Тимофеева М. К. Типология семантических отношений, выявляемых посредством инструмента RusVectōrēs / М. К. Тимофеева // Научный диалог. — 2018. — № 8. — С. 74—87. — DOI: 10.24224/2227-1295-2018-8-74-87.

Timofeeva, M. K. (2018). Typology of Semantic Relations Extracted by the Instrument RusVectōrēs. *Nauchnyy dialog, 8:* 74-87. DOI: 10.24224/2227-1295-2018-8-74-87. (In Russ.).

УДК 81’33+811.93
DOI: 10.24224/2227-1295-2018-8-74-87

Типология семантических отношений, выявляемых посредством инструмента RusVectōrēs

© **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.

Рассматривается вопрос о типах семантических отношений, выявляемых посредством Web-сервиса RusVectōrēs. Актуальность исследования обусловлена малой изученностью лингвистических возможностей данного инструмента и дистрибутивной семантики в целом. Новизна исследования видится в том, что впервые поставлена задача анализа видов семантических отношений, выявляемых на основе дистрибутивной семантики, то есть метода, опирающегося на автоматическую обработку очень больших объемов текстов, без использования лингвистической информации. Предлагается классификация семантических отношений, включающая 8 типов парадигматических и 3 типа синтагматических отношения. Показано, что в результате применения рассматриваемого метода выявляются преимущественно парадигматические отношения, причем чаще всего обнаруживаются гипонимы и согипонимы заданного слова. Рассматривается вопрос о том, зависят ли типы выявляемых семантических отношений от того, к какой семантической группе относится слово, для которого эти отношения ищутся. Показано, что такая зависимость просматривается, хотя и требует дополнительной проверки на более представительном материале. Сравнительный анализ пяти семантических групп слов позволил подтвердить общую закономерность (о преобладании гипонимов и согипонимов) и выдвинуть некоторые частные гипотезы относительно корреляций между семантическим типом задаваемого слова и типами выявляемых для него семантических отношений.

Ключевые слова: семантические отношения; лексика; дистрибутивная семантика; автоматический анализ текста; существительные; RusVectōrēs.

1. Введение

Дистрибутивная семантика в настоящее время является одним из наиболее активно исследуемых средств автоматического выявления семантических отношений между лексическими единицами. Данный метод состоит в анализе совместной встречаемости лексем в больших корпусах текстов. Алгоритмы анализа используют средства математической статистики и алгебры, лингвистическая информация присутствует только в морфологической разметке корпуса, причем разметка также осуществляется автоматически, поскольку вручную проанализировать такой большой объем текстов невозможно.

Разработкой и применением инструментов дистрибутивной семантики занимаются главным образом представители сферы IT-технологий, в таких исследованиях активно участвуют компьютерные лингвисты. Дистрибутивная семантика используется при автоматической обработке текстов естественного языка, причем область ее применения расширяется, включая, в частности, изучение процессов исторического изменения языка [Kutuzov et al., 2017], обнаружения текстовых заимствований [Belyy et al., 2018], извлечения коллокаций [Enikeeva et al., 2018], разрешения кореференций [Toldova et al., 2017].

Однако в других направлениях языкознания, вне сферы компьютерной лингвистики, дистрибутивная семантика пока малоизвестна, возможности использования данного инструмента для собственно лингвистических исследований неясны. Это обусловлено прежде всего тем, что данный метод получил распространение сравнительно недавно и создавался главным образом применительно к решению задач общего назначения. Он не предназначен специально для лингвистических исследований, как, например, лингвистические корпусы. Сервис RusVectōrēs делает данный метод доступным для любых пользователей.

Цель статьи: рассмотреть лингвистические аспекты дистрибутивной семантики на примере сервиса RusVectōrēs, проанализировав типы семантических отношений, выявляемые для существительных русского языка.

Необходимо иметь в виду, что корпусы, используемые для построения семантических моделей, могут быть больше по объему, чем в RusVectōrēs, и это потенциально способно повысить качество выявления семантических отношений, особенно для редких слов. Поэтому нижеследующий текст следует рассматривать как обсуждение определенных гипотез относительно лингвистических аспектов применения дистрибутивной семантики, сформулированных в результате анализа варианта этого метода, представленного указанным сервисом.

