материалы ежегодной конференции «Диалог». — Москва : РГГУ, 2013. — С. 827—846.

6. *Лукашевич Н. В*. Тезаурусы в задачах информационного поиска / Н. В. Лукашевич. — Москва : Издательство Московского университета, 2011. — 512 с.

7. *Мельчук И. А.* Опыт теории лингвистических моделей «СМЫСЛ ⇔ ТЕКСТ» : семантика, синтаксис / И. А. Мельчук. — Москва : Наука, 1974. — 315 с.

8. *Тимофеева М. К.* Типология семантических отношений, выявляемых посредством инструмента RusVectōrēs / М. К. Тимофеева // Научный диалог. — 2018. — № 8. — С. 74—87.

9. *A quantitative* philology of introspection / C. G. Diuk, D. F. Slezak, L. Raskovsky, M. Sigman, G. A. Cecchi // Frontiers in integrative neuroscience. — 2012. — Vol. 329, No. 5998. — P. 1541—1543.

10. *Kutuzov A*. Webectors: A Toolkit for Building Web Interfaces for Vector Semantic Models / A. Kutuzov, E. Kuzmenko // Ignatov D. et al. (eds). Analysis of Images, Social Networks and Texts. Series : Communications in Computer and Information Science : Proceedings 5th International Conference, AIST 2016, 7—9 April 2016 g., g. Yekaterinburg. — Springer, Cham, 2017. — Vol. 661 — P. 155—161.

11. *The emergence* of the modern concept of introspection : a quantitative linguistic analysis / L. Raskovsky, D. F. Slezak, C. G. Diuk, G. A. Cecchi // Young investigator workshop on computational approaches to languages of the Americas : Proceedings of the NAACL. — Los Angeles, California, 2010. — P. 68—75.

Possibility of Extracting Semantic Associates of Russian Verbs by the Instrument RusVectōrēs

© Timofeeva Mariya Kirillovna (2018), orcid.org/0000-0001-8999-2330, Scopus Author ID 57169993900, Doctor of Sciences in Philology, senior researcher of the Sobolev Institute of mathematics, Siberian Branch of the Russian Academy of Sciences; Head of the Section of Fundamental and Applied Linguistics of the Institute of Humanities, Novosibirsk State University (Novosibirsk, Russia), m.timofeeva@g.nsu.ru.

The paper presents the results of investigating distributional semantics' potential in its realization in the form Web-service RusVectōrēs. The question about applying the service for studying semantics of Russian verbs is considered. The research urgency is caused by insufficient level of information about linguistic possibilities of distributional semantics in whole. The novelty of the research consist in the fact that the question about applying the instrument RusVectōrēs for investigating verb semantics is posed for the first time. Classification of semantic relations extracted by RusVectōrēs for Russian verbs is proposed. The analyzed data include two lists of entry verbs and the set of semantic associates for each verb; the integral set of considered semantic associates consists of 468 verbs. Special attention is paid to semantic relations that can be treated as lexical functions because this sort of relations appeared to be the most frequent for Russian verbs; the whole number of extracted lexical functions is equal to 28. It is shown that lexical functions that correspond to aspectual variants of verbs, to synonymic relations and conversion are the most frequent; hyponyms and co-hyponyms are the most frequent among semantic relations that differ from lexical functions; situational relations and actant relations are comparatively rare. Special attention is paid to the details that are important for applying the instruments of lexical functions to semantic relations extracted for verbs by the service RusVectōrēs.

Key words: lexical semantics; distributional semantics; lexical functions; machine learning; Russian verbs; RusVectōrēs.

Material resources

SS — Evgenyeva, A. P. (ed.). (1970). *Slovar' sinonimov russkogo yazyka*. Leningrad: Nauka. (In Russ.).

Tezaurus russkogo yazyka RuWordNet. Available at: http://www.ruwordnet.ru/ru/. (In Russ.).

TKS — Melchuk, I. A., Zholkovskiy, A. K. (2016). *Tolkovo-kombinatornyy slovar' russkogo yazyka: opyty semantiko-sintaksicheskogo analiza russkoy leksiki*. Moskva: Global Kom: Yazyki slavyanskoy kultury. (In Russ.).

Veb-servis distributivno-semanticheskikh modeley dlya russkogo yazyka RusVectōrēs. Available at: http://rusvectores.org/ru/. (In Russ.).

References

Apresyan, Yu. D. (1966). *Idei i metody sovremennoy strukturnoy lingvistiki (kratkiy ocherk)*. Moskva: Prosveshcheniye. (In Russ.).

Diuk, C. G., Slezak, D. F., Raskovsky, L., Sigman, M., Cecchi, G. A. (2012). A quantitative philology of introspection. *Frontiers in integrative neuroscience, 329/5998:* 1541—1543.

Gumboldt, V. (2000). *Izbrannyye trudy po yazykoznaniyu*. Moskva: Progress. (In Russ.).

Kalinin, S. M. (2017). Obzor sovremennykh podkhodov k uluchsheniyu tochnosti neyronnogo mashinnogo perevoda. *Rhema. Rema, 2:* 70—79. (In Russ.).

Kutuzov, A., Kuzmenko, E. (2017). Webectors: A Toolkit for Building Web Interfaces for Vector Semantic Models. In: Ignatov, D. et al. (eds). *Analysis of Images, Social Networks and Texts. Series: Communications in Computer and Information Science, 661*: *Proceedings 5th International Conference, AIST 2016, 7—9 April 2016 g., g. Yekaterinburg*. Springer, Cham. 155—161.

Kuznetsov, I. O. (2016). *Avtomaticheskaya razmetka semanticheskikh roley v russkom yazyke: dissertatsiya…* kandidata filologicheskikh nauk. Moskva. (In Russ.).

Lashevskaya, O. N., Kashkin, E. V. (2013). Semanticheskiye roli i set' konstruktsiy v sisteme FrameBank. In: *Kompyuternaya lingvistika i intellektualnyye tekhnologii: materialy ezhegodnoy konferentsii «Dialog»*. Moskva: RGGU. 827—846. (In Russ.).

Lukashevich, N. V. (2011). *Tezaurusy v zadachakh informatsionnogo poiska*. Moskva: Izdatelstvo Moskovskogo universiteta. (In Russ.).

Melchuk, I. A. (1974). *Opyt teorii lingvisticheskikh modeley «SMYSL ⇔ TEKST»: semantika, sintaksis*. Moskva: Nauka. (In Russ.).

Raskovsky, L., Slezak, D. F., Diuk, C. G., Cecchi, G. A. (2010). The emergence of the modern concept of introspection: a quantitative linguistic analysis. In: *Young investigator workshop on computational approaches to languages of the Americas: Proceedings of the NAACL*. Los Angeles, California. 68—75.

Timofeeva, M. K. (2018). Tipologiya semanticheskikh otnosheniy, vyyavlyaemykh posredstvom instrumenta RusVectōrēs [Typology of Semantic Relations Extracted by the Instrument RusVectōrēs]. *Nauchnyy dialog, 8:* 74—87. (In Russ.).