UDC 81`322.2 UDC 373 **DOI: 10.18413/2313-8912-2023-9-1-0-3**

Sergei I. Monakhov[1]
Vladimir V. Turchanenko[2]
Dmitrii N. Cherdakov[3]

Terminology use in school textbooks: corpus analysis

[1] Friedrich Schiller University Jena
1 Fuerstengraben, Jena, 07743, Germany
E-mail: sergomon@gmail.com

[2] Institute of Russian Literature (Pushkinsky Dom) of the Russian Academy of Sciences
4 Makarov Emb., Saint Petersburg, 199034, Russia
E-mail: vladimir.turchanenko@mail.ru

[3] St Petersburg University
7-9 Universitetskaya Emb., Saint Petersburg, 199034, Russia
E-mail: dm.cherdakov@gmail.com

Received 23 January 2023; accepted 13 March 2023; published 30 March 2023

Acknowledgements. The reported study was funded by the Russian Foundation for Basic Research, Project number 19-29-14032 mk “Study of terminological subsystems of modern school textbooks in Russian with the help of word embedding models Word2Vec and neural networks”.

Abstract. The article presents the methods and results of the study that investigated the use of terminology in textbooks for secondary schools in Russia. The data were taken from a full-text DIY corpus of 207 textbooks for grades 5-11. The toolkit included models trained with the Word2Vec algorithms driven by the ideas of distributional semantics. The models were used to improve traditional automatic term extraction based on word frequency statistics. Numerical representation of word collocation patterns and their semantic similarity enabled the following: more effective automatic term extraction with a clear dividing line between terminology *per se* and high-frequency common words; comparative analysis of inventory and functioning of terms in textbooks for different school subjects and grades; analysis of the dynamics of new terms entering educational and methodological complexes and insights into terminological relations between textbooks for different grades. The study included another DIY corpus compiled of scholarly articles across the subjects taught at school. It was used to identify differences in term use in textbooks and scholarly texts as well as in non-specific and popular science contexts. The latter was facilitated by the RusVectōrēs word embedding model. The comprehensive analysis identified some patterns in term functioning relevant for particular school subjects or groups of subjects. The results were evaluated in view of the theory of text complexity, teaching methodology and didactics. The study found some contradictions between the expected and real text complexity. It also showed certain discrepancy between text complexity and basic didactic principles.
Keywords: Term; Terminology; School textbook; Text complexity; Word frequency; Vector representation; Word2Vec; Neural network

How to cite: Monakhov, S. I., Turchanenko, V. V. and Cherdakov, D. N. (2023). Terminology use in school textbooks: corpus analysis, *Research Result. Theoretical and Applied Linguistics*, 9 (1), 27-49. DOI: 10.18413/2313-8912-2023-9-1-0-3

УДК 81`322.2 УДК 373 **DOI: 10.18413/2313-8912-2023-9-1-0-3**

Монахов С. И.[1]
Турчаненко В. В.[2]
Чердаков Д. Н.[3]

Школьный учебный текст в аспекте терминоупотребления: корпусный анализ

[1] Йенский университет им. Ф. Шиллера
Фюрстенграбен, 1, Йена, 07743, Германия
E-mail: sergomon@gmail.com

[2] Институт русской литературы (Пушкинский Дом) РАН
наб. Макарова, 4, Санкт-Петербург, 199034, Россия
E-mail: vladimir.turchanenko@mail.ru

[3] Санкт-Петербургский государственный университет
Университетская наб., 7–9, Санкт-Петербург, 199034, Россия
E-mail: dm.cherdakov@gmail.com

Статья поступила 23 января 2023 г.; принята 13 марта 2023 г.; опубликована 30 марта 2023 г.

Информация об источниках финансирования или грантах: Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 19-29-14032 мк «Изучение терминологических подсистем современных школьных учебников на русском языке с помощью моделей анализа семантики естественных языков Word2Vec и нейронных сетей».

Аннотация. В статье излагаются методы и результаты анализа употребления терминологической лексики в современных школьных учебниках на русском языке. Основным материалом исследования является созданный исследовательский корпус, включающий тексты 207 учебников с 5-го по 11-й класс по 21 школьной дисциплине. Традиционный способ автоматического извлечения терминов, основанный на статистических показателях частотности словоупотребления, предлагается усовершенствовать с помощью создания моделей, обученных по алгоритмам Word2Vec, в основе которых лежат идеи дистрибутивной семантики. Применение этих алгоритмов, выражающее в числовом представлении сочетаемостное поведение слов и соответственно степень их семантической близости, позволило: в существенной мере устрожить результаты автоматического выделения терминов, отграничивая от них высокочастотные нетерминологические единицы; осуществить сопоставительную характеристику состава и употребления терминов в учебниках по разным предметам и разных ступеней обучения; проанализировать динамику пополнения терминологических систем внутри учебно-методических комплексов и охарактеризовать терминологические взаимосвязи между учебниками для отдельных классов. При помощи

специально созданного корпуса научных статей по тем дисциплинам, которые соответствуют предметам школьного обучения, были выявлены различия в употреблении терминов в школьной и научной сферах, а также (с использованием дистрибутивно-семантической модели, предоставляемой ресурсом RusVectōrēs) в сфере общеупотребительной и научно-популярной речи. Для каждого из отмеченных аспектов анализа обнаружены значимые признаки в функционировании терминов, свойственные отдельным школьным дисциплинам или их группам. Полученные результаты оценивались в том числе в свете положений теории сложности текста и принципов дидактики и методики. Отмечены, в частности, случаи противоречия между показателями сложности текста и его предполагаемой трудности, а также неоднозначный характер взаимодействия меры сложности текста с ключевыми дидактическими началами.