Научная новизна исследования состоит в том, что анализ типов семантических отношений, выявляемых на основе использования дистрибутивной семантики, ранее не осуществлялся. Фактически ставится задача ввести в арсенал лингвистических средств изучения лексической семантики новый метод, продемонстрировав некоторые из его возможностей.

Использованные при решении этой задачи методы включали прежде всего метод выявления семантически близких слов посредством сервиса RusVectōrēs. Кроме того, привлекались методы идентификации семантических отношений, разработанные в ходе построения электронных тезаурусов типа WordNet.

2. Общая характеристика методологии дистрибутивной семантики

Метод дистрибутивной семантики, по сути, можно рассматривать как развитие — на новом уровне технических возможностей — метода дистрибутивного анализа, использовавшегося в дескриптивном направлении американской лингвистики XX века [Глисон, 1959], а также более поздних модификаций данного метода во второй половине XX века, например, в статистико-комбинаторной модели Н. Д. Андреева [Андреев, 1967], теории замещаемости А. В. Гладкого [Гладкий, 1973, с. 314—342], понятии семейства как объединения дистрибутивно близких слов в модели И. И. Ревзина [Ревзин, 1967, с. 87—96], дешифровочной модели языка Б. В. Сухотина [Сухотин, 1984]. Модель Б. В. Сухотина была реализована в виде компьютерных программ, осуществляющих автоматическую сегментацию текста произвольного (неизвестного) языка, причем алгоритм анализа использовал только статистические характеристики заданного текста и не опирался ни на какие лингвистические сведения. Основу всех перечисленных теорий и моделей, в том числе основу дистрибутивной семантики, составляет так называемая дистрибутивная гипотеза, авторство которой обычно приписывают З. Харрису [Harris, 1962]. Согласно этой гипотезе любое явление языка можно обнаружить на основе исследования сочетаемости языковых единиц. Однако современная дистрибутивная семантика имеет существенные отличия от более ранних (лингвистических) подходов к анализу дистрибуции.

Во-первых, в дистрибутивной семантике в качестве контекста встречаемости лексемы выступает не непосредственный небольшой контекст (например, предложение), а контекст произвольно заданной длины, в который входит данная лексема. При этом порядок следования слов в текстах не учитывается, а словоформы заменяются на леммы (начальные формы лексем) посредством применения программ автоматического морфологического анализа.

Во-вторых, объем текстов, по которым выявляются дистрибутивные характеристики лексем, очень велик: в RusVectōrēs от 150 до 600 тысяч лемм, но другие модели дистрибутивной семантики могут строиться на основе более объемных корпусов. Автоматический анализ таких объемов текстов в дескриптивной лингвистике и указанных ее модификациях был невозможен ввиду отсутствия в то время соответствующей технической базы.

В-третьих, значительно усложнен алгоритм анализа дистрибуции лексем. В дистрибутивной семантике информация о том, в каких текстах (контекстах) встретилась рассматриваемая лексема, определяет параметры вектора в многомерном пространстве текстов. Каждой лексеме ставится в соответствие вектор, близость слов по смыслу оценивается с помощью метрик, определенных на векторах. На основе методов машинного обучения строятся семантические модели, обучаемые по корпусам текстов с морфологической разметкой. Семантические модели могут строиться по разным корпусам обучающих текстов и посредством разных алгоритмов обучения. В онлайновом режиме использования сервис RusVectōrēs предоставляет выбор из пяти семантических моделей, в частности, доступна модель, построенная на основе Национального корпуса русского языка (НКРЯ).

Функция «Вычисление семантических ассоциатов» позволяет выявить для заданного входного слова десять семантически близких к нему слов. Каждое из этих слов находится с входным словом в определенном семантическом отношении. Однако сервис не дифференцирует выявляемые семантические отношения, устанавливаются только числовые характеристики, которые можно трактовать как оценки интенсивности семантической связи между двумя лексемами. Содержательная типология семантических отношений не строится. На данном этапе сервис RusVectōrēs позволяет задавать поиск семантически близких слов только среди существительных, глаголов, имен собственным, прилагательных, наречий.

3. Анализируемый материал и основные понятия

Для того чтобы выяснить, какие семантические отношения для существительных с большей вероятностью выявляются посредством применения инструментов дистрибутивной семантики, были рассмотрены два вопроса: 1) каковы типы выявляемых семантических отношений для существительных; 2) как соотносятся выявляемые семантические отношения с данными словаря синонимов.

