УДК 81`33+373 https://www.doi.org/10.33910/2687-0215-2020-2-2-118-146

Изучение терминологических подсистем современных школьных учебников на русском языке с помощью модели анализа семантики естественных языков Word2Vec

С. И. Монахов[✉1], В. В. Турчаненко[1, 2], Е. А. Федюкова[3], Д. Н. Чердаков[1, 4]

[1] Российский государственный педагогический университет им. А. И. Герцена,
191186, Россия, г. Санкт-Петербург, наб. реки Мойки, д. 48
[2] Институт русской литературы (Пушкинский дом) РАН,
199034, Россия, г. Санкт-Петербург, наб. Макарова, д. 4
[3] Независимый исследователь, Россия, г. Санкт-Петербург
[4] Санкт-Петербургский государственный университет,
199034, Россия, г. Санкт-Петербург, Университетская наб., д. 7/9

Terminological subsystems of modern Russian school textbooks: A study based on Word2Vec and neural networks

S. I. Monakhov[✉1], V. V. Turchanenko[1, 2], E. A. Fedyukova[3], D. N. Cherdakov[1, 4]

[1] Herzen State Pedagogical University of Russia, 48 Moika Emb., Saint Petersburg 191186, Russia
[2] Institute of Russian Literature (Pushkinskij Dom), Russian Academy of Science,
4 Makarova Emb., Saint Petersburg 199034, Russia
[3] Independent researcher, Saint Petersburg, Russia
[4] Saint-Petersburg State University, 7/9 Universitetskaya Emb., Saint-Petersburg 199034, Russia

Аннотация. Цель исследования, первые результаты которого представлены в настоящей статье, — анализ состава и особенностей функционирования терминологической лексики в учебниках для средней школы Российской Федерации с помощью методов и средств компьютерной лингвистики. Количество терминов из разных областей знания, которое школьник должен усвоить за время обучения в средней школе, никогда не подвергалось оценке. По предварительным подсчетам, произведенным на материале Примерной основной образовательной программы общего и среднего образования 2015 года только в части предмета «Русский язык», ученик в 5–11 классах средней школы должен понимать, распознавать и уметь употреблять около 1000 терминов и терминологических сочетаний из этой сферы знания. Таким образом, учитывая количество школьных дисциплин, общее число единиц специальной лексики, изучаемых в общеобразовательной школе, измеряется тысячами. В то же время сопоставительные характеристики состава и функционирования терминов в учебниках для разных школьных предметов не изучены и остаются неизвестными. Неясна корреляция между

Abstract. The article reports the results of the study that explored the inventory and functioning of scientific terms and special lexemes in textbooks for Russia's secondary schools. The toolset included modern methods of natural language processing and deep learning. The number of terms from different fields of knowledge that a secondary school student should learn has never been evaluated. According to the preliminary evaluations based on the Model Basic Curriculum for General and Secondary Education 2015, a secondary school leaver is supposed to be able to understand, recognise and use about 1,000 terms and terminological combinations in the subject Russian Language alone. Thus, taking into account the number of school subjects, the total number of special vocabulary studied in general education schools is measured in thousands. At the same time, the comparative characteristics of the inventory and functioning of terms in textbooks for different school subjects are under-scrutinized and remain unknown. Besides, it is unclear how the terminological density of school textbooks for different subjects correlates with the place occupied by these subjects in the curriculum.

терминологической плотностью учебного текста в школьных учебниках по разным предметам и местом, занимаемым этими предметами в учебных планах. Традиционным способом вычленения терминов из специальных текстов является их просмотр и «ручное» формирование соответствующих перечней. При надежности такого способа в отношении интеллектуализации принципов отбора он плохо приложим к большим массивам данных и не отражает ни частотность употребления терминов, ни специфику их синтагматических связей, ни системные отношения между терминами, формируемые их сочетаемостным поведением. Реализация описываемого проекта предусматривает создание полнотекстового корпуса на материале текстов школьных учебников 5–11 классов, включенных в Федеральный перечень Министерства просвещения РФ, автоматическое вычленение и стратификацию терминов при помощи методов дистрибутивной семантики, создание и обучение глубокой нейросети, способной по поданной на вход группе векторных представлений терминов определить учебную дисциплину, уровень обучения и учебную тему. Результаты исследования могут представлять теоретический интерес в перспективе развития терминоведения и иметь практическое применение при создании школьной учебной литературы разных типов.

Ключевые слова: термин, терминология, векторное представление, учебник, общее образование, русский язык, коллокация, нейросеть, глубокое обучение, модель, Word2Vec, CBOW, skip-gram, ngram.

The traditional way of compiling lists of special terms is simply to glean them from special texts and write down manually. This method is reliable to gain insights into the best selection practices, however, it cannot be applied to large data sets and does not reflect the term frequency, the specificity of their syntagmatic connections, or the systemic relationship between them. Our project is aimed at filling this gap through: 1) creating a full-text corpus of school textbooks approved by the Ministry of Education for grades 5–11, 2) automatic extraction, stratification, and mapping of terms with the help of distribution semantics algorithms, 3) creation and training of a deep neural network capable of predicting the subject, level of education and educational topic given a group of vector theoretical development of terminology science. They may also find practical application, e. g., in the development of different types of educational literature.

Keywords: term, terminology, vector representation, school textbook, general education, Russian language, collocations, neural network, deep learning, Word2Vec, CBOW, skip-gram.

Вступление

Изучение терминологической лексики сопряжено с решением многих теоретических и прикладных вопросов. В числе наиболее важных проблем можно выделить следующие: уяснение специфики лексической семантики терминологических единиц, их структурных и семантических типов, их отграничения от терминоподобных единиц, их употребления в специальных и неспециальных текстах, выработка методов их неавтоматического и автоматического вычленения из текстов, формирование принципов составления терминологических словарей разных видов.

Учебная литература является одной из важных сфер бытования терминологической лексики. При этом школьная учебная литература, несмотря на обширный опыт отечественного терминоведения, исследована в аспекте терминоупотребления слабо. Между тем эта область терминологического поля русского языка интересна в собственно научном аспекте и, без сомнения, значима в социально-педагогическом отношении.

Терминологический состав школьных дисциплин формируется из разных источников: предметной терминологии Федерального государственного образовательного стандарта (в проекте нового ФГОС эта терминология включена в состав разделов «Требования к результатам освоения основной образовательной программы» и «Требования к предметным результатам освоения учебных предметов, выносимым на промежуточную и итоговую аттестацию»), предметной терминологии основных образовательных программ, предметной терминологии рабочих программ, терминологии собственно учебника, терминологии компонентов учебно-методических комплектов (рабочих тетрадей, контрольных работ, тестов и др.).

Количество терминов из разных областей знания, которое школьник должен усвоить за время обучения в средней школе, насколько нам известно, никогда не подвергалось оценке. Впрочем, можно полагать, что оно весьма велико. Так, по предварительным подсчетам, произведенным на материале Примерной основной образовательной программы общего и среднего образования 2015 года, только в части предмета «Русский язык» ученик в 5–11 классах средней школы должен понимать, распознавать и в идеале уметь употреблять около 1000 терминов и терминологических сочетаний из этой сферы знания. Таким образом, учитывая количество школьных дисциплин, общее число единиц специальной лексики, изучаемых в общеобразовательной школе, измеряется тысячами.

В то же время сопоставительные характеристики состава и функционирования терминов в учебниках для разных школьных предметов не изучены и остаются неизвестными. Неясна корреляция между терминологической плотностью учебного текста в школьных учебниках по разным предметам и местом, занимаемым этими предметами в учебных планах (имея в виду и объем отводимого на курс учебного времени, и значимость курса с точки зрения его положения в традиционной иерархии в системе знания).

Плохо изучена специфика школьной терминологии в ее сопоставлении с собственно научной, особенно современной. (Ср., например, используемое в некоторых учебниках по русскому языку обозначение «слова́ состояния» при научном «категория состояния», «слова́ категории состояния»; предложенное в проекте нового Федерального государственного образовательного стандарта сочетание «разговорный язык» при преимущественном употреблении в науке «разговорная речь» и др.).

Школьная терминология находит свое отражение в школьных терминологических (энциклопедических) словарях. Однако они не охватывают всей совокупности терминов, используемых в школьной учебной литературе, и не дают представления о реальной картине употребления терминов в школьных учебных текстах.

Традиционным способом вычленения терминов из специальных текстов является их просмотр и «ручное» формирование соответствующих перечней. При надежности такого способа в отношении интеллектуализации принципов отбора он плохо приложим к большим массивам данных и не выявляет ни частотность употребления терминов, ни специфику их синтагматических связей, ни системные отношения между терминами, формируемые их сочетаемостным поведением. В связи с этим перспективным представляется автоматизированное вычленение терминов из созданного корпуса текстов.

Извлечение терминологии из текстового корпуса (terminology extraction) — одна из наиболее актуальных сегодня задач машинной обработки естественного языка, важность решения которой была осознана еще в 1990-х годах, когда в лингвистике оформилось представление о терминологической лексике как об активной части словарного состава языка (Караулов 1991; Ментруп 1983; Шелов 1998). Принципиальная проблема, возникающая на пути к решению этой задачи, связана с лингвистической природой самой терминологической лексики — размытостью внутренних границ между различными группами терминов (зачастую один термин может относиться сразу к нескольким областям человеческого знания), а также между терминами и общеупотребительной лексикой. Согласно В. М. Лейчику, «граница между терминологической и общеупотребительной лексикой нестабильна <...> постоянно происходит как процесс превращения терминов в общеупотребительные слова, так и использование бытовой лексики для формирования терминологий, когда на основе представлений формируются понятия» (Лейчик 2007, 26).

Подавляющее большинство современных систем автоматического извлечения терминологии используют статистический подход — это позволяет сделать их работу независимой от языка исследуемого корпуса. Такие системы работают с двумя корпусами:

целевым, из которого необходимо извлечь терминологические единицы, и референционным — чаще всего национальным корпусом соответствующего языка, представляющим языковую систему в целом. Для слов в целевом корпусе подсчитывается частота встречаемости, которая затем сравнивается с частотой встречаемости этого же слова в референционном корпусе. В результате сравнения частот употребления система присваивает единицам, отличающимся неожиданно высокой частотностью, статистическую меру ключевого слова (keyness score). Те слова, у которых keyness score оказывается выше определенного порога, записываются в кандидаты на присвоение им терминологического статуса (Kilgarriff et al. 2014).

У этого подхода есть два основных недостатка: 1) неспособность извлекать низкочастотные термины (Cabré et al. 2001); 2) итоговым результатом обработки является список терминов, упорядоченный по частоте встречаемости, но не дающий никакого представления о терминологической системе, частью которой эти термины являются, то есть не эксплицирующий системные связи между этим терминами, не оценивающий степень их удаленности друг от друга.

Если с первым недостатком можно бороться, во-первых, с помощью оценки результатов компьютерной модели по критериям точности (способность системы отличать термины от нетерминов, которая рассчитывается как отношение количества извлеченных терминов к количеству извлеченных кандидатов в термины) и полноты (способность системы извлекать все термины из корпуса, которая рассчитывается как отношение количества извлеченных кандидатов в термины к общему количества терминов в корпусе), а во-вторых, с помощью использования гибридной (то есть сочетающей количественный подход с использованием набора неких специальных правил, чаще всего морфологической разметки корпуса), а не чисто статистической системы автоматического извлечения терминологии, то решение второй проблемы представляет собой в настоящее время нерешенную фундаментальную научную задачу.

В начале XXI века в изучении лексической семантики, и в том числе различных терминологических подсистем лексики, наметился новый поворот, подготовленный в первую очередь новейшими достижениями в сфере машинной обработки естественного языка. Все большее количество научных моделей лексической семантики, при всем разнообразии их технических деталей, исходят из одной общей предпосылки: окружение слова, его контекст несет в себе важнейшую информацию о значении самого этого слова. Иначе говоря, даже не зная ничего о семантике конкретных лексем, мы можем группировать их в любом большом корпусе по степени смысловой близости, учитывая бóльшую или меньшую вероятность их появления на определенном удалении друг от друга (см., например: Durda, Buchananan 2008; Jones, Mewhort 2007; Rohde, Gonnerman, Plaut 2006).