Ключевые слова: Термин; Терминология; Школьный учебник; Сложность текста; Частотность слова; Векторное представление; Word2Vec; Нейронная сеть

Информация для цитирования: Монахов С. И., Турчаненко В. В., Чердаков Д. Н. Школьный учебный текст в аспекте терминоупотребления: корпусный анализ // Научный результат. Вопросы теоретической и прикладной лингвистики. 2023. Т. 9. № 1. С. 27-49. DOI: 10.18413/2313-8912-2023-9-1-0-3

1. Introduction

Russian terminology science is rich in history and scope, however, functioning of terms in school textbooks is still under-scrutinized. The V. A. Tatarinov Comprehensive Encyclopedic Dictionary (Tatarinov, 2006) that embraces all possible advances in Soviet and Russian terminology science says nothing about the use of terminology in school-related textual contexts. At the same time, terms are used in school textbooks in thousands to reflect a system of scientific concepts. This makes the reported study relevant as a theoretical contribution to terminology science and research exploring didactic, social, and cultural issues. The study aimed to fill the gap in terminology science with the tools of corpus and computational linguistics that facilitate effective processing of big data.

Approaches to term extraction from large corpora vary (Korkontzelos, Ananiadou, 2014; Stepanova, 2017), however, top among them is the statistical approach that dates back to the 1960s (Piotrovskij, Yastrebova, 1969). This approach is based on the assumption that terms are considerably more frequent in specialized texts than in general texts. Thus, the algorithm compares word frequency in the target corpus meant for term extraction with that in the reference corpus generally representing a collection of non-specialized text (Kilgarriff et al., 2014). As the results of traditional statistical approach to term extraction are still not satisfactory (Cabré et al., 2001), researchers are looking for options to improve its effectiveness with other methods (Mitrofanova and Zakharov, 2009; Lukashevich and Logachev, 2010; Nokel, 2012). We believe that a considerable improvement in automatic term extraction may be achieved through the use of the Word2Vec algorithms (continuous-bag-of-words (CBOW) and skip-gram). The algorithms follow the underlying idea of distributional semantics: the meaning of a word is derived from its lexical context, while mathematically it is represented as a sum of occurrences of the word in various contexts (Rohde et al., 2006; Jones, Mewhort, 2007; Durda and Buchananan, 2008; Turney and Pantel, 2010). Word embedding models trained with the Word2Vec algorithms use vector representations to measure the semantic similarity of words (Mikolov et al., 2013a; Mikolov et al., 2013b; Levy and

Goldberg, 2014; Brownlee, 2017). They have been gaining momentum in the recent decade; however, we have found no evidence of their application to explore the terminology of school textbooks. The major advantage of the proposed methodology is that it allows to track the behavior of semantically related groups of terms. This, in turn, opens up an opportunity to explore term functioning in view of its key property, i.e., its relation to a particular terminology system (Lejchik, 2007: 98–129).

The study outcomes, i.e., the stratification of terminology obtained from school textbooks with the help of corpus and computational linguistics, may also speak to the theory of text complexity. Complexity theory is witnessing an extensive use of automatic text processing and analysis tools; see reviews in (Solovyev et al., 2022; Solnyshkina et al., 2022) and examples of relevant studies in (Flor et al., 2013; Iomdin and Morozov, 2021; Glazkova et al., 2021; Sharoff, 2022). These methods are used to measure the complexity of various educational texts – a major focus of complexity studies. See, for example, a study investigating the complexity of textbooks with evidence taken from text corpora (Solovyev et al., 2018; Martynova et al., 2020).

Terminology is considered one of the lexical indicators of text complexity (Shpakovsky, 2007). By their nature, terms increase text complexity due to a number of factors. First, beyond the boundaries of specialized texts, terms are generally low-frequency words. It is commonly assumed that a high rate of low-frequency words increases text complexity. The assumption has been proved by recent cutting-edge research (Laposhina et al., 2022). Second, despite their reference to specific real objects, terms lean towards conceptual abstraction (Tatarinov, 2006: 231–234). This makes a text more abstract and, hence, more complex (Schwanenflugel, 1991; Fisher et al., 2016). Third, terms, as a rule, have a complex semantics which is unlikely to be familiar to laymen (Mikk, 1981: 65). Semantic complexity is difficult to measure formally (Morozov and Iomdin, 2019). However, it should be borne in mind that terms, as part of a textbook, are the words that are supposed to become known. This is the reason why it is fair to describe them as “unknown”.

When it comes to textbooks or educational and methodological complexes, it is important to keep in mind the dichotomy found in complexity studies: absolute vs. relative text complexity. The former is a total of its objective features, while the latter depends on external factors, namely, cognitive abilities of the reader (Solnyshkina and Kiselnikov, 2015: 86-87; Solnyshkina et al., 2022: 20). High frequency of terms increases text complexity, however, their regular occurrence in a textbook reduces it (Mikk, 1981: 67). The same passage containing terms that are supposed to be mastered is likely to be perceived as more or less difficult depending on its position at the beginning or at the end of the text, respectively. This is due to the knowledge and certain familiarity with terminology that students accumulate over the course of study.

As regards didactics, textbook complexity and difficulty are subject to the dichotomy of didactic principles, i.e., scientificity and comprehensibility. A textbook should be in line with state of the art in science and research. This, *inter alia*, includes terminology which makes an educational text inevitably complex. While being complex, a textbook should be comprehendible. A failure to observe these principles affects educational effectiveness of a textbook.

2. Materials and methods

2.1. Development of target corpora

The reported study included several stages. Among them is the development of research/target corpora, their vectorization, and data clustering.