Для решения первого вопроса сформировано множество семантических отношений, включающее отношения синонимии и антонимии, а так-

же базовые отношения существительных в электронных тезаурусах типа WordNet: гипонимия / гиперонимия (родовидовое отношение, род — гипероним, вид — гипоним, согипонимы — гипонимы с общим гиперонимом), меронимия / холонимия (отношение «часть — целое»). Кроме этого, добавлено еще четыре типа отношений: признак, операция, ситуационная связанность, словообразовательный вариант.

Для рассмотрения второго вопроса использовался «Словарь синонимов русского языка» [Словарь синонимов …, 1970]. Предпочтение было отдано профессионально составленному словарю синонимов, несмотря на то что по своему объему он значительно меньше интернет-ресурсов, позиционирующих себя как «Словари синонимов». Причина в том, что такие ресурсы обычно не содержат никакой информации о принципах их составления и не проводят различия между отношением синонимии и другими отношениями семантической близости. Например, онлайн-словарь русских синонимов, размещенный на ресурсе synonymonline.ru [CC], предлагает считать *свисток* синонимом слова *арбитр*, а *авторикша*, *полиграф*, *я*, *мы*, *песенник* — синонимами слова *автор*. Очевидно, что перечисленные слова не являются синонимами заданных лексем, а либо являются гипонимами (*песенник* по отношению к слову *автор*), либо связаны лишь тематически, либо не имеют даже тематической связи (*авторикша* по отношению к слову *автор*).

Поскольку [Словарь синонимов …, 1970] относительно невелик по количеству представленных в нем лексических единиц, для оценки возможности выявления синонимов на основе RusVectōrēs были взяты лексемы, имеющиеся в этом словаре синонимов.

Заранее было неясно, какие типы семантических отношений (кроме синонимии) выявляет рассматриваемый сервис, поэтому анализ проводился в два этапа. Сначала была взята сплошная выборка слов из фрагмента словаря синонимов, затем проявившиеся для этих слов закономерности проверялись дополнительно на других словах аналогичного типа.

В качестве первой выборки были взяты все опорные слова на букву «а» из «Словаря синонимов русского языка» [Словарь синонимов …, 1970] (сплошная выборка первых по порядку опорных слов-существительных), таких слов оказалось 29. Каждому из них сервис RusVectōrēs, поставил в соответствие по 10 семантически близких слов, таким образом, общее количество первоначально рассмотренных пар слов равно 290. В число этих 10 слов теоретически могли попасть существительные, прилагательные, глаголы, наречия, имена собственные, так как опция ограничения частей речи не использовалась. Дополнительная выборка состояла

из 35 слов, им были поставлены в соответствие 350 семантически близких слов.

Сервис RusVectōrēs базируется на автоматической морфологической разметке корпуса, поэтому определения частей речи и форм слов не всегда корректны. Это приводит к тому, что в число слов, семантически близких заданному, попадают словоизменительные варианты. Такие варианты из рассмотрения исключались. Были исключены также аномальные формы, которые затруднительно идентифицировать, например, *противоположенить*, *хозяинута*. Слова с незначительными легко распознаваемыми ошибками / опечатками и лексемы, представленные не начальной или нестандартной формой, не исключались (например, *шагомя* идентифицировано как гипоним слова *аллюр*, а *гротесок* — как согипоним слова *аллегория*).

После удаления всех словоизменительных вариантов и аномальных форм в первой выборке осталось 239 пар семантически близких слов, во второй — 281 пара. Суммарно в обеих выборках — 520 пар слов.

Использовались следующие критерии идентификации типов семантических отношений.

Для первой выборки синонимия устанавливалась по словарю синонимов, гипонимия / гиперонимия, меронимия / холонимия — по данным тезауруса RuWordNet (если нужные слова в нем присутствовали) или на основе критериев установления этих отношений, используемых в тезаурусах WordNet. Эти критерии таковы [Лукашевич, 2011, с. 63—64]:

— синсет X считается гипонимом синсета Y, если носитель языка считает нормальными предложения типа «X — это вид Y-а»;

— синсет X считается меронимом синсета Y, если можно сказать, что X есть часть Y-а.

Например, *метафора* является согипонимом по отношению к слову *аллегория*, *мелодия* — гиперонимом по отношению к слову *аккомпанемент*, *цирк* — холонимом по отношению к слову *арена*. Отношение согипонимии связывает два слова в том случае, если существует более общее понятие (гипероним), охватывающее оба значения. Считалось, что собственные имена (одноэлементные множества) также могут быть гипонимами, например, *Леонард Эйлер* — гипоним по отношению к слову *математик*.