С момента разработки набора компьютерных алгоритмов Word2Vec (Mikolov, Chen, Corrado, Dean 2013; Mikolov, Sutskever, Chen et al. 2013) — continuous-bag-of-words (CBOW) и skip-gram, — векторные представления слов стали широко использоваться при обучении нейросетей тому, чтобы либо предсказывать искомое слово по данному контекстному окружению (CBOW), либо, напротив, предсказывать контекстное окружение по данному слову (skip-gram). В процессе обучения нейросеть получает входной one-hot вектор и пытается предсказать результат в виде распределения вероятностей каждого слова в словаре оказаться в контекстном окне входного слова. Основы процесса могут быть описаны следующим образом. Сперва для введенного (целевого) слова в embedding layer находится соответствующий вектор. Этот вектор подается на вход нейросети, которая пытается предсказать правильное выходное (контекстное) слово. После сравнения предсказанного слова

и того слова, которое на самом деле находится в контекстном окне, вычисляется функция потерь, которая вместе со стохастическим градиентным спуском используется для оптимизации нейросети и embedding layer (Mikolov, Yih, Zweig 2013).

Алгоритмы Word2Vec, запатентованные Google, доказали свою высокую эффективность, совершив значительный прорыв в области семантики, использующей для анализа векторные представления слов. CBOW и skip-gram в ходе тестирования продемонстрировали существенно более высокие показатели корректной семантической и синтаксической интерпретации текста, чем многие альтернативные модели (Mikolov, Sutskever, Chen et al. 2013).

Одной из характеристик, которая делает алгоритмы Word2Vec уникальными, является использование плотных (dense) векторов. Если в большинстве контекстных лексических моделей слова представлены one-hot векторами (то есть векторами, длина которых равняется длине словаря, при этом все позиции в них, кроме одной, заполнены нулями), то алгоритм skip-gram использует более короткие векторные представления, обычно длиной в несколько сотен информационных ячеек, кодирующих вероятности прогнозируемых контекстов. Это говорит о том, что векторы Word2Vec представляют семантическое пространство корпуса текстов более прозрачно, выбирая из общего множества только те отношения, которые имеют непосредственное значение для предсказания семантических различий между словами и их контекстами.

Было доказано, что модели, обученные по алгоритмам Word2Vec, дают на удивление хорошие результаты при сложении и вычитании векторных представлений даже семантически неблизких слов. Хрестоматийным стал следующий пример: [вектор для английского слова king ‘король’] – [вектор для английского слова man ‘мужчина’] + [вектор для английского слова woman ‘женщина’] = [вектор для английского слова queen ‘королева’]. И это далеко не исключительный случай, а скорее прямое следствие того, что алгоритмы Word2Vec позволяют улавливать закодированные в тексте понятийные связи, представляющие собой проекцию связей онтологических (Levy, Goldberg 2014). Не менее важно то, что, хотя сами модели Word2Vec представляют собой мелкие нейронные сети (shallow network), состоящие всего из двух скрытых слоев нейронов, созданные ими векторные представления слов могут использоваться в качестве матрицы весов входных сигналов при создании и обучении глубоких нейросетей (Brownlee 2017).

Стоит отметить, что, хотя модели Word2Vec являются сейчас чрезвычайно актуальным и востребованным способом семантического анализа естественного языка, они, насколько нам известно, до сих пор не использовались в качестве инструмента извлечения терминологии и моделирования терминологических подсистем, тем более применительно к учебной литературе. Между тем применительно к проблеме состава и функционирования терминов в школьном учебном тексте указанные методы предоставляют эвристические возможности, недоступные традиционным описательным процедурам. В частности: 1) появляется возможность проследить за частотностью того или иного термина в разных типах школьного учебного текста и оценить их терминологическую плотность; 2) становится возможным сопоставление прототипического и ближайшего для термина семантического окружения — его дефиниции — и разновидностей его семантического окружения в многообразии контекстов употребления; 3) становится возможно охарактеризовать текстуально реализуемые синонимические и антонимические связи терминов, а также терминов и слов нетерминологического характера, используемых в школьных учебниках; 4) выявляется терминологическое и нетерминологическое употребление одного и того же слова внутри одного учебника и учебников, относящихся к разным школьным предметам; 5) формируется база для сопоставительно-типологической характеристики употребления различных терминов

внутри одной дисциплины и в разных дисциплинах школьного обучения. Кроме того, становится возможным сопоставление поведения одних и тех же терминов в школьных учебниках, в собственно научных текстах и неспециальных текстах.

Указанные явления могут быть охарактеризованы как синхронно-описательно в рамках отдельно взятого учебника, так и сопоставительно — при сравнении учебников по разным школьным предметам. Отдельным аспектом исследования является характеристика данных явлений в динамике — при последовательном анализе учебников разных классов, от 5-го к 11-му.

Далее в настоящей статье описываются основные результаты работы коллектива авторов над: 1) созданием корпуса школьных учебников на русском языке по основным учебным дисциплинам в соответствии с федеральным перечнем, утвержденным приказом Министерства просвещения РФ № 345 от 28 декабря 2018 года (с изменениями, утвержденными приказами Министерства просвещения РФ № 233 от 8 мая 2019 года, № 632 от 22 ноября 2019 года, № 249 от 18 мая 2020 года); 2) составлением списков автоматически извлеченных из корпуса школьных учебников терминов и специальной лексики по областям знания, соответствующим рассматриваемым учебным дисциплинам (здесь и ниже сочетание «специальная лексика» используется в значении ‘лексика, не являющаяся собственно терминологической, но близкая к ней по частотным характеристикам’; подробнее см. далее); 3) созданием комплекта дистрибутивно-семантических моделей Word2Vec, содержащих векторные представления извлеченных из корпуса школьных учебников терминов и специальной лексики по областям знания, соответствующим рассматриваемым учебным дисциплинам.

1. Подготовка корпуса школьных учебников

Создание исследовательского корпуса школьных учебников на русском языке в качестве основы для всей последующей работы было первоочередной и ключевой задачей. Научным коллективом был предпринят анализ федерального перечня учебников, утвержденного приказом Министерства просвещения РФ № 345 от 28 декабря 2018 года (с учетом изменений, утвержденных приказами Министерства просвещения РФ № 233 от 8 мая 2019 года, № 632 от 22 ноября 2019 года, № 249 от 18 мая 2020 года). По итогам анализа был составлен репрезентативный список учебников (21 дисциплина, 212 позиций), рекомендуемых для включения в исследовательский корпус (см. Приложение). Научным коллективом получено согласие АО «Издательство “Просвещение”» на использование учебных изданий в исследовательских целях (письмо № 1300/20 от 24 августа 2020 года).

Членами научного коллектива было осуществлено сканирование и последующее распознавание текстов 212 учебников. Распознанные тексты прошли предварительную обработку (препроцессинг), включающую удаление небуквенных символов и знаков препинания, приведение букв к нижнему регистру и проч. С использованием программных средств были осуществлены лемматизация (приведение слов к начальной форме — лемме) и POS-тегирование (частеречная разметка) предварительно обработанных текстов.

Приведем список учебных дисциплин с указанием количества учебников: алгебра — 18 учебников; астрономия — 2 учебника; биология — 21 учебник; всеобщая история и история России — 17 учебников; география — 8 учебников; геометрия — 8 учебников; естествознание — 2 учебника; изобразительное искусство — 8 учебников; информатика — 6 учебников; литература — 36 учебников; математика — 10 учебников; математический анализ — 14 учебников; мировая художественная культура — 2 учебника; музыка — 4 учебника; обществознание — 12 учебников; право — 2 учебника; русский язык — 24 учебника; технология — 4 учебника; физика — 15 учебников; физическая культура — 7 учебников; химия — 13 учебников.

Подготовленный в соответствии с федеральным перечнем корпус школьных учебников на русском языке был поэтапно загружен на платформу Sketch Engine (https://www.sketchengine.eu). Выбор программного обеспечения обусловлен возможностью параллельной работы нескольких исследователей и широким набором инструментов, доступных исследователю после создания корпуса. Внутри корпуса учебников были сформированы подкорпусы по основным учебным дисциплинам; каждый предметный подкорпус был также разделен внутри по годам обучения. В связи с требованиями правообладателя корпус закрыт и доступен для работы только исследовательской группе.

Общий объем корпуса — около 14 370 000 слов, из них по подкорпусам: алгебра — 1 144 089 слов (код Algebra), астрономия — 89 574 слова (код Astronomy), биология — 1 125 648 слов (код Biology), всеобщая история и история России — 973 498 слов (код History), география — 512 173 слова (код Geography), геометрия — 370 054 слова (код Geometry), естествознание — 158 665 слов (код NatSci), изобразительное искусство — 283 608 слов (код Art), информатика — 284 683 слова (код Informatics), литература — 3 939 054 слова (код Literature), математика — 525 035 слов (код Maths), математический анализ — 1 134 786 слов (код Matan), мировая художественная культура — 33 130 слов (код ArtisticCulture), музыка — 76 241 слово (код Music), обществознание — 505 822 слова (код Sociology), право — 171 349 слов (код Law), русский язык — 1 131 575 слов (код Ruslang), технология — 163 574 слова (код Crafts), физика — 1 098 625 слов (код Physics), физическая культура — 301 371 слово (код PhysicalEd), химия — 543 283 слова (код Chemistry).

2.1. Автоматическое извлечение терминологии: метрика keyness score

Следующей исследовательской задачей на данном этапе было составление списков терминов и специальной лексики по областям знания, соответствующим рассматриваемым учебным дисциплинам.

Из подкорпусов созданного корпуса школьных учебников на русском языке были автоматически извлечены кандидаты в термины соответствующих областей знания и уровней обучения посредством сравнения относительной частоты лексем целевого корпуса учебников с относительной частотой аналогичных лексем референционного корпуса, в качестве которого выступил Russian Web 2011 Sample (ruTenTen11), основанный на текстах русскоязычного сегмента сети Интернет — он содержит более 900 миллионов слов и доступен в Sketch Engine для использования в готовом виде. Важно отметить, что для кандидатов — однословных единиц (keywords в нотации Sketch Engine) и кандидатов — неоднословных сочетаний (terms в нотации Sketch Engine) технология выделения несколько различалась.

В случае с однословными единицами процедура носила следующий характер: для каждой лексемы, встречающейся в соответствующем подкорпусе не менее трех раз, высчитывалась метрика keyness score по следующей формуле:

$$((Lt \times 1\,000\,000 / Ct) + 1) / ((Lr \times 1\,000\,000 / Cr) + 1),$$

где Lt — частота употребления лексемы в целевом корпусе, Ct — общее количество токенов в целевом корпусе, Lr — частота употребления лексемы в референционном корпусе, Cr — общее количество токенов в референционном корпусе.

В том случае если значение метрики keyness score превышало 1, данная лексема включалась в список терминологических кандидатов. Так, для термина «многочлен», встречающегося в целевом подкорпусе по учебной дисциплине «Алгебра» (7–9 классы), было

получено следующее значение метрики keyness score: 1003,785 + 1 / 0,352 + 1 = 743,18, что свидетельствует об очень высоком статусе данного терминологического кандидата.

Алгоритм выделения неоднословных терминологических кандидатов включал два этапа. На первом этапе из всех возможных сочетаний лексем, встречающихся в соответствующем подкорпусе не менее трех раз, были выделены коллокации — сочетания, характеризующиеся положительным значением метрики Log-Dice score, рассчитываемой по следующей формуле:

$$14 + \log(2(|X \cap Y|) / (|X| + |Y|)),$$

где |X| — абсолютная частота первого элемента сочетания в подкорпусе, |Y| — абсолютная частота второго элемента сочетания в подкорпусе, |X∩Y| — абсолютная частота всего сочетания в подкорпусе.

На втором этапе для выделенных коллокаций была рассчитана метрика keyness score — по той же формуле, что и в случае с однословными терминологическими кандидатами. Так, для терминологического сочетания «график функции», встречающегося в целевом подкорпусе по учебной дисциплине «Алгебра» (7–9 классы), было получено следующее значение метрики keyness score: 832,366 + 1 / 0,148 + 1 = 725,930, что свидетельствует об очень высоком статусе данного терминологического кандидата.