The priority task was to develop a target corpora of school textbooks. The corpus included 207 textbooks for grades 5-11 published by Prosveshcheniye Publishers. The

research team obtained a permission from the publisher to use their texts for research purposes. The corpus was compiled in 2020. As of 2020, all the indexed textbooks were approved for the use at schools by the Ministry of Education of Russia. The corpus was developed in several stages: scanning, OCR, processing to delete non-letter and punctuation symbols and harmonize the letter case. Special software was used for POS tagging and lemmatization. The corpus totaled a little more than 13 965 000 words. It was then split into subcorpora matching school subjects (a total of 21 subjects): Algebra (18 textbooks; 1 144 089 words), Astronomy (2; 89 574), Biology (21; 1 125 648), General History and History of Russia (15; 973 498), Geography (8; 512 173), Geometry (8; 370 054), Natural Science (2; 158 665), Visual Arts (8; 283 608), Computer Science (6; 284 683), Literature (18; 3 939 054), Mathematics (10; 525 035), Mathematical Analysis (14; 1 134 786), History of World Art (2; 33 130), Music (4; 76 241), Social Science (12; 505 822), Law (2; 171 349), Russian Language (18; 1 131 575), Arts and Crafts (4; 163 574), Physics (15; 1 098 625), Physical Education (7; 301 371), Chemistry (13; 543 283). Then, every subject-specific subcorpus was further split by school grade.

As textbook terminology was investigated comparatively, the reported study needed another research corpus (see Section 3.2 below). This corpus included relevant scholarly articles selected according to a set of principles. We chose articles published in the period from 2016 to 2021 in high-citation wide-scope journals. Each subject area was covered with 2-5 journals. The share of articles from each of the journals was determined by the journal citation index. The corpus included titles, abstracts, and the main body of articles and reports published in key journal sections. For example, out of 100 plus journals in geography indexed in the Russian Index of Science Citation, we only selected three wide-scope journals with strong citation rates: Geography and Natural Resources; Moscow University Bulletin. Series 5, Geography; Proceedings of the Russian Academy of Sciences, Geography Series. The journals have similar average citation rate of 8.51, 9.72, and 9.00, respectively. For this reason, their share in the corpus was almost equal and accounted for 50, 46 and 40 articles, respectively (136 articles in total). The corpus of scholarly articles was split into subcorpora. On the whole, they match the subcorpora of school textbooks. The exceptions are few: the corpora of scholarly articles did not include the textbook subcorpora in Natural Science and Arts and Crafts; the school subcorpora in Mathematics, Algebra, Geometry, and Mathematical Analysis corresponded to a single subcorpus of scholarly articles called Mathematics. The same is true for the Visual Arts and History of World Art textbook subcorpora that corresponded to the Art scholarly subcorpus. The processing of texts for the corpus of scholarly articles followed the same stages as that for the textbook corpus. The size of each subcorpus of scholarly articles was no less than 75% the size of the corresponding thematic subcorpus of school textbooks. E.g., the size of the Geography subcorpus was, respectively, 434 000 words and about 512 000 words for the scholarly and textbook corpus; Biology 853 000 words and about 1 126 000 words; History about 902 000 words and about 973 000 words. The corpus of scholarly articles totaled about 10 795 500 words.

Once the corpora were ready, they were uploaded on the Sketch Engine at https://www.sketchengine.eu. The platform was used for automatic extraction of term candidates based on the comparative analysis of word frequency in target and reference corpora (see above for details). The reference corpus was the Russian Web 2011 Sample (ruTenTen11), available in the Sketch Engine and containing over 900 million word uses from Russian-language Internet texts.

2.2. Automatic term extraction and consequent data vectorization

One-word and multi-word term candidates followed different extraction

algorithms. The keyness score was calculated for every one-word lexical unit with the minimum frequency threshold of three according to the formula ((Lt * 1,000,000 / Ct) +1) / ((Lr * 1,000,000 / Cr) +1), where Lt is the word frequency in the target corpus, Ct the total number of tokens in the target corpus, Lr the word frequency in the reference corpus, Cr the total number of tokens in the reference corpus. A one-word lexical unit became a term candidate if the keyness score was higher than 1. Let us compare, as an example, the keyness score for the words from different subcorpora of the textbook corpus. The Algebra subcorpus (grades 7-9): *многочлен* (polynomial) – 743.2, *множитель* (multiplier) – 380.4, *парабола* (parabola) – 322.3; the Russian Language subcorpus (grade 5): *существительное* (noun) – 479.4, *падеж* (case) – 231.4, *антоним* (antonym) – 170.4; the Biology subcorpus (grade 9): *фотосинтез* (photosynthesis) – 562.3, *фенотип* (phenotype) – 66.1, *цитоплазма* (cytoplasm) – 11.2; the Chemistry subcorpus in the corpus of scholarly articles: *макромолекула* (macromolecule) – 306.5, *адсорбция* (adsorption) – 103.042, *полимеризация* (polymerization) – 67.9; the Astronomy subcorpus: *полуось* (semiaxis) – 197.4, *галактика* (galaxy) – 94.6, *цефеиды* (Cepheids) – 31.3.

The extraction of multi-word term candidates was in two stages. First, we identified word combinations with the minimum frequency threshold of three that had positive Log-Dice scores. This was calculated according to the formula 14 + log(2(|X⋂Y|) / (|X| + |Y|)), where |X| is the absolute frequency of the first item in the word combination in the subcorpus, |Y| the absolute frequency of the second item in the word combination in the subcorpus, and |X⋂Y| the absolute frequency of the word combination in the subcorpus. Once the term candidates were selected, their keyness score was calculated according to the above formula. Let us compare, as an example, the keyness score for the collocations from different subcorpora of the textbook corpus. The Algebra Subcorpus (grades 7-9): *график функции* (function graph) – 725.9, *натуральное число* (natural number) – 200.9, *линейная функция* (linear function) – 97.6; the Russian language subcorpus (grade 5): *часть речи* (part of speech) – 428.2, *единственное число* (singular number) – 222.4, *прошедшее время* (past tense) – 76.1; the Biology subcorpus (grade 9): *бесполое размножение* (asexual reproduction) – 240.4, *пищевая цепь* (food chain) – 190.9, *генная инженерия* (genetic engineering) – 67.0; the Chemistry subcorpus of the corpus of scholarly articles: *элементный анализ* (elemental analysis) – 145.0, *реакционная масса* (reacting mass) – 75.4, *буферный раствор* (buffer solution) – 65.7; the Astronomy subcorpus: *красное смещение* (red shift) – 120.4, *дыра Локмана* (Lockman Hole) –81.0, *солнечный ветер* (solar wind) – 51.2.