Посредством RusVectōrēs словосочетания, как правило, не выявляются. Поэтому двухсловные гипонимы / гиперонимы обнаружены не будут, но могут быть выявлены частично (пословно). Например, в число слов, являющихся близкими по смыслу к слову *автомобиль*, попало прилагательное *легковой*. Понятно, что это часть словосочетания *легковой авто-*

мобиль, представляющего собой гипоним заданной лексемы *автомобиль*. Однако в отличие от однословных гипонимов того же слова (*иномарка*, *джип*, *форд* и др.), являющихся парадигматическими вариантами заданной лексемы, слово легковой характеризует синтагматику этой лексемы, а в роли парадигматического варианта может выступать только словосочетание *легковой автомобиль* как целое. Аналогично слово *непререкаемый* характеризует синтагматику лексемы *авторитет*, и в качестве гипонима (то есть парадигматического варианта) можно рассматривать только словосочетание *непререкаемый авторитет*. Необязательно в такой функции выступает прилагательное, например, в число семантически близких слов, выявленных для лексемы *аромат*, попало слово *тубероз* (форма слова *туберо́за* — название цветка), что, скорее всего, является частью словосочетания *аромат тубероз*. Форма *тубероз* характеризует синтагматику слова *аромат*, а словосочетание в целом является его гипонимом. Аналогично для слова *шофёр* было выявлено слово *газика* (*шофёр газика*). При идентификации типов семантических отношений таким синтагматическим характеристикам был приписан тип «признак».

К числу ситуационных семантических отношений слова X отнесены слова, которые могут встретиться в той ситуации, применительно к которой было использовано слово X. Например, слова *аплодисменты* и *шиканье* оказались семантически близкими, по-видимому, потому, что обозначенные ими действия могут происходить в одной и той же ситуации (в театре, на концерте, лекции и т. п.). По этой же причине семантически близким к слову *арена* оказалось слово *пикадор*, а аббревиатуры *fis* (Federation Internationale de Ski, FIS), *fig* (Federation Internationale de Gymnastique, FIG), *уефа* (от UEFA, Union of European Football Association) — тематически связанными с лексемой *арбитр* (могут быть использованы совместно в одной и той же ситуации). К этому же виду семантической связи отнесена лексическая функция конверсии [Мельчук, 1976, с. 78—83] (например, между лексемами *продавец* и *покупатель*).

В отдельный тип семантических отношений выделены операции: действия, которые могут быть выполнены с объектами, обозначаемыми входным словом. Например, в число семантически близких слов для лексемы *абзац* попал глагол *вычёркивание* (возможная операция, производимая с абзацем). Аналогичной семантической связью связаны слова *авантюра* и *втравлять*, *аромат* и *запахнуть*, *ветер* и *подуть*.

Таким образом, при идентификации семантической связи между входным словом и словом, выявленным посредством RusVectōrēs, рассматриваются следующие возможности: выявленное слово может быть по отноше-

нию к заданному слову синонимом, антонимом, гипонимом, гиперонимом, холонимом, меронимом, признаком, операцией, ситуационно связанным понятием, словообразовательным вариантом.

4. Анализ результатов

Числовые характеристики типов выявленных семантических отношений (табл. 1) свидетельствуют о том, что три наиболее часто обнаруживаемых типа отношений таковы: согипонимия, гипонимия, ситуационная связь. В каждой выборке эти три типа отношений составляют более 50 % от числа всех выявленных отношений.

Количество выявленных парадигматических отношений значительно превышает количество выявленных синтагматических отношений: в первой выборке 178 (74,5 %) и 61 (25,5 %) соответственно, во второй выборке 219 (77,94 %) и 62 (22,06 %). Эта закономерность заслуживает внимания. Дистрибутивная семантика основана на анализе частотных характеристик совместной встречаемости слов в текстах: чем чаще две лексемы встречаются в одном и том же тексте, тем больше вероятность того, что они семантически близки. Совместная встречаемость — это, вообще говоря, синтагматический признак, для выявления же парадигматических отношений на основе контекста обычно используется отношение взаимозаме-