Оба типа получившихся списков кандидатов — однословных и неоднословных — были отсортированы по убыванию значения метрики keyness score; первые 1000 позиций в обоих случаях были сохранены для дальнейшей обработки.

2.2. Автоматическое извлечение терминологии: word embedding models

Основная проблема описанного выше подхода, который является традиционным для современного корпусного терминоведения, заключается в том, что в списки терминологических кандидатов проникает не только терминологическая, но и просто нечастотная лексика, по каким-либо причинам широко представленная в целевом корпусе. Так, для подкорпуса учебников русского языка характерно активное использование языковых примеров из произведений русской классической литературы, насыщенных архаизмами, а также большого количества пейзажных зарисовок. В результате в списке терминологических кандидатов для данной учебной дисциплины оказались такие слова, как «роща», «кафтан», «вьюга», «туча», «шалаш», «ландыш», «журавль», «груша», «сумрак», «огурец», «гумно», «овраг» и многие другие.

В рамках подхода к автоматическому извлечению терминов, основанного на сравнении относительной частоты встречаемости леммы в целевом и референционном корпусе, невозможно установить надежные различия между словами «груша» и «суффикс»: оба будут характеризоваться высоким значением метрики keyness score. С целью усовершенствования этих результатов была проведена векторизация корпуса, после чего для каждой из рассматриваемых областей знания (для каждой из представленных учебных дисциплин) были созданы и обучены дистрибутивно-семантические модели (word embedding models), позволяющие выявить относительную семантическую близость единиц изучаемых терминологических подсистем.

Для каждой дисциплины было получено две модели: одна для одиночных лексем, другая для двух- и трехсловных сочетаний (биграмм и триграмм). При обучении моделей использовался набор алгоритмов Word2Vec, предполагающий следующую организацию процесса работы: 1) рассчитывается встречаемость каждого слова в корпусе; 2) массив слов сортируется по частоте, редкие слова удаляются; 3) для кодирования словаря строится

дерево Хаффмана (Huffman Binary Tree), что значительно снижает вычислительную сложность алгоритма; 4) с учетом заданного параметра окна́ контекстов (максимальная дистанция между текущим и предсказываемым словом в предложении) для каждого слова в корпусе строится вектор, элементы которого представляют собой обозначения количества случаев, когда данное слово оказывается в одном окне с другими наиболее частотными словами данного корпуса; 5) получившиеся векторы подаются на вход нейросети прямого распространения (feedforward neural network), которая обучается предсказывать либо контекст по заданному слову, либо слово по заданному контексту.

Благодаря векторному представлению становится возможным оценивать степень семантической близости каждой пары слов как косинусной меры их векторов. Эта мера может принимать значения в промежутке [0, 1]. Значение 1 свидетельствует о том, что векторы слов ортогональны, перпендикулярны друг другу, а значит, у этих слов нет похожих контекстов и общих сем. Значение 0, напротив, свидетельствует о практически полной идентичности контекстов и, следовательно, о почти тождественной семантике слов.

Так, для векторов упомянутых выше слов «груша» и «суффикс» косинусная мера составляет 0,82, а для векторов слов «приставка» и «суффикс» — 0,17, что свидетельствует о редкой контекстной встречаемости первой пары и частой встречаемости второй. Показательно сравнение двух групп наиболее близких в векторном пространстве слов и словосочетаний из целевого подкорпуса по учебной дисциплине «Русский язык»:

— группа слова «суффикс»: [(‘приставка’, 0,17), (‘производный_слово’, 0,18), (‘исходный_слово’, 0,18), (‘окончание’, 0,20), (‘корень’, 0,22), (‘основа_слово’, 0,22), (‘суффиксальный’, 0,22), (‘образовывать_слово’, 0,22)],

— группа слова «груша»: [(‘ваза’, 0,04), (‘сыпаться’, 0,04), (‘тощий’, 0,04), (‘уголь’, 0,04), (‘свернуться’, 0,04), (‘сизый’, 0,04), (‘великан’, 0,04), (‘скашивать’, 0,04), (‘луковица’, 0,05)].

В первом случае очевидна высокая степень тематической близости группы, во втором — ее разнородность и случайность. Кроме того, для слов группы «суффикс» характерна хотя и высокая, но не экстремально высокая мера косинусной близости: это свидетельствует о том, что они часто встречаются вместе, но могут употребляться и независимо друг от друга в разных контекстах. Что касается слов группы «груша», аномально высокие показатели косинусной близости говорят о том, что эти слова в целевом подкорпусе имеют ограниченное употребление и встречаются лишь в нескольких контекстах.

2.3. Автоматическое извлечение терминологии: t-SNE и кластерный анализ

Различия в векторных представлениях терминологической и нетерминологической лексики позволили нам спроектировать алгоритм отсеивания некорректных терминологических кандидатов и усовершенствовать таким образом результаты автоматического извлечения терминов, полученные благодаря сравнению относительных частот слов целевого и референционного корпусов. Это было сделано в несколько этапов.

Построение синхронных (по дисциплинам) и диахронных (по уровням обучения) карт взаимного расположения терминологических кандидатов в полученных дистрибутивно-семантических моделях и проекция этих карт из векторного пространства высокой размерности в двухмерную плоскость с целью визуализации полученных результатов были осуществлены с использованием алгоритма стохастического вложения соседей с t-распределением (t-SNE), предназначенного для вложения данных высокой размерности в двух- или трехмерное пространство.

Из множества полученных для каждой учебной дисциплины точек на плоскости были аннотированы только представления слов и словосочетаний, входящих в списки первоначальных терминологических кандидатов с высоким значением метрики keyness score (см. схему 1).

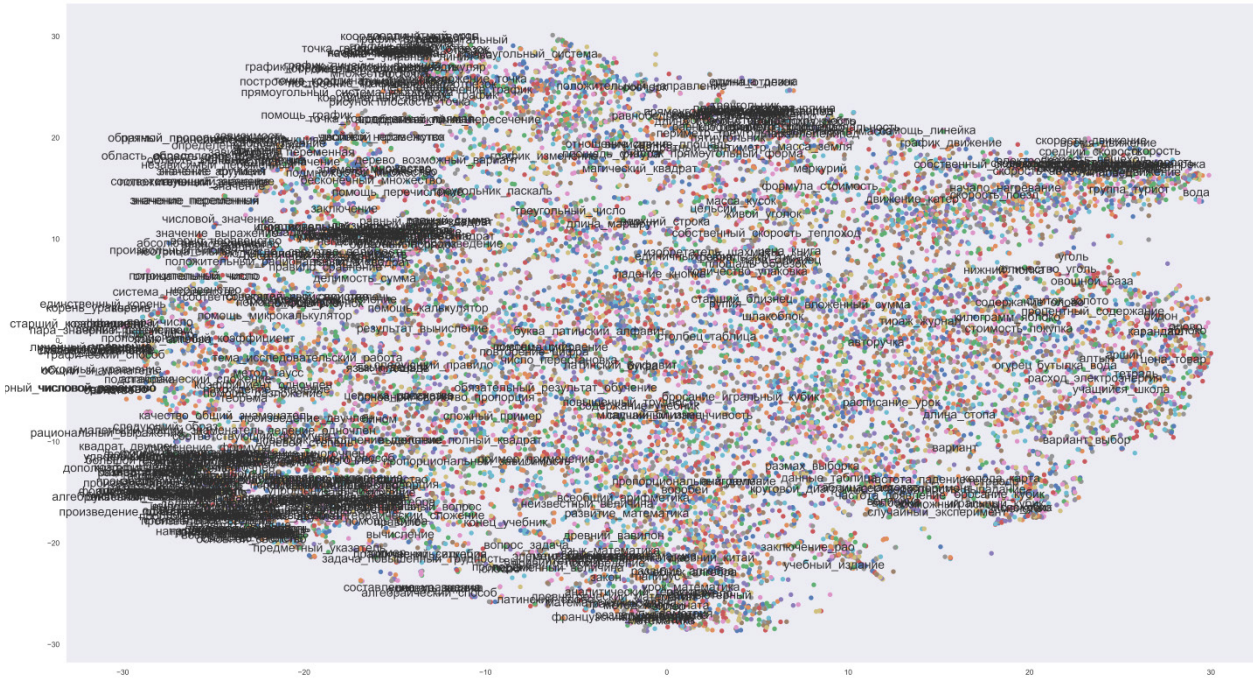

Схема 1. Векторное пространство терминологических кандидатов подкорпуса «Алгебра. 7 класс»

Образующие на схеме черноты скопления терминологических кандидатов, согласно нашей теории, с высокой долей вероятности являются терминами, поскольку группируются в кластеры, объединенные небольшой косинусной дистанцией. Что касается терминологических кандидатов, беспорядочно разбросанных на остальном пространстве карты, то они, скорее всего, являются лжетерминами, так как не входят в состав ни одной из терминологических подсистем этого корпуса.

После составления таких карт по всем областям знания и всем уровням обучения мы высчитали для каждой из расположенных на них точек две координаты: по горизонтальной и вертикальной осям. Затем с помощью метода вычисления k-средних (k-means) была осуществлена кластеризация точек на плоскости по их координатам; для каждой учебной дисциплины было условно задано 20 кластеров, по которым и распределились все имеющиеся точки.

Наконец, каждый кластер был маркирован или как содержащий терминологическую лексику, или как не содержащий терминологическую лексику с учетом следующих факторов: 1) удельная доля лексем, встречающихся внутри кластера и в качестве отдельных единиц, и в составе словосочетаний (гипотеза заключалась в том, что для терминологических кластеров характерна более высокая степень повторяемости, чем для нетерминологических кластеров); 2) удельная доля неоднословных сочетаний внутри кластера (гипотеза заключалась в том, что в терминологических кластерах количество неоднословных единиц больше, поскольку автоматическое выделение терминологических сочетаний характеризуется более высоким уровнем точности, чем выделение отдельных терминов); 3) удельная доля терминологических кандидатов внутри кластера, находящих соответствие в списках терминов, предусмотренных стандартами обязательного содержания образования (федеральными государственными образовательными стандартами) (гипотеза заключалась в том, что для терминологических кластеров характерно более высокое число совпадений). С учетом этих трех факторов высчитывалась одна общая метрика, значение которой для каждого кластера варьируется от 1 до 7200. Значение 1 свидетельствует о том, что терминологические кандидаты внутри данного кластера с высокой вероятностью являются терминами.

Значение 7200 свидетельствует о том, что терминологические кандидаты внутри данного кластера с высокой вероятностью являются лжетерминами (см. табл. 1 и табл. 2).