The obtained lists of one-word and multi-word terms were sorted in discerning order from the highest to the lowest keyness score. Further processing was made for the first 1 000 terms from the textbook corpus and the first 2 000 terms from the corpus of scholarly articles.

One of the key challenges of the proposed framework is the identification in the obtained word lists of terms *per se* and non-terms that show term-like behavior in the textbook corpus, i.e., have high frequency in the target corpus and low frequency in the reference corpus. Here and in what follows such units will be referred to as pseudoterms. This designation is conventional since automatic delineation between terms and pseudoterms may end up labelling some of the terms *per se* as pseudoterms.

To optimize the obtained results, we used the Word2Vec algorithms. They facilitated vectorization of target corpora as well as development and training of word embedding models. The models, in turn, were used to identify the degree of syntagmatic similarity among automatically extracted terms in each of the DIY subcorpora. Each

subcorpus had two models. One was used for one-word terms, the other for multi-word terms (bigrams and trigrams). The training of models involved the following stages: 1) frequency analysis for each word in the corpus; 2) frequency-based sorting, rare words deleted; 3) Huffman binary tree coding to reduce computational complexity of the algorithm; 4) vectorization of every single word of the corpus. Vectors show the number of cases a given word occurs in the same context window with other high-frequency words from a given corpus. The context window indicates the maximum range between the target and the predicted word in a sentence; 5) using the obtained vectors as an input for a feedforward neural network. The neural network was trained to predict the context by a target word or predict a word by a target context.

Vector representation of words is a tool to evaluate semantic similarity of any pair of words through calculating the cosine measure between their vectors. We calculated the cosine similarity CS = u * v / (||u|| * ||v||) for each word pair. CS was within the range [0,1], where 1 denotes the identity of vectors, i.e., identical contexts of target words implying their semantic similarity; and 0 denotes vector orthogonality, implying a lack of common contexts and, hence, common semes. Compare, as an example, cosine similarity of two pairs of words in the Russian Language subcorpus textbook: *суффикс* (suffix) and *окончание* (ending) – 0.80; *суффикс* (suffix) and *груша* (pear) – 0.18.

The results of vectorization were used to enhance the effectiveness of automatic term extraction. This was done in two ways depending on the target corpus (textbooks or scholarly articles). The reason behind this differentiation is the different nature of texts in each of the corpora – the corpus of scholarly articles is more consistent lexically and structurally.

Semantic mapping was chosen for terms extracted from the textbook corpus. The subcorpus-related maps of term candidate distribution in trained word embedding models were built using t-distributed stochastic neighbor embedding (t-SNE). For visualization, the maps were projected onto a two-dimensional plane from high-dimensional vector space. See examples in Figures 1-2.

We assumed a high probability of black spots on the map denoting clusters with small cosine distance of terms *per se*. Pseudoterms, on the opposite, were expected to be scattered across the rest of the map. K-means were calculated to cluster the points in the plane by their coordinates. Each of the obtained clusters was labeled as either containing terms or pseudoterms. Each semantic map had about 20 clusters and all available points were distributed across them. The following factors were accounted for in cluster labelling: 1) specific share of words that occur within the cluster as independent words or as part of bigrams or trigrams. Assumingly, terminology-rich clusters have a higher word recurrence; 2) specific share of multi-word lexical units within a cluster. It is assumed that terminology-rich clusters have more multi-word units as automatic term extraction is more effective with multi-word candidates; 3) specific share of term candidates within the cluster that match the terminology used in the Federal State Educational Standards of Russia. It is assumed that terminology-rich clusters have more such matches. Taking into account the outlined factors, we calculated a single metric varying from 1 to 7 200 for each cluster with a high probability of a cluster to contain terms or pseudoterms, respectively. As an example, compare the obtained lists of terms for different subcorpora. The Russian Language subcorpus (grade 9) with the metric equaling 1: *русский_язык* (Russian language), *синтаксис* (syntax), *фонетика* (phonetics), *орфография* (orthography), *история_язык* (history_language), *слово* (word), *неологизм* (neologism), *морфема* (morpheme), *этимология* (etymology), *старославянский_язык* (Old Slavic_language), *значение* (meaning), *древнерусский_язык* (Old Russian_language), *современный_русский_язык* (modern_Russian_language), *славянский_язык* (Slavic_language).

Figure 1. Semantic map representing the distribution of term candidates. The Algebra Subcorpus (grade 9)
Рисунок 1. Семантическая карта распределения кандидатов в термины. Подкорпус «Алгебра» (9 класс)