Таблица 1

Типы семантических отношений

№	Первая выборка			Вторая выборка	
	Тип семантического отношения	Количество	%	Количество	%
1.	Антоним	1	0,4	2	1
2.	Гипероним	15	6,3	13	4,63
3.	Гипоним	49	20,5	104	37,01
4.	Мероним	6	2,5	2	1
5.	Операция	6	2,5	4	1,42
6.	Признак	25	10,5	23	8,19
7.	Синоним	20	8,4	17	6,05
8.	Ситуационная связь	30	12,6	35	12,46
9.	Словообразование	19	7,9	29	10,32
10.	Согипоним	62	25,9	51	18,15
11.	Холоним	6	2,5	1	0
	Итого	239		281	

щаемости [Гладкий, 1973]. Соответственно можно было бы ожидать, что метод выявления семантических отношений посредством дистрибутивной семантики позволит обнаружить значительную часть синтагматических отношений. Однако это оказалось не так. Причина видится в том, что в работах по теории замещаемости тексты рассматриваются как упорядоченные последовательности словоформ, в дистрибутивной семантике порядок не учитывается. Поэтому решающее значение, по-видимому, начинает оказывать не непосредственное окружение рассматриваемых слов в текстах (как в теории замещаемости), а общность тематики текстов.

Для слов первой выборки в [Словарь синонимов …, 1970] приведено 54 синонима. Из них в списке семантически близких слов, составленном посредством RusVectōrēs, оказалось только 20 (37 %) синонимов. Сравнение со [Словарь синонимов …, 1970] для второй выборки не проводилось, так как данный словарь сильно ограничивает выбор рассматриваемых лексем. Для второй выборки были привлечены сведения о синонимии из тезауруса RuWordNet.

При анализе данных первой выборки было замечено, что меньшее количество неинформативных результатов (аномальных форм и словоизменительных вариантов) получается для слов определенных семантических групп. Это названия сообществ людей, профессий, ролей, физических объектов / явлений / процессов. Вторая выборка была предназначена для того, чтобы получить данные об использовании RusVectōrēs для аналогичных слов.

Во вторую выборку были включены слова, относящиеся к следующим семантическим группам: 1) большие группы людей, объединяемых по роду профессиональных занятий (*художник*, *биолог*, *математик*, *лингвист*, *спортсмен*, *шофёр*, *строитель*); 2) внешние или внутренние качества человека (*хитрец*, *мудрец*, *скряга*, *брюнет*, *толстяк*, *мошенник*, *стиляга*); 3) роль (*автор*, *хозяин*, *продавец*, *лектор*, *путешественник*, *посетитель*, *пассажир*); 4) физические объекты и явления (*ветер*, *дождь*, *мороз*, *страна*, *гора*, *море*, *улица*); 5) процессы (*спор*, *соревнование*, *экзамен*, *путешествие*, *спектакль*, *эксперимент*, *ходьба*).

После устранения аномальных форм и словоизменительных вариантов выявленные семантически близкие слова образовали соответственно пять групп семантических ассоциатов (табл. 2). В таблице для каждой группы выделены цветом наиболее частотные семантические отношения. Можно заметить, что для всех групп чаще всего выявляются гипонимы, согипонимы, ситуационная связь. Для физических объектов и явлений этот вид семантических отношений далеко опережает остальные. Во всех груп-

пах крайне малочисленны антонимы, меронимы, холонимы, операции. Их доля была невелика в первой выборке, во второй еще уменьшилась. Ситуационные связи более всего проявились для третьей и пятой групп. Это было ожидаемо, так как слова из этих групп, как правило, используются для обозначения ситуаций с несколькими участниками или разнообразным физическим окружением / взаимодействием. Например, для слова *продавец* выявлено три ситуационно связанных с ним слова: *покупатель*, *товар*, *прилавок*, а для слова *море* — слова *предгорья*, *ущелье*, *долина*. Согипонимы оказались характерны для первых двух групп, здесь выявились названия сходных видов профессий, в частности, для слова *лингвист* выявлены согипонимы *этнолог*, *литературовед*, *антрополог*, *логик*, *социолог*, *культуролог*.