Табл. 1. Пример кластерного деления терминологических кандидатов подкорпуса «Алгебра. 7 класс»

Терминологический кандидат	№ кластера	Наличие в текстах ФГОС	Метрика	Вывод
значение	9	+	1	термин
аргумент	9	+	1	термин
зависимость	9	+	1	термин
прямой_пропорциональность	9	+	1	термин
обратный_пропорциональность	9	+	1	термин
соответствующий_значение	9	—	1	термин
значение_переменный	9	—	1	термин
числовой_значение	9	—	1	термин
значение_переменная	9	—	1	термин
помощь_график	9	—	1	термин
маленький_значение	9	+	1	термин
функция	9	+	1	термин
линейный_функция	9	+	1	термин
положительный_значение	9	—	1	термин
определение_функция	9	—	1	термин
независимый_переменный	9	—	1	термин
значение_функция	9	+	1	термин
значение_аргумент	9	—	1	термин
область_определение_функция	9	—	1	термин
область_определение	9	+	1	термин
способ_задание	9	+	1	термин
таблица_значение	9	+	1	термин
степенный_функция	9	+	1	термин
область_значение_функция	9	—	1	термин
область_значение	9	—	1	термин
зависимость_переменная	9	—	1	термин
учащийся_школа	10	—	1900	лжетермин
цена_товар	10	—	1900	лжетермин
яблоко	10	—	1900	лжетермин
бутылка_вода	10	—	1900	лжетермин
олово	10	—	1900	лжетермин
тетрадь	10	—	1900	лжетермин
карандаш	10	—	1900	лжетермин
вариант_выбор	10	—	1900	лжетермин
аршин	10	—	1900	лжетермин
расход_электроэнергия	10	—	1900	лжетермин
процентный_содержание	10	—	1900	лжетермин
бидон	10	—	1900	лжетермин
алтын	10	—	1900	лжетермин

Табл. 2. Пример кластерного деления терминологических кандидатов подкорпуса «Русский язык. 6 класс»

Терминологический кандидат	№ кластера	Наличие в текстах ФГОС	Метрика	Вывод
пословица	6	+	2	термин
орфоэпический_словарь	6	—	2	термин
словесный_ударение	6	—	2	термин
постановка_ударение	6	+	2	термин
эпитет	6	+	2	термин
правило_орфография	6	+	2	термин
толкование_слово	6	—	2	термин
случай_затруднение	6	+	2	термин
историзм	6	+	2	термин
архаизм	6	+	2	термин
словообразовательный_гнездо	6	—	2	термин
омонимия	6	+	2	термин
лексика_русский_язык	6	—	2	термин
значение_фразеологизм	6	+	2	термин
слияние	6	+	2	термин
небольшой_текст	6	—	2	термин
фрагмент_текст	6	—	2	термин
пример_употребление	6	—	2	термин
система_русский_язык	6	+	2	термин
специальный_словарь	6	—	2	термин
принцип_русский_орфография	6	—	2	термин
графика	6	+	2	термин
правило_произношение	6	—	2	термин
усечение	6	—	2	термин
олицетворение	6	+	2	термин
синонимия	6	+	2	термин
признак_текст	6	+	2	термин
норма_литературный_язык	6	+	2	термин
орфографический_ошибка	6	—	2	термин
алфавитный_порядок	6	—	2	термин
доступный_источник_информация	6	—	2	термин
каламбур	6	—	2	термин
красный_строка	6	—	2	термин
словообразовательный_словарь	6	+	2	термин
ограниченный_употребление	6	—	2	термин
принадлежность_слово	6	+	2	термин
помощник	6	—	2	термин
нейтральный_лексика	6	—	2	термин
высокий_лексика	6	—	2	термин
связный_текст	6	—	2	термин
диктовка	6	—	2	термин
сочетаемость	6	+	2	термин

Продолжение табл. 2

сочетаемость_слово	6	+	2	термин
разговорный_лексика	6	+	2	термин
книжный_лексика	6	—	2	термин
функциональный_разновидность_язык	6	+	2	термин
начало_текст	6	—	2	термин
лингвистический_словарь	6	+	2	термин
словарь_синоним	6	+	2	термин
скороговорка	6	—	2	термин
пересказ_текст	6	—	2	термин
слоговой_принцип	6	—	2	термин
фонетический_разбор	6	—	2	термин
словарь_антоним	6	—	2	термин
основной_информация	6	+	2	термин
основа_предложение	6	—	2	термин
сказуемое	6	+	2	термин
сложный_предложение	6	+	2	термин
предложение	6	+	2	термин
противопоставление	6	+	2	термин
алфавит	6	+	2	термин
разбор	6	+	2	термин
абзац	6	+	2	термин
междометие	6	+	2	термин
последний_абзац	6	—	2	термин
скобка	6	+	2	термин
слово_словосочетание	6	+	2	термин
предыдущий_упражнение	6	—	2	термин
относительный_местоимение	6	+	2	термин
препинание	6	+	2	термин
словосочетание	6	+	2	термин
подлежащее	6	+	2	термин
называть_признак	6	+	2	термин
вопросительный_предложение	6	+	2	термин
именной_сказуемое	6	—	2	термин
способ_выражение	6	+	2	термин
простой_предложение	6	+	2	термин
глагольный_сказуемое	6	+	2	термин
синтаксический_функция	6	+	2	термин
конец_предложение	6	+	2	термин
сравнительный_оборот	6	+	2	термин
сравнительный_союз	6	—	2	термин
смысловой_отношение	6	+	2	термин
словосочетание_предложение	6	+	2	термин
правило_постановка	6	—	2	термин
особенность_правописание	6	—	2	термин

Продолжение табл. 2

помощь_союз	6	—	2	термин
четверостишие	6	—	2	термин
разбор_предложение	6	—	2	термин
эмоциональный_окраска	6	+	2	термин
синтаксический_разбор	6	—	2	термин
последний_предложение	6	—	2	термин
неполный_предложение	6	+	2	термин
выражение_подлежащее	6	+	2	термин
форма_сравнительный_степень	6	—	2	термин
разряд_местоимение	6	+	2	термин
пример_местоимение	6	—	2	термин
устный_разбор	6	—	2	термин
совет_помощник	6	—	2	термин
ботфорт	1	—	2900	лжетермин
мичман	1	—	2900	лжетермин
крестьянский_изба	1	—	2900	лжетермин
канитель	1	—	2900	лжетермин
ключик	1	—	2900	лжетермин
полный_сила	1	—	2900	лжетермин
веретено	1	—	2900	лжетермин
образец_рассуждение	1	—	2900	лжетермин
соя	1	—	2900	лжетермин
голубок	1	—	2900	лжетермин
жилой_здание	1	—	2900	лжетермин
зеница	1	—	2900	лжетермин
постелька	1	—	2900	лжетермин
орешек	1	—	2900	лжетермин
замочек	1	—	2900	лжетермин
лисий_нора	1	—	2900	лжетермин
колбасник	1	—	2900	лжетермин
фрагмент_картина	1	—	2900	лжетермин
бирюзовый_небо	1	—	2900	лжетермин
светец	1	—	2900	лжетермин
медвежий_берлога	1	—	2900	лжетермин

Результаты этого анализа, даже с учетом некоторых сохраняющихся ошибок, позволили не только радикально улучшить качество автоматического извлечения терминологической лексики по сравнению с результатами, основанными на применении метрики keyness score, но и создать задел для последующего тематического моделирования текстов корпуса школьных учебников и соотнесения полученных результатов с эксплицированной структурой параграфов этих учебников с целью распределения анализируемых терминологем по тематическим группам и создания базы знаний по русской терминологической и специальной лексике, соответствующей основному содержанию общего образования в соответствии с федеральными стандартами.

3. Составление итогового списка терминов по областям знания и уровням обучения

Последней задачей текущего этапа исследования было сопоставление списков извлеченных терминов по всем областям знания (учебным дисциплинам) и уровням обучения со списками терминов, предусмотренных стандартами обязательного содержания образования (федеральными государственными образовательными стандартами), а также формирование списков специальной лексики, не являющейся в строгом смысле терминологической, но употребляющейся в целевом корпусе с частотой, значительно превышающей референционный корпус.

Для выполнения подзадачи сопоставления извлеченных терминов со списками терминов, предусмотренных стандартами обязательного содержания образования, была прежде всего проанализирована текущая ситуация с нормативными документами. Она выглядит следующим образом.

1) Новый стандарт основного общего образования (5–9 классы) был фактически (но не формально) утвержден в Министерстве просвещения РФ в ноябре 2019 года: его хотели ввести в действие с лета 2021 года, но смена правительства внесла коррективы в эти планы. Новый министр просвещения в январе 2020 года взял паузу для анализа ситуации и обсуждения ключевых вопросов с профессиональным сообществом. Однако в связи с тем, что в старых нормативных документах (о них ниже) отсутствует разбиение по годам обучения, исследовательский коллектив принял решение в любом случае учитывать текст этого проекта наряду с продолжающим формально действовать старым стандартом основного общего образования (5–9 классы) (https://base.garant.ru/55170507/53f89421bbdaf741eb2d1ecc4ddb4c33/), в котором нет ни детального описания предметов, ни разбиения по годам обучения. Предметное содержание для данного стандарта детализируется в примерной основной образовательной программе основного общего образования (https://fgosreestr.ru), одобренной 8 апреля 2015 года; для целей проекта использовалась редакция этой программы от 4 февраля 2020 года. Разделение по годам обучения в данном документе также отсутствует.

2) О состоянии работы над проектом нового стандарта для среднего общего образования (10–11 классы) никакой актуальной достоверной информации найти не удалось, поэтому исследовательский коллектив пользовался формально действующим стандартом (https://base.garant.ru/70188902/8ef641d3b80ff01d34be16ce9bafc6e0/), а также примерной программой среднего общего образования, детализирующей содержание курсов для этого стандарта (https://fgosreestr.ru) и одобренной 28 июня 2016 года. Здесь также нет разделения по годам обучения, но в данном случае оно и не нужно, так как в рамках проекта 10–11 классы рассматривались как единое целое.

Все вышеупомянутые нормативные документы были приведены в электронный формат, обработаны в соответствии с теми же принципами, что и тексты учебников: деление текста на предложения; токенизация, деление предложений на слова; удаление знаков препинания; POS-тегирование, лемматизация/стемминг корпуса; построение списков частотных лемм, а также биграмм и триграмм. После препроцессинга было осуществлено автоматическое сравнение полученных списков с итоговыми списками терминологических кандидатов.

Результаты по учебным дисциплинам и классам выглядят следующим образом.

Общее количество терминологических кандидатов — 48 911 единиц. Из них: уникальных единиц — 24 650; терминологической лексики — 26 282 единицы; специальной

лексики — 22 629 единиц; совпадений со списками ФГОС — 19 199 единиц. В том числе терминологических кандидатов по уровням обучения: 5 класс — 4 637 единиц; 6 класс — 5 425 единиц; 7 класс — 7 305 единиц; 8 класс — 8 341 единица; 9 класс — 7 425 единиц; 10–11 классы — 12 104 единицы. (Отдельно учитывались случаи объединения нескольких классов в одном учебнике: 5–6 классы — 1 235 единиц; 5–7 классы — 842 единицы; 8–9 классы — 1 607 единиц).

По рассмотренным учебным дисциплинам распределение выглядит следующим образом:

1) алгебра: общее количество терминологических кандидатов — 1 655 единиц; всего терминологической лексики — 1 526 единиц; специальной лексики — 129 единиц; совпадений со списками ФГОС — 652 единицы;

2) астрономия: общее количество терминологических кандидатов — 721 единица; всего терминологической лексики — 456 единиц; специальной лексики — 265 единиц; совпадений со списками ФГОС — 183 единицы;

3) биология: общее количество терминологических кандидатов — 3 863 единицы; всего терминологической лексики — 2 324 единицы; специальной лексики — 1 539 единиц; совпадений со списками ФГОС — 1 149 единиц;

4) всеобщая история и история России: общее количество терминологических кандидатов — 5 198 единиц; всего терминологической лексики — 2 491 единица; специальной лексики — 2 707 единиц; совпадений со списками ФГОС — 1 950 единиц;

5) география: общее количество терминологических кандидатов — 3 352 единицы; всего терминологической лексики — 1 635 единиц; специальной лексики — 1 717 единиц; совпадений со списками ФГОС — 1 402 единицы;

6) геометрия: общее количество терминологических кандидатов — 806 единиц; всего терминологической лексики — 570 единиц; специальной лексики — 236 единиц; совпадений со списками ФГОС — 296 единиц;

7) естествознание: общее количество терминологических кандидатов — 875 единиц; всего терминологической лексики — 198 единиц; специальной лексики — 677 единиц; совпадений со списками ФГОС — 398 единиц;

8) изобразительное искусство: общее количество терминологических кандидатов — 2 875 единиц; всего терминологической лексики — 808 единиц; специальной лексики — 2 067 единиц; совпадений со списками ФГОС — 911 единиц;

9) информатика: общее количество терминологических кандидатов — 1 294 единицы; всего терминологической лексики — 682 единицы; специальной лексики — 612 единиц; совпадений со списками ФГОС — 621 единица;

10) литература: общее количество терминологических кандидатов — 5 094 единицы; всего терминологической лексики — 2 306 единиц; специальной лексики — 2 788 единиц; совпадений со списками ФГОС — 1 124 единицы;

11) математика: общее количество терминологических кандидатов — 1 133 единицы; всего терминологической лексики — 903 единицы; специальной лексики — 230 единиц; совпадений со списками ФГОС — 410 единиц;

12) математический анализ: общее количество терминологических кандидатов — 670 единиц; всего терминологической лексики — 635 единиц; специальной лексики — 35 единиц; совпадений со списками ФГОС — 294 единицы;

13) мировая художественная культура: общее количество терминологических кандидатов — 722 единицы; всего терминологической лексики — 215 единиц; специальной лексики — 507 единиц; совпадений со списками ФГОС — 182 единицы;

14) музыка: общее количество терминологических кандидатов — 1 566 единиц; всего терминологической лексики — 46 единиц; специальной лексики — 1 520 единиц; совпадений со списками ФГОС — 951 единица;

15) обществознание: общее количество терминологических кандидатов — 3 972 единицы; всего терминологической лексики — 2 286 единиц; специальной лексики — 1 686 единиц; совпадений со списками ФГОС — 2 461 единица;

16) право: общее количество терминологических кандидатов — 876 единиц; всего терминологической лексики — 404 единицы; специальной лексики — 472 единицы; совпадений со списками ФГОС — 396 единиц;

17) русский язык: общее количество терминологических кандидатов — 3 780 единиц; всего терминологической лексики — 2 633 единицы; специальной лексики — 1 147 единиц; совпадений со списками ФГОС — 1 668 единиц;

18) технология: общее количество терминологических кандидатов — 2 554 единицы; всего терминологической лексики — 406 единиц; специальной лексики — 2 148 единиц; совпадений со списками ФГОС — 1 186 единиц;

19) физика: общее количество терминологических кандидатов — 3 254 единицы; всего терминологической лексики — 2 836 единиц; специальной лексики — 418 единиц; совпадений со списками ФГОС — 1 235 единиц;

20) физическая культура: общее количество терминологических кандидатов — 2 556 единиц; всего терминологической лексики — 1 161 единица; специальной лексики — 1 395 единиц; совпадений со списками ФГОС — 765 единиц;

21) химия: общее количество терминологических кандидатов — 2 095 единиц; всего терминологической лексики — 1 807 единиц; специальной лексики — 288 единиц; совпадений со списками ФГОС — 865 единиц.