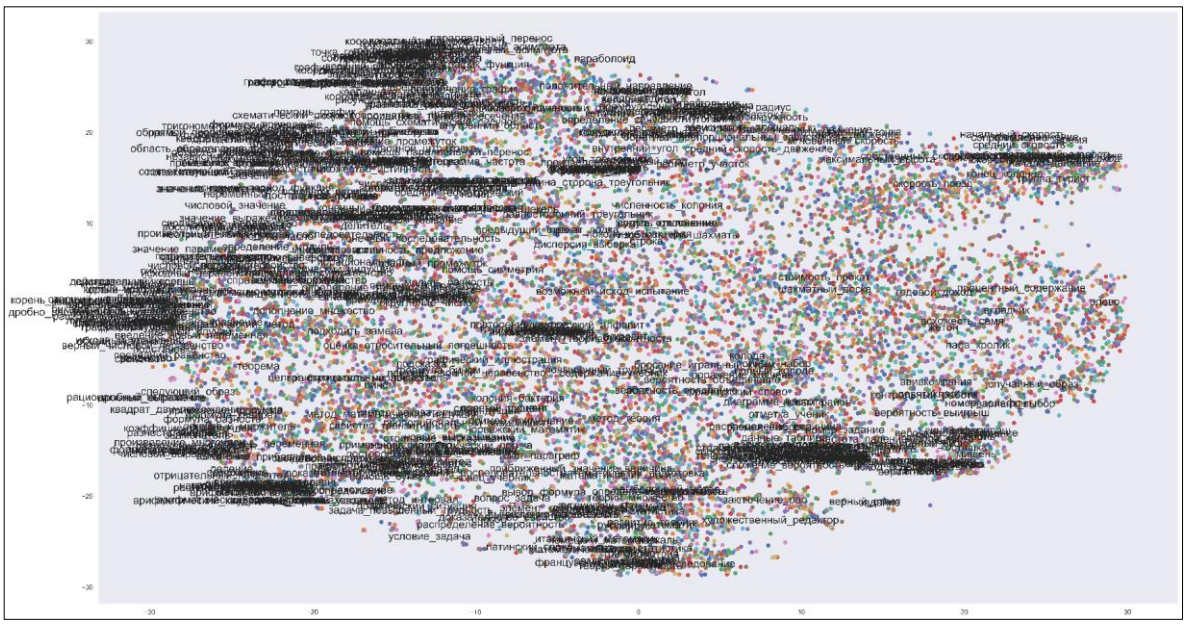
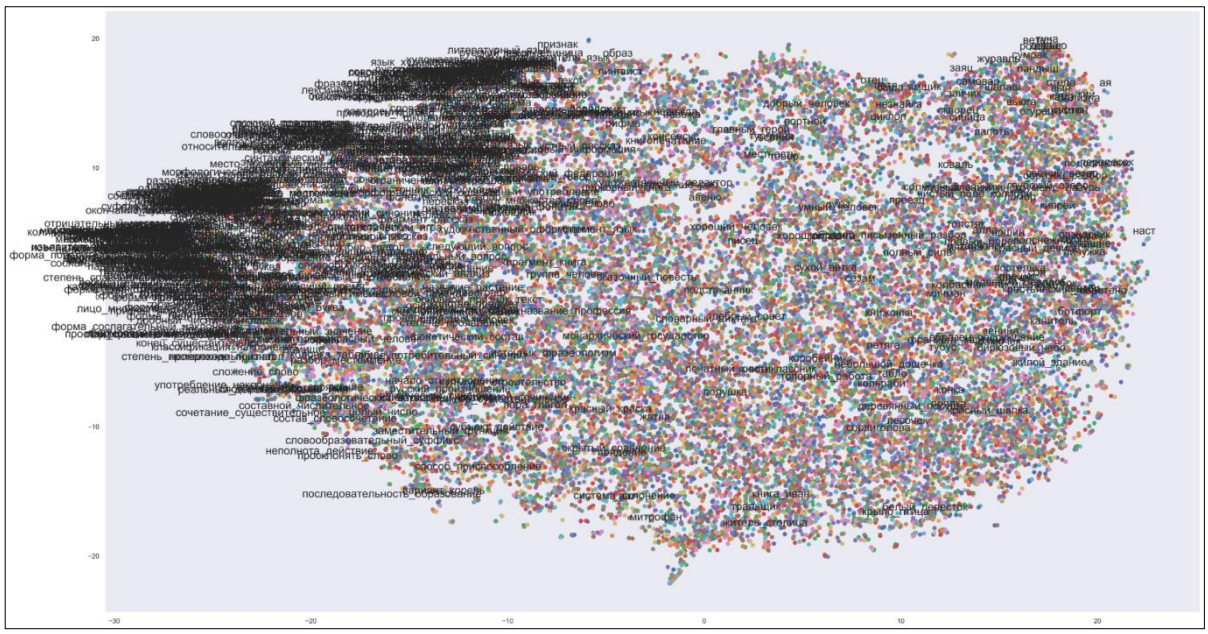


Figure 2. Semantic map representing the distribution of term candidates. The Russian Language Subcorpus (grade 6)
Рисунок 2. Семантическая карта распределения кандидатов в термины. Подкорпус «Русский язык» (6 класс)



For the same subcorpus, the metric with the value 3 600 returned the following words: *ветер* (wind), *дерево* (tree), *осина* (aspen), *оса* (wasp), *колокольчик* (bell), *ель* (spruce), *соловей* (nightingale), *ямщик* (coachman), *рябина* (ashberry), *пирог* (pie), *туман* (fog), *гром* (thunder), *туча* (thundercloud), *роса* (dew). Another example is the Geography subcorpus (grade 7) with the metric equaling 4.6. The list of terms included *воздушный_масса* (air_mass), *форма_рельеф* (shape_relief), *котловина* (basin), *высотный_поясность* (altitude_zonality), *высотный_пояс* (altitude_zone), *землетрясение* (earthquake), *кристаллический_фундамент* (crystalline _foundation), *платформа* (platform), *муссон* (monsoon). For the same subcorpus with the metric value 16.8 the word list contained *причина_образование* (reason_formation), *французский_язык* (French_language), *карта_приложение* (map_application), *сочетание_фактор* (combination_factor), *карта_евразия* (map_Eurasia), *деление_земля* (demarcation_Earth), *бразильский_карнавал* (Brazilian_carnival), *благосостояние_население* (wellbeing_population), *главный_занятие* (major_activity).