Таблица 2

Типы отношений для семантических групп слов

Семантическое отношение	Процент от общего количества слов в группе (%)				
	Группа 1	Группа 2	Группа 3	Группа 4	Группа 5
Антоним		4			
Гипероним	4,62	3,7	3,51	8	3,64
Гипоним	47,69	27,78	31,58	44	32,73
Мероним				4	
Операция	1,54			6	
Признак	10,77	3,7	5,26	10	10,91
Синоним	3,08	11,11	5,26	2	9,09
Ситуационная связь	3,08	1,85	24,56	10	23,64
Словообразование	4,62	14,81	17,54	8	7,27
Согипоним	24,62	33,33	10,53	8	12,73
Холоним			2		
Общее количество слов в группе	65	54	57	50	55

5. Выводы

1. Множества семантических отношений, выявленных посредством сервиса RusVectōrēs, обладают определенными закономерностями, характеризующими как всю совокупность рассмотренных слов в целом, так и особенности семантических ассоциатов разных семантических групп слов. Проделанная работа позволила выявить следующие закономерности.

2. Среди семантических отношений, выявляемых посредством RusVectōrēs, преобладают парадигматические: количество пар слов, связанных такими отношениями, оказалось примерно в три раза больше, чем

синтагматически связанных пар. Для анализа возможностей обсуждаемого метода было использовано 11 типов семантических отношений: 8 парадигматических и 3 синтагматических. Слова, соотносящиеся с заданным словом парадигматически, могут быть его гипонимами, гиперонимами, меронимами, холонимами, синонимами, антонимами, словообразовательными вариантами. Слова, синтагматически связанные с заданным словом, могут именовать признаки обозначаемых им объектов, возможные операции с этими объектами, ситуационно связанные с ними объекты.

3. Для определенных семантических групп слов синтагматические отношения также проявились заметным образом. Это характерно для слов, обозначающих ситуации или процессы со сложной структурой, разнообразными элементами окружения или взаимодействиями.

4. Наиболее часто посредством RusVectōrēs выявляются гипонимы или согипонимы. Это характерно для всех рассмотренных семантических групп слов.

5. Анализ результатов работы сервиса RusVectōrēs для опорных слов, выбранных из словаря синонимов русского языка [Словарь синонимов …, 1970], показал, что сервис выявляет только около одной трети синонимов, представленных в словаре.

Анализ особенностей работы RusVectōrēs для разных семантических групп слов позволит использовать данный сервис как вспомогательный инструмент для исследований в области лексической семантики. Для развития такой области применения представляется перспективным проверить и уточнить обнаруженные закономерности на более обширном материале, а также. провести аналогичный анализ типов выявляемых семантических отношений для глаголов и наречий.

Источники

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

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

3. СС — Словарь синонимов [Электронный ресурс]. — Режим доступа : http://synonymonline.ru/.

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

Литература

1. *Андреев Н. Д.* Статистико-комбинаторные методы в теоретическом и прикладном языковедении / Н. Д. Андреев. — Ленинград : Наука, 1967. — 403 с.

2. *Гладкий А. В.* Формальные грамматики и языки / А. В. Гладкий. — Москва : Наука, 1973. — 368 с.

3. *Глисон Г.* Введение в дескриптивную лингвистику / Г. Глисон. — Москва : Иностранная литература, 1959. — 487 с.

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

5. *Мельчук И. А.* Опыт теории лингвистических моделей «СМЫСЛ ⇔ТЕКСТ» / И. А. Мельчук. — Москва : Школа Языки русской культуры, 1999. — 346 с.

6. *Ревзин И. И.* Метод моделирования и типология славянских языков / И. И. Ревзин. — Москва : Наука, 1967. — 300 с.

7. *Сухотин Б. В*. Оптимизационные методы исследования языка / Б. В. Сухотин. — Москва : Наука, 1976. — 169 с.

8. *Belyy A. V.* Framework for Russian Plagiarism Detection Using Sentence Embedding Similarity and Negative Sampling / A. V. Belyy, M. A. Dubova // Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Международной конференции «Диалог». — Москва : Издательский центр РГГУ, 2018. — Выпуск 17 (24). — С. 96—109.

9. *Enikeeva E. V*. Russian collocation extraction based on word embeddings / E. V. Enikeeva, O. A. Mitrofanova // Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Международной конференции «Диалог». — Москва : Издательский центр РГГУ, 2017. — Вып. 16 (23). — Т. 1. — С. 52—64.

10. *Harris Z. S.* From phoneme to morpheme / Z. Harris // Language. — 1955. — Vol. 31. — № 2. — P. 190—222.