Результаты анализа (см. схему 2) показывают существенные расхождения между учебными дисциплинами в том, что можно назвать «коэффициентом терминологической насыщенности», — в отношении числа собственно терминов к числу всех лексем, употребляющихся в целевом корпусе с относительной частотой, значительно превышающей общеязыковую.

На схеме видно отчетливое противопоставление гуманитарных и естественнонаучных дисциплин: для первых характерно преобладание доли специальной лексики, для вторых — доли лексики собственно терминологической. Эти наблюдения должны быть учтены на следующих этапах проекта исследования терминологических подсистем современных школьных учебников на русском языке, в ходе которых в том числе предполагается: 1) сопоставить полученные данные по корпусу школьных учебников с данными предобученных дистрибутивно-семантических моделей, предоставляемых сервисом RusVectores (Национальный корпус русского языка и Википедия) (Kutuzov, Kuzmenko 2017), и данными модели, обученной на корпусе специальных научных текстов, с целью сравнения принципов организации анализируемых терминологем в векторном пространстве целевого корпуса и корпусов, содержащих аналогичные понятия в бытовой, научно-популярной и собственно научной сферах; 2) провести тематическое моделирование текстов корпуса школьных учебников и соотнесение полученных результатов с эксплицированной структурой параграфов этих учебников с целью распределения анализируемых терминологем по тематическим группам; 3) создать базу знаний по русской терминологической и специальной лексике, соответствующей основному содержанию общего образования в соответствии с федеральными стандартами; 4) создать и обучить глубокую нейросеть, способную по поданной на вход группе векторных представлений терминов определить учебную дисциплину, уровень обучения и учебную тему.

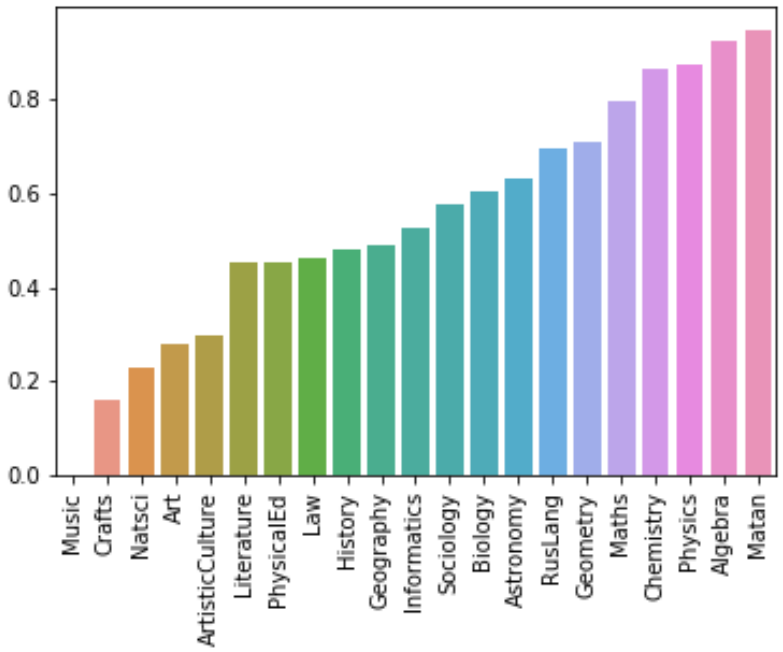

Схема 2. Коэффициенты терминологической насыщенности для разных областей знания

Заключение

Описанные в настоящей статье методы автоматического извлечения терминологии представляют, применительно к проблеме состава и функционирования терминов в школьном учебном тексте, эвристические возможности, недоступные традиционным описательным процедурам. В частности: 1) появляется возможность проследить за частотностью того или иного термина в разных типах школьного учебного текста и оценить их терминологическую плотность; 2) становится возможным сопоставление прототипического и ближайшего для термина семантического окружения — его дефиниции — и разновидностей его семантического окружения в многообразии контекстов употребления; 3) становится возможно охарактеризовать текстуально реализуемые синонимические и антонимические связи терминов, а также терминов и слов нетерминологического характера, используемых в школьных учебниках; 4) выявляется терминологическое и нетерминологическое употребление одного и того же слова внутри одного учебника и учебников, относящихся к разным школьным предметам; 5) формируется база для сопоставительно-типологической характеристики употребления различных терминов внутри одной дисциплины и в разных дисциплинах школьного обучения. Кроме того, становится возможным сопоставление поведения одних и тех же терминов в школьных учебниках, в собственно научных текстах и неспециальных текстах.

Указанные явления могут быть охарактеризованы как синхронно-описательно в рамках отдельно взятого учебника, так и сопоставительно — при сравнении учебников по разным школьным предметам. Отдельным аспектом исследования является характеристика данных явлений в динамике — при последовательном анализе учебников разных классов, от 5-го к 11-му.

Полученные результаты исследования могут представлять фундаментальный теоретический интерес для терминоведения. В прикладном отношении эти результаты могут

быть использованы для составления рекомендаций авторам школьных учебных пособий и для создания школьного толкового словаря, включающего в себя весь терминологический состав, актуальный для современной школьной учебной литературы.

Примечания

Все материалы, необходимые для воспроизведения результатов и верификации выводов статьи, за исключением текстов учебников, являющихся собственностью правообладателя, размещены в постоянном научном хранилище Zenodo и доступны по адресу: https://zenodo.org/record/4079198#.X4Mrfy1h29Y.

Литература

Караулов, Ю. Н. (1991) *О состоянии русского языка современности: Доклады на конференции «Русский язык и современность. Проблемы и перспективы развития русистики» и материалы почтовой дискуссии, в которой приняли участие Ю. Д. Апресян и др*. М.: Институт русского языка РАН, 66 с.

Лейчик, В. М. (2007) *Терминоведение: предмет, методы, структура*. 3-е изд. М.: Изд-во ЛКИ, 256 с.

Ментруп, В. К. (1983) К проблеме лексикографического описания общенародного языка и международных языков. В кн.: Н. Н. Попов (ред.). *Новое в зарубежной лингвистике. Вып. 14. Проблемы и методы лексикографии*. М.: Прогресс, с. 301–333.

Шелов, С. Д. (1998) *Определение терминов и понятийная структура терминологии*. СПб.: Изд-во СПбГУ, 236 с.

Brownlee, J. (2017) *Deep learning for natural language processing: Develop deep learning models for your natural language problems*. Vermont: Machine Learning Mastery Publ., 414 p.

Cabré, M. T., Estopà, R., Vivaldi, J. (2001) Automatic term detection: A review of current systems. In: D. Bourigault, Ch. Jacquemin, M.-C. L'Homme (eds.). *Recent advances in computational terminology*. Amsterdam: John Benjamins Publ., pp. 53–87. https://doi.org/10.1075/nlp.2.04cab

Durda, K., Buchanan, L. (2008) Windsors: Windsor improved norms of distance and similarity of representations of semantics. *Behavior Research Methods*, 40 (3): 705–712. https://www.doi.org/10.3758/BRM.40.3.705

Jones, M. N., Mewhort, D. J. K. (2007) Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, 114 (1): 1–37. https://doi.org/10.1037/0033-295X.114.1.1

Kilgarriff, A., Jakubíček, M., Kovář, V. et al. (2014) Finding terms in corpora for many languages with the sketch engine. In: *Proceedings of the Demonstrations at the 14th Conference the European Chapter of the Association for Computational Linguistics, Sweden, April 2014*. Gothenburg: Association for Computational Linguistics Publ., pp. 53–56. https://www.doi.org/10.3115/v1/E14-2014

Kutuzov, A., Kuzmenko, E. (2017) WebVectors: A toolkit for building web interfaces for vector semantic models. In: D. Ignatov et al. (eds.). *Analysis of images, social networks and texts. AIST 2016. Communications in computer and information science. Vol. 661.* Cham: Springer Publ., pp. 155–161. https://www.doi.org/10.1007/978-3-319-52920-2_15

Levy, O., Goldberg, Y. (2014) Linguistic regularities in sparse and explicit word representations. In: R. Morante, S. W-t. Yih (eds.). *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*. Stroudsburg, PA: Association for Computational Linguistic Publ., pp. 171–180. https://www.doi.org/10.3115/v1/W14-1618

Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013) Efficient estimation of word representations in vector space. In: *Proceedings of Workshop at ICLR, 2013.* [Online]. Available at: https://arxiv.org/abs/1301.3781 (accessed 20.02.2021).

Mikolov, T., Sutskever, I., Chen, K. et al. (2013) Distributed representations of words and phrases and their compositionality. In: *NIPS'13: Proceedings of the 26th International Conference on Neural Information Processing Systems. Vol. 2.* Red Hook: Curran Associates Publ., pp. 3111–3119.

Mikolov, T., Yih, W.-t., Zweig, G. (2013) Linguistic regularities in continuous space word representations. In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* Atlanta: Association for Computational Linguistics Publ., pp. 746–751.

Rohde, D. L., Gonnerman, L. M., Plaut, D. C. (2006) An improved model of semantic similarity based on lexical co-occurrence. *Communications of the ACM*, 8: 627–633.