To enhance the effectiveness of automatic term extraction from the corpus of scholarly articles, another approach was used. The automatically extracted term candidates were sorted by their semantic distance from the hypothetical center of the lexical system used for general communication purposes. The center was the calculated average value O of all vector representations of the word embedding model that was trained with the Russian National Corpus and is now available on the RusVectōrēs platform (Kutuzov, Kuzmenko, 2017). (1) Vector representation of each term candidate Ci in the original list {LCj}, so that Ci ∈ {LCj}, was compared with O through calculating the cosine distance between the vectors θ(Ci) = cos(Ci, O); (2) the candidate with the cosine distance θ was assigned index 1 indicating a high term probability. It was subsequently deleted from the list so that Ci => K1 and K1 ∈ {KC} ∉ {LCj+1}, if θ(Ci) = argmax(θi … θn); (3) steps 1 and 2 repeated for Ci+1 ∈ {LCj+1} until the list was empty so that {LCn} = ∅ and {KC} = {LCj}. Ultimately, from among the hierarchy of indexes {i … n} in the list {KC}, the index k (i < k < n) was chosen so that the subset of candidates {Kk … Kn} could be excluded from the list {KC} as containing the least probable terms. At this stage of the study, the cut-off point for each subject area was determined by expert decision. Compare, as an example, the first and the last 15 term candidates from the list of the Russian Language subcorpus compiled following the outlined approach. The first 15 candidates are *экспликация* (explication), *предикативность* (predicativity), *именование* (nomination), *модус* (mode), *денотат* (denotatum), *интенция* (intention), *лексема* (lexeme), *актант* (actant), *пресуппозиция* (presupposition), *дескрипция* (description), *модальность* (modality), *семантика* (semantics), *референция* (reference), *предикат* (predicate), *пропозиция* (proposition). Among the last 15 candidates from the list are *мальчик* (boy), *варвара* (varvara), *петя* (petya), *нацu* (nazi), *бенефициант* (beneficiary), *господин* (gentleman), *скотина* (cattle), *парная* (steam room), *зеница* (pupil), *макар* (makar), *жучок* (bug), *обида* (offence), *скука* (boredoom), *тополь* (poplar), *червяк* (worm).

Once the outlined methodology was applied to the original list of term candidates, the total number of one-word and multi-word term candidates for the textbook corpus accounted for 26 328 with the following subcorpora distribution: Algebra – 1 526, Astronomy – 456, Biology – 2 324, General History and History of Russia – 2 491, Geography – 1 635, Geometry – 570, Natural Science – 198, Visual Arts – 808, Computer Science – 682, Literature – 2 306, Mathematics – 903, Mathematical Analysis – 635, History of World Culture – 215, Music – 46, Social Science – 2 286, Law – 404, Russian Language – 2 633, Arts and Crafts – 406, Physics – 2 836, Physical Education – 1 161, Chemistry – 1 807. The same indicator

for the corpus of scholarly articles was 15 247 with the following subcorpora distribution: Astronomy – 1 060, Biology – 1 157, Geography – 1 112, Computer Science – 896, Art – 1 182, History – 891, Literature Studies – 1 101, Mathematics – 753, Musicology – 955, Social Science – 1 116, Law – 1 169, Russian Language / Linguistics – 945, Physics – 999, Physical Education – 892, Chemistry – 1 019.

3. Results and discussion

3.1. Terms *per se* and high-frequency non-terms

Data vectorization increases the effectiveness of automatic term extraction. It also creates a foundation for further analysis of term functioning in school textbooks and beyond.

Notably, automatic term extraction from the textbook corpus generated a considerable number of pseudoterms, i.e., lexical units with high relative frequency in the target corpus that were discarded after vectorization. Pseudoterms show different behavior in subject-specific subcorpora both by quantity and thematically. Used in school textbooks in different subjects, pseudoterms comprise part of special lexical and semantic groups. These lexical groups may be of interest to scholars focusing on teaching and learning methodology for schools and other educational settings.

For obvious reasons, a range of textbooks in particular subjects have pseudoterms that describe subject-specific real-life phenomena. Below is an excerpt from an extensive list of high-frequency plant names from the Biology subcorpus: *абрикос* (apricot), *акация* (acacia), *арахис* (peanut), *астра* (aster), *бегония* (begonia), *белена* (henbane), *бузина* (elderberry), *вишня* (cherry), *георгин* (dahlia), *горох* (pea), *дуб* (oak), *дурман* (thorn apple), *ель* (spruce), *земляника* (wild strawberry), *ива* (willow), *капуста* (cabbage), *картофель* (potato), *кипарис* (cypress), *кислица* (oxalis), *клевер* (clover), *кукуруза* (corn), *ландыш* (lily-of-the-valley), *лещина* (hazelnut), *липа* (linden), *лиственница* (larch), *люпин* (lupine), *люцерна* (medick), *малина* (raspberry), *можжевельник* (juniper), *нарцисс* (daffodil), *одуванчик* (dandelion), *ольха* (alder), *орешник* (hazel tree), *орхидея* (orchid), *осина* (aspen), *осока* (sedge), *пальма* (palm), *папоротник* (fern), *пеларгония* (pelargonium), *пихта* (fir tree), *подорожник* (plantago), *подсолнечник* (sunflower), *пшеница* (wheat), *пырей* (elytrigia), *редис* (radish), *редька* (winter radish), *репа* (turnip), *рыжик* (orange milk cap), *рябина* (ashberry), *саксаул* (saxaul), *сирень* (lilac), *слива* (plum), *сосна* (pine), *томат* (tomato), *тополь* (poplar), *тюльпан* (tulip), *фасоль* (bean), *фиалка* (violet), *фикус* (ficus), *хлопчатник* (cotton plant), *хризантема* (chrysanthemum), *цикорий* (chicory), *шиповник* (rose hip), *эвкалипт* (eucalyptus), *яблоня* (apple tree), *ясень* (ash), *ячмень* (barley).