11. *Kutuzov A*. Two centuries in two thousand words: neural embedding models in detecting diachronic lexical changes / A. Kutuzov, E. Kuzmenko // Quantitative Approaches to the Russian Language. — New York : Routledge, 2017. — P. 95—112.

12. *Toldova S.* Coreference resolution for Russian: the impact of semantic features / S. Toldova, M. Ionov // Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Международной конференции «Диалог». — Москва : Издательский центр РГГУ, 2017. — Вып. 16 (23). — Т. 1. — С. 339—348.

Typology of Semantic Relations Extracted 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 question about the types of semantic relations extractable by the Web-service RusVectōrēs is considered. The research urgency is caused by insufficient state of knowledge about linguistic potential of this instrument and distributive semantics in whole. The novelty of the research is seen in the fact that the problem concerning typology of semantic relations extractable on the base of distributional semantics — that is the method that uses automatic processing of big textual data without taking into account linguistic information — is posed for the first time. Classification of semantic relations including eight types of paradigmatic and three types of syntagmatic relations is proposed. It is shown that application of the method under consideration provides the set of semantic associates that consists predominantly of the words that stand in paradigmatic relations with the entry word; and among them hyponyms or co-hyponyms are the most frequent. The question about an effect that the semantic group of the entry word may have on the types of extractable semantic relations is studied. It is shown that the effect can be traced, although it needs additional investigations on more representative data for further specification. The comparative analysis of five semantic groups of the words gave evidence for the abovementioned general trend (about predominance of hyponyms and co-hyponyms) and allowed to produce several specific hypothesis about correlations between the semantic type of an entry word and the types of semantic relations extracted for the word.

Key words: semantic relations; lexis; distributional semantics; automatic text processing; nouns; RusVectōrēs.

Material resources

SS — *Slovar' sinonimov*. Available at: http://synonymonline.ru/. (In Russ.).

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

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

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

References

Andreyev, N. D., (1967). *Statistiko-kombinatornyye metody v teoreticheskom i prikladnom yazykovedenii*. Leningrad: Nauka. (In Russ.).

Belyy, A. V., Dubova, M. A. (2018). Framework for Russian Plagiarism Detection Using Sentence Embedding Similarity and Negative Sampling. In: *Kompyuternaya lingvistika i intellektualnyye tekhnologii. Po materialam ezhegodnoy Mezhdunarodnoy konferentsii «Dialog».* Moskva: Izdatelskiy tsentr RGGU. *17 (24)*: 96—109. (In Russ.).

Enikeyeva, E. V., Mitrofanova, O. A. (2017). Russian collocation extraction based on word embeddings. In: *Kompyuternaya lingvistika i intellektualnye tekhnologii. Po materialam ezhegodnoy Mezhdunarodnoy konferentsii «Dialog».* Moskva: Izdatelskiy tsentr RGGU. *16 (23)/1*: 52—64. (In Russ.).

Gladkiy, A. V. (1973). *Formalnye grammatiki i yazyki*. Moskva: Nauka. (In Russ.).

Gleason, H. (1959). *Vvedeniye v deskriptivnuyu lingvistiku*. Moskva: Inostrannaya literatura. (In Russ.).

Harris, Z. S. (1955). From phoneme to morpheme. *Language*, *31 (2)*: 190—222.

Kutuzov, A., Kuzmenko, E. (2017). Two centuries in two thousand words: neural embedding models in detecting diachronic lexical changes. In: *Quantitative Approaches to the Russian Language.* New York: Routledge. 95—112.

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

Melchuk, I. A. (1999). *Opyt teorii lingvisticheskikh modeley «SMYSL — TEKST»*. Moskva: Shkola Yazyki russkoy kultury. (In Russ.).

Revzin, I. I. (1967). *Metod modelirovaniya i tipologiya slavyanskikh yazykov*. Moskva: Nauka. (In Russ.).

Sukhotin, B. V. (1976). *Optimizatsionnyye metody issledovaniya yazyka*. Moskva: Nauka. (In Russ.).

Toldova, S., Ionov, M. (2017). Coreference resolution for Russian: the impact of semantic features. In: *Kompyuternaya lingvistika i intellektualnye tekhnologii. Po materialam ezhegodnoy Mezhdunarodnoy konferentsii «Dialog»*. Moskva: Izdatelskiy tsentr RGGU. *16 (23)/1*: 339—348. (In Russ.).