References

Brownlee, J. (2017) *Deep learning for natural language processing: Develop deep learning models for your natural language problems*. Vermont: Machine Learning Mastery Publ., 414 p. (In English)

Cabré, M. T., Estopà, R., Vivaldi, J. (2001) Automatic term detection: A review of current systems. In: D. Bourigault, Ch. Jacquemin, M.-C. L'Homme (eds.). *Recent advances in computational terminology*. Amsterdam: John Benjamins Publ., pp. 53–87. https://doi.org/10.1075/nlp.2.04cab (In English)

Durda, K., Buchanan, L. (2008) Windsors: Windsor improved norms of distance and similarity of representations of semantics. *Behavior Research Methods*, 40 (3): 705–712. https://www.doi.org/10.3758/BRM.40.3.705 (In English)

Jones, M. N., Mewhort, D. J. K. (2007) Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, 114 (1): 1–37. https://doi.org/10.1037/0033-295X.114.1.1 (In English)

Karaulov, Yu. N. (1991) *O sostoyanii russkogo yazyka sovremennosti: Doklady na konferentsii "Russkij yazyk i sovremennost'. Problemy i perspektivy razvitiya rusistiki" i materialy pochtovoj diskussii, v kotoroj prinyali uchastie Yu. D. Apresyan i dr*. Moscow: V. V. Vinogradov Russian Language Institute of the Russian Academy of Sciences Publ., 66 p. (In Russian)

Kilgarriff, A., Jakubíček, M., Kovář, V. et al. (2014) Finding terms in corpora for many languages with the sketch engine. In: *Proceedings of the Demonstrations at the 14th Conference the European Chapter of the Association for Computational Linguistics, Sweden, April 2014*. Gothenburg: Association for Computational Linguistics Publ., pp. 53–56. https://www.doi.org/10.3115/v1/E14-2014 (In English)

Kutuzov, A., Kuzmenko, E. (2017) WebVectors: A toolkit for building web interfaces for vector semantic models. In: D. Ignatov et al. (eds.). *Analysis of images, social networks and texts. AIST 2016. Communications in computer and information science. Vol. 661.* Cham: Springer Publ., pp. 155–161. https://www.doi.org/10.1007/978-3-319-52920-2_15 (In English)

Lejchik, V. M. (2007) *Terminovedenie: predmet, metody, struktura*. Moscow: URSS Publ., 256 p. (In Russian)

Levy, O., Goldberg, Y. (2014) Linguistic regularities in sparse and explicit word representations. In: R. Morante, S. W-t. Yih (eds.). *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*. Stroudsburg: Association for Computational Linguistic Publ., pp. 171–180. https://www.doi.org/10.3115/v1/W14-1618 (In English)

Mentrup, V. K. (1983) K probleme leksikograficheskogo opisaniya obshchenarodnogo yazyka i mezhdunarodnykh yazykov. In: N. N. Popov (ed.). *Novoe v zarubezhnoj lingvistike. Vyp. 14. Problemy i metody leksikografii*. Moscow: Progress Publ., pp. 301–333. (In Russian)

Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013) Efficient estimation of word representations in vector space. In: *Proceedings of Workshop at ICLR,* 2013. [Online]. Available at: https://arxiv.org/abs/1301.3781 (accessed 20.02.2021). (In English)

Mikolov, T., Sutskever, I., Chen, K. et al. (2013) Distributed representations of words and phrases and their compositionality. In: *NIPS'13: Proceedings of the 26th International Conference on Neural Information Processing Systems. Vol. 2.* Red Hook: Curran Associates Publ., pp. 3111–3119. (In English)

Mikolov, T., Yih, W.-t., Zweig, G. (2013) Linguistic regularities in continuous space word representations. In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* Atlanta: Association for Computational Linguistics Publ., pp. 746–751. (In English)

Rohde, D. L., Gonnerman, L. M., Plaut, D. C. (2006). An improved model of semantic similarity based on lexical co-occurrence. *Communications of the ACM*, 8: 627–633. (In English)

Shelov, S. D. (1998) *Opredelenie terminov i ponyatijnaya struktura terminologii*. Saint Petersburg: Saint Petersburg University Press, 236 p. (In Russian)

Приложение. Список школьных учебников на русском языке, составивших исследовательский корпус

Номер	Название	Авторы	Класс
1.2.1.1.1.1	Русский язык (в 2 частях)	Чердаков Д. Н., Дунев А. И., Вербицкая Л. А. и др. / Под ред. Л. А. Вербицкой	5
1.2.1.1.1.2	Русский язык (в 2 частях)	Чердаков Д. Н., Дунев А. И., Пугач В. Е. и др. / Под ред. Л. А. Вербицкой	6
1.2.1.1.1.3	Русский язык	Чердаков Д. Н., Дунев А. И., Пугач В. Е. / Под ред. Л. А. Вербицкой	7
1.2.1.1.1.4	Русский язык	Чердаков Д. Н., Дунев А. И., Пугач В. Е. / Под ред. Л. А. Вербицкой	8
1.2.1.1.1.5	Русский язык	Чердаков Д. Н., Дунев А. И., Пугач В. Е. / Под ред. Л. А. Вербицкой	9
1.2.1.1.3.1	Русский язык (в 2 частях)	Ладыженская Т. А., Баранов М. Т., Тростенцова Л. А. и др.	5
1.2.1.1.3.2	Русский язык (в 2 частях)	Баранов М. Т., Ладыженская Т. А., Тростенцова Л. А. и др.	6
1.2.1.1.3.3	Русский язык	Баранов М. Т., Ладыженская Т. А., Тростенцова Л. А. и др.	7
1.2.1.1.3.4	Русский язык	Бархударов С. Г., Крючков С. Е., Максимов Л. Ю. и др.	8
1.2.1.1.3.5	Русский язык	Бархударов С. Г., Крючков С. Е., Максимов Л. Ю. и др.	9
1.2.1.1.5.1	Русский язык (в 2 частях)	Рыбченкова Л. М., Александрова О. М., Глазков А. В. и др.	5
1.2.1.1.5.2	Русский язык (в 2 частях)	Рыбченкова Л. М., Александрова О. М., Загоровская О. В. и др.	6
1.2.1.1.5.3	Русский язык (в 2 частях)	Рыбченкова Л. М., Александрова О. М., Загоровская О. В. и др.	7
1.2.1.1.5.4	Русский язык	Рыбченкова Л. М., Александрова О. М., Загоровская О. В. и др.	8
1.2.1.1.5.5	Русский язык	Рыбченкова Л. М., Александрова О. М., Загоровская О. В. и др.	9
1.2.1.2.2.1	Литература (в 2 частях)	Коровина В. Я., Журавлёв В. П., Коровин В. И.	5
1.2.1.2.2.2	Литература (в 2 частях)	Полухина В. П., Коровина В. Я., Журавлёв В. П. и др. / Под ред. В. Я. Коровиной	6
1.2.1.2.2.3	Литература (в 2 частях)	Коровина В. Я., Журавлёв В. П., Коровин В. И.	7
1.2.1.2.2.4	Литература (в 2 частях)	Коровина В. Я., Журавлёв В. П., Коровин В. И.	8
1.2.1.2.2.5	Литература (в 2 частях)	Коровина В. Я., Журавлёв В. П., Збарский И. С. и др. / Под ред. В. Я. Коровиной	9
1.2.1.2.5.1	Литература (в 2 частях)	Чертов В. Ф., Трубина Л. А., Ипполитова Н. А. и др. / Под ред. В. Ф. Чертова	5
1.2.1.2.5.2	Литература (в 2 частях)	Чертов В. Ф., Трубина Л. А., Ипполитова Н. А. и др. / Под ред. В. Ф. Чертова	6
1.2.1.2.5.3	Литература (в 2 частях)	Чертов В. Ф., Трубина Л. А., Ипполитова Н. А. и др. / Под ред. В. Ф. Чертова	7
1.2.1.2.5.4	Литература (в 2 частях)	Чертов В. Ф., Трубина Л. А., Антипова А. М. и др. / Под ред. В. Ф. Чертова	8
1.2.1.2.5.5	Литература (в 2 частях)	Чертов В. Ф., Трубина Л. А., Антипова А. М. и др. / Под ред. В. Ф. Чертова	9
1.2.3.2.1.1	Всеобщая история. История Древнего мира	Вигасин А. А., Годер Г. И., Свенцицкая И. С. / Под ред. А. А. Искендерова	5
1.2.3.2.1.2	Всеобщая история. История Средних веков	Агибалова Е. В., Донской Г. М. / Под ред. А. А. Сванидзе	6

Продолжение приложения

1.2.3.2.1.3	Всеобщая история. История Нового времени	Юдовская А. Я., Баранов П. А., Ванюшкина Л. М. / Под ред. А. А. Искендерова	7
1.2.3.2.1.4	Всеобщая история. История Нового времени	Юдовская А. Я., Баранов П. А., Ванюшкина Л. М. и др. / Под ред. А. А. Искендерова	8
1.2.3.2.1.5	Всеобщая история. Новейшая история	Юдовская А. Я., Баранов П.А., Ванюшкина Л. М. и др. / Под ред. А. А. Искендерова	9
1.2.3.2.3.1	Всеобщая история. Древний мир	Уколова В. И.	5
1.2.3.2.3.2	Всеобщая история. Средние века	Ведюшкин В. А., Уколова В. И.	6
1.2.3.2.3.3	Всеобщая история. Новое время	Ведюшкин В. А., Бовыкин Д. Ю.	7
1.2.3.2.3.4	Всеобщая история. Новое время	Бовыкин Д. Ю., Ведюшкин В. А.	8
1.2.3.2.3.5	Всеобщая история. Новое время	Медяков А. С., Бовыкин Д. Ю.	9
1.2.3.3.1.1	Обществознание	Боголюбов Л. Н., Виноградова Н. Ф., Городецкая Н. И. и др.	6
1.2.3.3.1.2	Обществознание	Боголюбов Л. Н., Иванова Л. Ф., Городецкая Н. И. и др.	7
1.2.3.3.1.3	Обществознание	Боголюбов Л. Н., Лазебникова А. Ю., Городецкая Н. И. и др.	8
1.2.3.3.1.4	Обществознание	Боголюбов Л. Н., Лазебникова А. Ю., Матвеев А. И. и др.	9
1.2.3.3.2.1	Обществознание	Котова О. А., Лискова Т. Е.	6
1.2.3.3.2.2	Обществознание	Котова О. А., Лискова Т. Е.	7
1.2.3.3.2.3	Обществознание	Котова О. А., Лискова Т. Е.	8
1.2.3.3.2.4	Обществознание	Котова О. А., Лискова Т. Е.	9
1.2.3.4.1.1	География	Алексеев А. И., Николина В. В., Липкина Е. К. и др.	5–6
1.2.3.4.1.2	География	Алексеев А. И., Николина В. В., Липкина Е. К. и др.	7
1.2.3.4.1.3	География	Алексеев А. И., Николина В. В., Липкина Е. К. и др.	8
1.2.3.4.1.4	География	Алексеев А. И., Николина В. В., Липкина Е. К. и др.	9
1.2.4.1.2.1	Математика	Бунимович Е. А., Дорофеев Г. В., Суворова С. Б. и др.	5
1.2.4.1.2.2	Математика	Бунимович Е. А., Кузнецова Л. В., Минаева С. С. и др.	6
1.2.4.1.3.1	Математика (в 2 частях)	Виленкин А. Н., Жохов В. И., Чесноков А. С. и др.	5
1.2.4.1.3.2	Математика (в 2 частях)	Виленкин А. Н., Жохов В. И., Чесноков А. С. и др.	6
1.2.4.1.6.1	Математика	Дорофеев Г. В., Шарыгин И. Ф., Суворова С. Б. и др. / Под ред. Г. В. Дорофеева, И. Ф. Шарыгина	5
1.2.4.1.6.2	Математика	Дорофеев Г. В., Шарыгин И. Ф., Суворова С. Б. и др. / Под ред. Г. В. Дорофеева, И. Ф. Шарыгина	6
1.2.4.1.9.1	Математика	Никольский С. М., Потапов М. К., Решетников Н. Н. и др.	5
1.2.4.1.9.2	Математика	Никольский С. М., Потапов М. К., Решетников Н. Н. и др.	6
1.2.4.1.10.1	Математика	Ткачёва М. В.	5
1.2.4.1.10.2	Математика	Ткачёва М. В.	6
1.2.4.2.1.1	Алгебра	Бунимович Е. А., Кузнецова Л. В., Минаева С. С. и др.	7
1.2.4.2.1.2	Алгебра	Бунимович Е. А., Кузнецова Л. В., Минаева С. С. и др.	8
1.2.4.2.1.3	Алгебра	Бунимович Е. А., Кузнецова Л. В., Минаева С. С. и др.	9
1.2.4.2.2.1	Алгебра	Дорофеев Г. В., Суворова С. Б., Бунимович Е. А. и др.	7