Another reason for groups of pseudoterms to appear in subject-specific subcorpora is a methodological tradition. The Literature subcorpus (grades 10-11) is marked by a group of automatically extracted abstract nouns with suffixes common for high-flown or purely bookish vocabulary: *безверие* (faithlessness), *возмездие* (retribution), *высокий_предназначение* (lofty_mission), *дарование* (talent), *духовный_возрождение* (spiritual_revival), *жертвенность* (beneficence), *искание* (pursuit), *искренность* (sincerity), *личный_достоинство* (personal_dignity), *мечтание* (dreaming), *мироздание* (universe), *мироощущение* (philosophy of life), *миросозерцание* (worldview), *нравственный_чувство* (moral_feeling), *обличение* (denunciation), *поэтический_вдохновение* (poetic_inspiration), *предчувствие* (presentiment), *преображение* (transfiguration), *прозрение* (insight), *простодушие* (guilelessness), *раздумье* (reflection), *самопожертвование* (self-sacrifice), *скитание* (wandering), *сострадание* (compassion), *сочувствие* (empathy), *страдание* (suffering), *счастие* (happiness), *тщеславие* (vanity), *человечность* (humanity), *чужбина* (foreign lands). These words are selected as term candidates due to their high frequency in

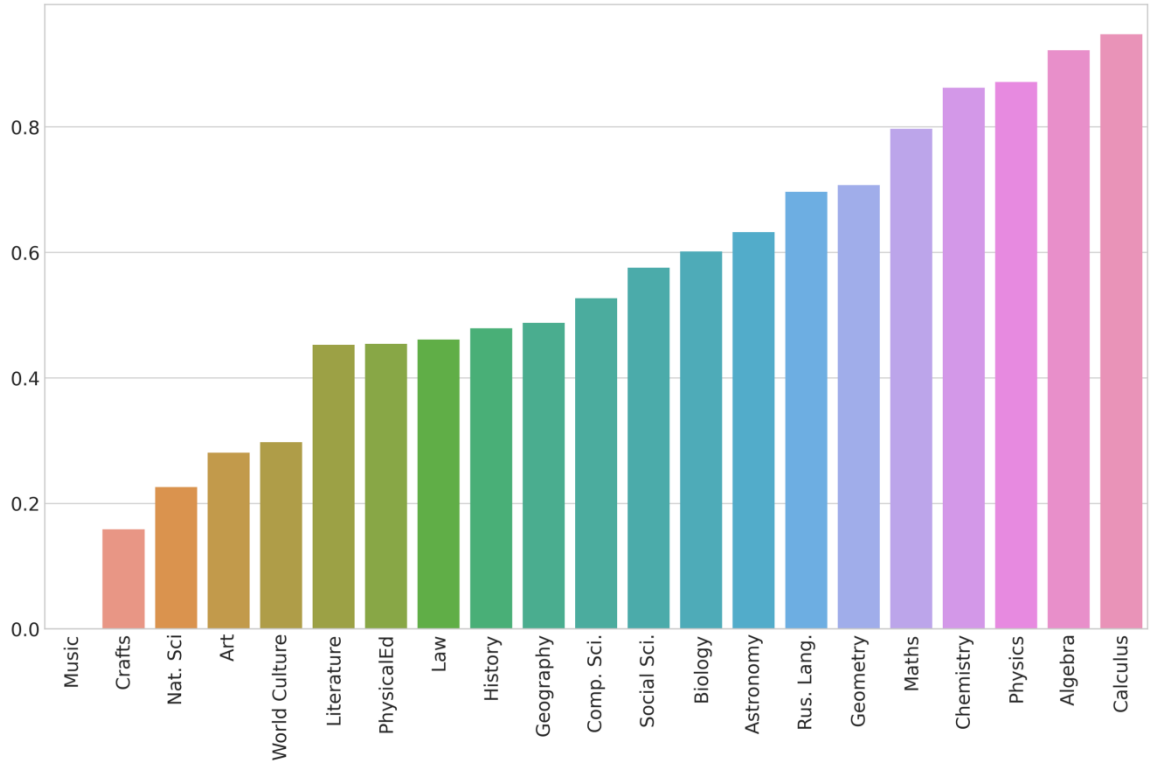
literature textbooks, where they are used to discuss the artistic meaning of literary works or certain aspects of writers' biographies.

The Russian Language subcorpus (grades 5-9) is even more remarkable when it comes to pseudoterms. They include numerous words and word combinations that describe landscapes or natural phenomena. Among them such words and phrases as *аист* (stork), *былинка* (blade of grass), *верба* (American willow), *ветер* (wind), *воробей* (sparrow), *восход_солнце* (sunrise_sun), *вьюга* (blizzard), *глубокий_озеро* (deep_lake), *голубой_небо* (blue_sky), *гроза* (thunderstorm), *гром* (thunder), *дождик* (drizzle), *долгий_зима* (long_winter), *дубрава* (oak forest), *дымок* (smoke), *ель* (spruce), *жаворонок* (lark), *журавль* (crane), *заяц* (hare), *зимний_утро* (winter__morning), *зяблик* (chaffinch), *ива* (weeping willow), *изморозь* (rime ice), *иней* (hoarfrost), *камыш* (reed), *крапива* (nettle), *кукушка* (cuckoo), *лазурь* (azure), *ландыш* (lily-of-the-valley), *лесной_озеро* (forest_lake), *лесной_поляна* (forest_glade), *лесок* (small forest), *липа* (linden), *лиса* (fox), *листопад* (leaf fall), *метель* (blizzard), *наст* (snow crust), *начало_осень* (beginning_fall), *облачко* (cloud), *овраг* (ravine), *озимь* (winter sowing), *орешник* (hazel tree), *оса* (wasp), *осина* (aspen), *осока* (sedge), *перелесок* (shaw), *песок* (sand), *пичужка* (bird), *подосиновик* (aspen bolete), *подснежник* (snowdrop), *поздний_осень* (late_autumn), *пороша* (dusting of snow), *пригорок* (hillock), *проталина* (thaw patch), *ракита* (riverside willow), *родной_природа* (Russian_nature), *роса* (dew), *роща* (grove), *рябина* (ashberry), *свежий_ветер* (breeze), *синий_небо* (blue_sky), *синица* (great tit), *сирень* (lilac), *скворец* (starling), *снегирь* (bullfinch), *снежный_буря* (snow_storm), *сова* (owl), *соловей* (nightingale), *старый-_дуб* (old_oak), *стужа* (cold), *сумрак* (dusk), *туман* (fog), *туча* (thunder cloud), *фиалка* (violet), *холодный_ветер* (cold_wind), *чаща (*thicket), *чибис* (peewit), *шалаш* (hut), *шмель* (bumblebee). It was found that primary school textbooks are dominated by vocabulary describing nature and natural phenomena (Laposhina et al., 2019). It should be noted, however, that the Russian Language textbooks are supposed to contain thematically balanced, emotionally relevant and stylistically diverse vocabulary to effectively fulfill their educational potential. In this regard and despite the methodological tradition, it is fair to suppose that intuitive and now mathematically proven thematic consistency of vocabulary found in textbooks in the subject Russian language does not promise much effectiveness in terms of educational outcomes.