Продолжение приложения

1.2.4.2.2.2	Алгебра	Дорофеев Г. В., Суворова С. Б., Бунимович Е. А. и др.	8
1.2.4.2.2.3	Алгебра	Дорофеев Г. В., Суворова С. Б., Бунимович Е. А. и др.	9
1.2.4.2.3.1	Алгебра	Колягин Ю. М., Ткачёва М. В., Фёдорова Н. Е. и др.	7
1.2.4.2.3.2	Алгебра	Колягин Ю. М., Ткачёва М. В., Фёдорова Н. Е. и др.	8
1.2.4.2.3.3	Алгебра	Колягин Ю. М., Ткачёва М. В., Фёдорова Н. Е. и др.	9
1.2.4.2.4.1	Алгебра	Макарычев Ю. Н., Миндюк Н. Г., Нешков К. И. и др. / Под ред. С. А. Теляковского	7
1.2.4.2.4.2	Алгебра	Макарычев Ю. Н., Миндюк Н. Г., Нешков К. И. и др. / Под ред. С. А. Теляковского	8
1.2.4.2.4.3	Алгебра	Макарычев Ю. Н., Миндюк Н. Г., Нешков К. И. и др. / Под ред. С. А. Теляковского	9
1.2.4.2.5.1	Алгебра (углубленный уровень)	Макарычев Ю. Н., Миндюк Н. Г., Нешков К. И. и др. / Под ред. С. А. Теляковского	7
1.2.4.2.5.2	Алгебра (углубленный уровень)	Макарычев Ю. Н., Миндюк Н. Г., Нешков К. И. и др. / Под ред. С. А. Теляковского	8
1.2.4.2.5.3	Алгебра (углубленный уровень)	Макарычев Ю. Н., Миндюк Н. Г., Нешков К. И. и др. / Под ред. С. А. Теляковского	9
1.2.4.2.10.1	Алгебра	Никольский С. М., Потапов М. К., Решетников Н. Н. и др.	7
1.2.4.2.10.2	Алгебра	Никольский С. М., Потапов М. К., Решетников Н. Н. и др.	8
1.2.4.2.10.3	Алгебра	Никольский С. М., Потапов М. К., Решетников Н. Н. и др.	9
1.2.4.3.1.1	Геометрия	Атанасян Л. С., Бутузов В. Ф., Кадомцев С. Б. и др.	7–9
1.2.4.3.2.1	Геометрия	Берсенев А. В., Сафонова Н. В.	7
1.2.4.3.2.2	Геометрия	Берсенев А. В., Сафонова Н. В.	8
1.2.4.3.2.3	Геометрия	Берсенев А. В., Сафонова Н. В.	9
1.2.4.3.3.1	Геометрия	Бутузов В. Ф., Кадомцев С. Б., Прасолов В. В. / Под ред. В. А. Садовничего	7
1.2.4.3.3.2	Геометрия	Бутузов В. Ф., Кадомцев С. Б., Прасолов В. В. / Под ред. В. А. Садовничего	8
1.2.4.3.3.3	Геометрия	Бутузов В. Ф., Кадомцев С. Б., Прасолов В. В. / Под ред. В. А. Садовничего	9
1.2.4.3.7.1	Геометрия	Погорелов А. В.	7–9
1.2.5.1.1.1	Физика	Белага В. В., Ломаченков И. А., Панебратцев Ю. А.	7
1.2.5.1.1.2	Физика	Белага В. В., Ломаченков И. А., Панебратцев Ю. А.	8
1.2.5.1.1.3	Физика	Белага В. В., Ломаченков И. А., Панебратцев Ю. А.	9
1.2.5.1.4.1	Физика	Громов С. В., Родина Н. А., Белага В. В. и др. / Под ред. Ю. А. Панебратцева	7
1.2.5.1.4.2	Физика	Громов С. В., Родина Н. А., Белага В. В. и др. / Под ред. Ю. А. Панебратцева	8
1.2.5.1.4.3	Физика	Громов С. В., Родина Н. А., Белага В. В. и др. / Под ред. Ю. А. Панебратцева	9
1.2.5.1.6.1	Физика	Кабардин О. Ф.	7
1.2.5.1.6.2	Физика	Кабардин О. Ф.	8
1.2.5.1.6.3	Физика	Кабардин О. Ф.	9
1.2.5.2.2.1	Биология	Пасечник В. В., Суматохин С. В., Калинова Г. С. и др. / Под ред. В. В. Пасечника	5–6

Продолжение приложения

1.2.5.2.2.2	Биология	Пасечник В. В., Суматохин С. В., Калинова Г. С. и др. / Под ред. В. В. Пасечника	7
1.2.5.2.2.3	Биология	Пасечник В. В., Каменский А. А., Швецов Г. Г. / Под ред. В. В. Пасечника	8
1.2.5.2.2.4	Биология	Пасечник В. В., Каменский А. А., Швецов Г. Г. / Под ред. В. В. Пасечника	9
1.2.5.2.4.1	Биология	Сивоглазов В. И., Плешаков А. А.	5
1.2.5.2.4.2	Биология	Сивоглазов В. И., Плешаков А. А.	6
1.2.5.2.4.3	Биология	Сивоглазов В. И., Сарычева Н. Ю., Каменский А. А.	7
1.2.5.2.4.4	Биология	Сивоглазов В. И., Каменский А. А., Сарычева Н. Ю.	8
1.2.5.2.4.5	Биология	Сивоглазов В. И., Каменский А. А., Сарычева Н. Ю. и др.	9
1.2.5.3.1.1	Химия	Габриелян О. С., Остроумов И. Г., Сладков С. А.	8
1.2.5.3.1.2	Химия	Габриелян О. С., Остроумов И. Г., Сладков С. А.	9
1.2.5.3.3.1	Химия	Журин А. А.	8
1.2.5.3.3.2	Химия	Журин А. А.	9
1.2.5.3.5.1	Химия	Рудзитис Г. Е., Фельдман Ф. Г.	8
1.2.5.3.5.2	Химия	Рудзитис Г. Е., Фельдман Ф. Г.	9
1.2.6.1.1.1	Изобразительное искусство	Горяева Н. А., Островская О. В. / Под ред. Б. М. Неменского	5
1.2.6.1.1.2	Изобразительное искусство	Неменская Л. А. / Под ред. Б. М. Неменского	6
1.2.6.1.1.3	Изобразительное искусство	Питерских А. С., Гуров Г. Е. / Под ред. Б. М. Неменского	7
1.2.6.1.1.4	Изобразительное искусство	Питерских А. С. / Под ред. Б. М. Неменского	8
1.2.6.1.2.1	Изобразительное искусство	Шпикалова Т. Я., Ершова Л. В., Поровская Г. А. и др. / Под ред. Т. Я. Шпикаловой	5
1.2.6.1.2.2	Изобразительное искусство	Шпикалова Т. Я., Ершова Л. В., Поровская Г. А. и др. / Под ред. Т. Я. Шпикаловой	6
1.2.6.1.2.3	Изобразительное искусство	Шпикалова Т. Я., Ершова Л. В., Поровская Г. А. и др. / Под ред. Т. Я. Шпикаловой	7
1.2.6.1.2.4	Изобразительное искусство	Шпикалова Т. Я., Ершова Л. В., Поровская Г. А. и др. / Под ред. Т. Я. Шпикаловой	8
1.2.6.2.1.1	Музыка	Сергеева Г. П., Критская Е. Д.	5
1.2.6.2.1.2	Музыка	Сергеева Г. П., Критская Е. Д.	6
1.2.6.2.1.3	Музыка	Сергеева Г. П., Критская Е. Д.	7
1.2.6.2.1.4	Музыка	Сергеева Г. П., Критская Е. Д.	8
1.2.7.1.1.1	Технология	Казакевич В. М., Пичугина Г. В., Семёнова Г. Ю. и др. / Под ред. В. М. Казакевича	5
1.2.7.1.1.2	Технология	Казакевич В. М., Пичугина Г. В., Семёнова Г. Ю. и др. / Под ред. В. М. Казакевича	6
1.2.7.1.1.3	Технология	Казакевич В. М., Пичугина Г. В., Семёнова Г. Ю. и др. / Под ред. В. М. Казакевича	7
1.2.7.1.1.4	Технология	Казакевич В. М., Пичугина Г. В., Семёнова Г. Ю. и др. / Под ред. В. М. Казакевича	8–9
1.2.8.1.1.1	Физическая культура	Виленский М. Я., Туревский И. М., Торочкова Т. Ю. и др. / Под ред. М. Я. Виленского	5–7

Продолжение приложения

1.2.8.1.1.2	Физическая культура	Лях В. И.	8–9
1.2.8.1.2.1	Физическая культура	Матвеев А. П.	5
1.2.8.1.2.2	Физическая культура	Матвеев А. П.	6–7
1.2.8.1.2.3	Физическая культура	Матвеев А. П.	8–9
1.3.1.1.5.1	Русский язык (базовый уровень)	Рыбченкова Л. М., Александрова О. М., Нарушевич А. Г. и др.	10–11
1.3.1.1.6.1	Русский язык (базовый уровень)	Чердаков Д. Н., Дунев А. И., Вербицкая Л. А. и др. / Под общ. ред. академика РАО Л. А. Вербицкой	10
1.3.1.1.6.2	Русский язык (базовый уровень)	Чердаков Д. Н., Дунев А. И., Вербицкая Л. А. и др. / Под общ. ред. академика РАО Л. А. Вербицкой	11
1.3.1.3.2.1	Литература (базовый уровень) (в 2 частях)	Лебедев Ю. В.	10
1.3.1.3.2.2	Литература (базовый уровень) (в 2 частях)	Михайлов О. Н., Шайтанов И. О., Чалмаев В. А. и др. / Под ред. В. П. Журавлёва	11
1.3.1.3.3.1	Литература (базовый уровень) (в 2 частях)	Свирина Н. М., Фёдоров С. В., Обухова М. Ю. и др. (1 ч.), Фёдоров С. В., Ачкасова Г. Л., Гордиенко Л. Л. и др. (2 ч.) / Под общ. ред. академика РАО Вербицкой Л. А.	10
1.3.1.3.3.2	Литература (базовый уровень) (в 2 частях)	Абелюк Е. С., Поливанов К. М. / Под общ. ред. академика РАО Л. А. Вербицкой	11
1.3.1.3.5.1	Литература (базовый, углубленный уровни) (в 2 частях)	Чертов В. Ф., Трубина Л. А., Ипполитова Н. А. и др./ Под ред. В. Ф. Чертова	10
1.3.1.3.5.2	Литература (базовый, углубленный уровни) (в 2 частях)	Чертов В. Ф., Трубина Л. А., Ипполитова Н. А. и др. / Под ред. В. Ф. Чертова	11
1.3.1.4.1.1	Литература (углубленный уровень) (в 2 частях)	Коровин В. И., Вершинина Н. Л., Капитанова Л. А. и др. / Под ред. В. И. Коровина	10
1.3.1.4.1.2	Литература (углубленный уровень) (в 2 частях)	Коровин В. И., Вершинина Н. Л., Гальцова Е. Д. и др. / Под ред. В. И. Коровина	11
1.3.3.1.1.1	Всеобщая история. Новейшее время (базовый уровень)	Белоусов Л. С., Смирнов В. П., Мейер М. С.	10
1.3.3.1.3.1	История России (базовый уровень) (в 2 частях)	Горинов М. М., Данилов А. А., Моруков М. Ю. и др. / Под ред. А. В. Торкунова	10
1.3.3.1.9.1	История. Всеобщая история. Новейшая история (базовый и углубленный уровни)	Сороко-Цюпа О. С., Сороко-Цюпа А. О. / Под ред. А. А. Искендерова	10
1.3.3.1.10.1	История. Всеобщая история (базовый уровень)	Уколова В. И., Ревякин А. В. / Под ред. А. О. Чубарьяна	10
1.3.3.1.10.2	История. Всеобщая история (базовый уровень)	Улунян А. А., Сергеев Е. Ю. / Под ред. А. О. Чубарьяна	11
1.3.3.3.2.1	География (базовый уровень)	Гладкий Ю. Н., Николина В. В.	10

Продолжение приложения

1.3.3.3.2.2	География (базовый уровень)	Гладкий Ю. Н., Николина В. В.	11
1.3.3.3.5.1	География (базовый уровень)	Лопатников Д. Л.	10–11
1.3.3.3.7.1	География (базовый уровень)	Максаковский В. П.	10–11
1.3.3.8.1.1	Право (углубленный уровень)	Боголюбов Л. Н., Лукашева Е. А., Матвеев А. И. и др. / Под ред. А. Ю. Лазебниковой, Е. А. Лукашевой, А. И. Матвеева	10
1.3.3.8.1.2	Право (углубленный уровень)	Боголюбов Л. Н., Абова Т. Е., Матвеев А. И. и др. / Под ред. А. Ю. Лазебниковой, Т. Е. Абовой, А. И. Матвеева	11
1.3.3.9.1.1	Обществознание (базовый уровень)	Боголюбов Л. Н., Лазебникова А. Ю., Матвеев А. И. и др. / Под ред. Л. Н. Боголюбова, А. Ю. Лазебниковой	10
1.3.3.9.1.2	Обществознание (базовый уровень)	Боголюбов Л. Н., Городецкая Н. И., Лазебникова А. Ю. и др. / Под ред. Л. Н. Боголюбова, А. Ю. Лазебниковой	11
1.3.3.9.2.1	Обществознание (базовый уровень)	Котова О. А., Лискова Т. Е.	10
1.3.3.9.2.2	Обществознание (базовый уровень)	Котова О. А., Лискова Т. Е.	11
1.3.4.1.1.1	Математика: алгебра и начала математического анализа, геометрия. Алгебра и начала математического анализа (базовый и углубленный уровни)	Алимов Ш. А., Колягин Ю. М., Ткачёва М. В. и др.	10–11
1.3.4.1.2.1	Математика: алгебра и начала математического анализа, геометрия. Геометрия (базовый и углубленный уровни)	Атанасян Л. С., Бутузов В. Ф., Кадомцев С. Б. и др.	10–11
1.3.4.1.3.1	Математика: алгебра и начала математического анализа, геометрия. Геометрия (базовый и углубленный уровни)	Бутузов В. Ф., Прасолов В. В. / Под ред. В. А. Садовничего	10–11
1.3.4.1.4.1	Математика: алгебра и начала математического анализа. Геометрия. (базовый уровень)	Вернер А. Л., Карп А. П.	10
1.3.4.1.4.2	Математика: алгебра и начала математического анализа. Геометрия. (базовый уровень)	Вернер А. Л., Карп А. П.	11
1.3.4.1.7.1	Математика: алгебра и начала математического анализа, геометрия. Алгебра и начала математического анализа (базовый и углубленный уровни)	Колягин Ю. М., Ткачёва М. В., Фёдорова Н. Е. и др.	10

Продолжение приложения

1.3.4.1.7.2	Математика: алгебра и начала математического анализа, геометрия. Алгебра и начала математического анализа (базовый и углубленный уровни)	Колягин Ю. М., Ткачёва М. В., Фёдорова Н. Е. и др.	11
1.3.4.1.11.1	Математика: алгебра и начала математического анализа, геометрия. Алгебра и начала математического анализа (базовый и углубленный уровни)	Никольский С. М., Потапов М. К., Решетников Н. Н. и др.	10
1.3.4.1.11.2	Математика: алгебра и начала математического анализа, геометрия. Алгебра и начала математического анализа (базовый и углубленный уровни)	Никольский С. М., Потапов М. К., Решетников Н. Н. и др.	11
1.3.4.1.12.1	Математика: алгебра и начала математического анализа, геометрия. Геометрия (базовый и углубленный уровни)	Погорелов А. В.	10–11
1.3.4.2.1.1	Математика: алгебра и начала математического анализа, геометрия. Геометрия (углубленный уровень)	Александров А. Д., Вернер А. Л., Рыжик В. И.	10
1.3.4.2.1.2	Математика: алгебра и начала математического анализа, геометрия. Геометрия (углубленный уровень)	Александров А. Д., Вернер А. Л., Рыжик В. И.	11
1.3.4.2.4.1	Математика: алгебра и начала математического анализа, геометрия. Алгебра и начала математического анализа (углубленный уровень)	Пратусевич М. Я., Столбов К. М., Головин А. Н.	10
1.3.4.2.4.2	Математика: алгебра и начала математического анализа, геометрия. Алгебра и начала математического анализа (углубленный уровень)	Пратусевич М. Я., Столбов К. М., Головин А. Н.	11
1.3.4.3.2.1	Информатика (базовый уровень)	Гейн А. Г., Юнерман Н. А.	10
1.3.4.3.2.2	Информатика (базовый уровень)	Гейн А. Г., Гейн А. А.	11
1.3.4.3.3.1	Информатика (базовый и углубленный уровень)	Гейн А. Г., Ливчак А. Б., Сенокосов А. И. и др.	10
1.3.4.3.3.2	Информатика (базовый и углубленный уровень)	Гейн А. Г., Сенокосов А. И.	11

Продолжение приложения

1.3.5.1.1.1	Физика (базовый уровень)	Белага В. В., Ломаченков И. А., Панебратцев Ю. А.	10
1.3.5.1.1.2	Физика (базовый уровень)	Белага В. В., Ломаченков И. А., Панебратцев Ю. А.	11
1.3.5.1.7.1	Физика (базовый уровень)	Мякишев Г. Я., Буховцев Б. Б., Сотский Н. Н. / Под ред. Н. А. Парфентьевой	10
1.3.5.1.7.2	Физика (базовый уровень)	Мякишев Г. Я., Буховцев Б. Б., Чаругин В. М. / Под ред. Н. А. Парфентьевой	11
1.3.5.2.1.1	Физика (углубленный уровень)	Кабардин О. Ф., Орлов В. А., Эвенчик Э. Е. и др. / Под ред. А. А. Пинского, О. Ф. Кабардина	10
1.3.5.2.1.2	Физика (углубленный уровень)	Кабардин О. Ф., Глазунов А. Т., Орлов В. А. и др. / Под ред. А. А. Пинского, О. Ф. Кабардина	11
1.3.5.3.2.1	Астрономия (базовый уровень)	Левитан Е. П.	11
1.3.5.3.3.1	Астрономия (базовый уровень)	Чаругин В. М.	10–11
1.3.5.4.1.1	Химия (базовый уровень)	Габриелян О. С., Остроумов И. Г., Сладков С. А.	10
1.3.5.4.1.2	Химия (базовый уровень)	Габриелян О. С., Остроумов И. Г., Сладков С. А.	11
1.3.5.4.3.1	Химия (базовый уровень)	Журин А. А.	10–11
1.3.5.4.5.1	Химия (базовый уровень)	Рудзитис Г. Е., Фельдман Ф. Г.	10
1.3.5.4.5.2	Химия (базовый уровень)	Рудзитис Г. Е., Фельдман Ф. Г.	11
1.3.5.5.2.1	Химия (углубленный уровень)	Пузаков С. А., Машнина Н. В., Попков В. А.	10
1.3.5.5.2.2	Химия (углубленный уровень)	Пузаков С. А., Машнина Н. В., Попков В. А.	11
1.3.5.6.2.1	Биология (базовый уровень)	Беляев Д. К., Дымшиц Г. М., Кузнецова Л. Н. и др. / Под ред. Д. К. Беляева, Г. М. Дымшица	10
1.3.5.6.2.2	Биология (базовый уровень)	Беляев Д. К., Дымшиц Г. М., Бородин П. М. и др. / Под ред. Д. К. Беляева, Г. М. Дымшица	11
1.3.5.6.4.1	Биология (базовый уровень)	Каменский А. А., Касперская Е. К., Сивоглазов В. И.	10
1.3.5.6.4.2	Биология (базовый уровень)	Каменский А. А., Касперская Е. К., Сивоглазов В. И.	11
1.3.5.6.5.1	Биология (базовый уровень)	Пасечник В. В., Каменский А. А., Рубцов А. М. и др. / Под ред. В. В. Пасечника	10
1.3.5.6.5.2	Биология (базовый уровень)	Пасечник В. В., Каменский А. А., Рубцов А. М. и др. / Под ред. В. В. Пасечника	11
1.3.5.6.7.1	Биология (базовый уровень)	Сухорукова Л. Н., Кучменко В. С., Иванова Т. В.	10
1.3.5.6.7.2	Биология (базовый уровень)	Сухорукова Л. Н., Кучменко В. С.	11
1.3.5.7.2.1	Биология (углубленный уровень)	Высоцкая Л. В., Дымшиц Г. М., Рувинский А. О. и др. / Под ред. В. К. Шумного, Г. М. Дымшица	10
1.3.5.7.2.2	Биология (углубленный уровень)	Бородин П. М., Саблина О. В. и др. / Под ред. В. К. Шумного, Г. М. Дымшица	11
1.3.5.7.3.1	Биология. Углубленный уровень (для медицинских классов)	Пасечник В. В., Каменский А. А., Рубцов А. М. и др. / Под ред. В. В. Пасечника	10
1.3.5.7.3.2	Биология. Углубленный уровень (для медицинских классов)	Пасечник В. В., Каменский А. А., Рубцов А. М. и др. / Под ред. В. В. Пасечника	11

1.3.5.8.1.1	Естествознание (базовый уровень)	Алексашина И. Ю., Галактионов К. В., Дмитриев И. С. и др. / Под ред. И. Ю. Алексашиной	10
1.3.5.8.1.2	Естествознание (базовый уровень)	Алексашина И. Ю., Ляпцев А. В., Шаталов М. А. и др. / Под ред. И. Ю. Алексашиной	11
1.3.6.1.2.1	Физическая культура (базовый уровень)	Лях В. И.	10–11
1.3.6.1.3.1	Физическая культура (базовый уровень)	Матвеев А. П.	10–11
1.3.6.2.1.1	Экология (базовый уровень)	Аргунова М. В., Моргун Д. В., Плюснина Т. А.	10–11
2.2.3.5.2.1	Основы финансовой грамотности	Чумаченко В. В., Горяев А. П.	8–9
2.2.4.2.2.1	Информатика	Семёнов А. Л., Рудченко Т. А.	5
2.2.4.2.2.2	Информатика	Семёнов А. Л., Рудченко Т. А.	6
2.2.7.1.1.1	Искусство	Сергеева Г. П., Кашекова И. Э., Критская Е. Д.	8–9
2.3.1.1.1.1	Российское порубежье: мы и наши соседи	Бабурин В. Л., Даньшин А. И., Елховская Л. И. и др.	10–11
2.3.1.1.3.1	Дизайн	Гуров Г. Е.	10–11
2.3.1.1.9.1	Мировая художественная культура	Солодовников Ю. А.	10
2.3.1.1.9.2	Мировая художественная культура	Солодовников Ю. А.	11

Сведения об авторе:

Сергей Игоревич Монахов, SPIN-код: 2620-7258, Scopus Author ID: 57194019779, ORCID: 0000-0002-0759-9998, e-mail: sergomon@gmail.com

Владимир Владимирович Турчаненко, SPIN-код: 2464-8109, e-mail: vladimir.turchanenko@mail.ru

Екатерина Алексеевна Федюкова, e-mail: katefedyukova@gmail.com

Дмитрий Наилевич Чердаков, SPIN-код: 5347-6154, ORCID: 0000-0003-1533-4284, e-mail: dm.cherdakov@gmail.com

Для цитирования: Монахов, С. И., Турчаненко, В. В., Федюкова, Е. А., Чердаков, Д. Н. (2020) Изучение терминологических подсистем современных школьных учебников на русском языке с помощью модели анализа семантики естественных языков Word2Vec. *Journal of Applied Linguistics and Lexicography*, 2 (2): 118–146. https://www.doi.org/10.33910/2687-0215-2020-2-2-118-146

Получена 28 февраля 2021; прошла рецензирование 2 апреля 2021; принята 9 апреля 2021.

Финансирование: Работа выполнена при поддержке гранта РФФИ № 19-29-14032 мк.

Authors

Sergey I. Monakhov, SPIN: 2620-7258, Scopus Author ID: 57194019779, ORCID: 0000-0002-0759-9998, e-mail: sergomon@gmail.com

Vladimir V. Turchanenko, SPIN: 2464-8109, e-mail: vladimir.turchanenko@mail.ru

Ekaterina A. Fedyukova, e-mail: katefedyukova@gmail.com

Dmitry N. Cherdakov, SPIN: 5347-6154, ORCID: 0000-0003-1533-4284, e-mail: dm.cherdakov@gmail.com

For citation: Monakhov, S. I., Turchanenko, V. V., Fedyukova, E. A., Cherdakov, D. N. (2020) Terminological subsystems of modern Russian school textbooks: A study based on Word2Vec and neural networks. *Journal of Applied Linguistics and Lexicography*, 2 (2): 118–146. https://www.doi.org/10.33910/2687-0215-2020-2-2-118-146

Received 28 February 2021; reviewed 2 April 2021; accepted 9 April 2021.

Funding: The reported study was funded by the Russian Foundation for Basic Research, project No. 19-29-14032 mk.