It is of interest to compare the quantity of terms *per se* and pseudoterms in the total high-frequency vocabulary found in different subject-specific subcorpora. Here, textbooks in exact and natural sciences form a juxtaposition with those in arts and humanities with terms *per se* prevailing in the former and pseudoterms more common for the latter (See Figure 3). Interestingly, the Russian Language subcorpus is unexpectedly term-rich with almost as many terms as in the subcorpora comprising textbooks in exact and natural sciences. This is indicative of extensive terminology contained in the Russian Language textbooks. Along with this, the Literature subcorpus in Figure 3 is shown to have fewer terms than the Russian Language textbooks despite the obvious focus of both disciplines on linguistic issues.

The reasons explaining the breakdown in Figure 3 are not trivial. The selection of terms from among high-frequency candidates was facilitated by special algorithms that took into account (a) regular occurrence of terms in the text, which (b) showed syntagmatic patterns similar to those of the terms from the same subject area. A failure to observe the requirement (a) may lead to a failure in the identification of a term from among high-frequency candidates in the target corpus, while a failure to observe (b) may result in misidentification of a term as a pseudoterm.

Figure 3. Share of terms in the total number of high-frequency words in the textbook subcorpora
Рисунок 3. Доля терминов от общего числа высокочастотной лексики в подкорпусах корпуса школьных учебников



With the above said, it may seem paradoxical that the Natural Science subcorpus ranks so low as regards the number of terms *per se*. The Natural Science subcorpus (Nat. Sci.) includes basic upper secondary school textbooks in physics, astronomy, chemistry, and biology. A low share of terms in the subcorpus is not only due to the overall low number of terms, but also due to the specifics of their functioning. These textbooks provide a very brief overview of relevant subject areas which explains low term frequency and lack of patterns in the contextual behavior of groups of terms.

Another salient example is the Music subcorpus. Semantic mapping showed a very low number of terms in comparison with other textbook subcorpora. This indicated that groups of terms extracted automatically with the keyness score (in particular, names of music genres, e.g., cantata, symphony, suite, etc.) do not show similar contextual behavior and are not counted as terms.

Similar cases may be observed in some parts of a subject-specific subcorpus that includes vocabulary of a particular educational and methodological complex. Let us consider as an example the Russian Language educational and methodological complex (grades 5-11) edited by L. Verbitskaya[1]. The textbook for grade 5 features the word combination *орфографический_правило* (spelling_rule) which is not identified as a term in this particular textbook. However, it is identified as such in textbooks for grades 6-7 where it is frequently used and, importantly, belongs to the term cluster that includes such words as *орфограмма* (orthogram),

[1] Russian Language: textbooks for educational organizations / Under the general editorship of Academician of the Russian Academy of Education L. Verbitskaya, Prosveshchenie, Moscow, Saint Petersburg, 2018–2019.

правописание_гласный (spelling_vowel), *правописание_приставка* (spelling_prefix), *правописание_слово* (spelling_word), *правописание_суффикс* (spelling_suffix), *ударение* (stress), *условие_выбор_буква* (condition_choice_letter), etc.

Needless to say, a high share of terms among high-frequency textbook vocabulary is indicative of considerable text complexity. On the other hand, absolute complexity of an educational text resulting from abundant terminology is somewhat compensated by a regular and systemic use of groups of terms that form a semantic whole. By contrast, the presence of terms with insufficient frequency and/or contextual similarity with the words close in meaning may impede understanding of the text while simultaneously reducing its absolute complexity.

3.2. Terminological links

In view of the above, of special importance are the terminological links established both between different parts of one and the same textbook and different textbooks within one and the same educational and methodological complex. These links implement prospection and retrospection fundamental to any text and correspond to the didactic principles of "advance training" and "revision and consolidation", respectively. At the outset, it is interesting to assess the dynamics of term accumulation and specific contribution that textbooks for different grades make into the development of subject-specific terminology systems. The following algorithm was developed to solve this task: 1) every educational and methodological complex was broken down into passages matching the length of one thematic section (1 000 words on average); 2) the number of terms and term combinations previously assigned to terminology systems of different grades were calculated for each passage; 3) the total number of new terms that a particular textbook contributes to the general terminology system of the educational and methodological complex was calculated; 4) the number of terms from the terminology systems of all grades was calculated for each passage.

Below is an example that shows the dynamics of new terms entering the terminology system of the Russian Language educational and methodological complex edited by L. Verbitskaya. Figure 4 depicts the share of new terms for every grade, while Figure 5 shows the same dynamics plus the distribution of terms in different parts of textbooks for different grades (textbooks for grades 10-11 are shown together with the black line).

As is seen, the textbook for grade 5 is very much the leader by the number of terms found in each of the four passages of the educational and methodological complex. Put otherwise, the knowledge of subject-specific terminology is developed at an early stage of education, while every new stage facilitates its revision. Upper secondary school textbooks almost never introduce new terms as they are primarily meant for consolidation and revision.

The study compared the above dynamics with that in the educational and methodological complex in Literature developed by V. Korovina et al. (grades 5-9) and V. Korovin et al. (grades 10-11)[2]. The terminology system of upper secondary school textbooks was found to be very much updated as compared to the choice of terms for earlier grades (see Figure 6). The number of new terms entering the Russian Language and Literature upper secondary textbooks is different as these textbooks are meant for different levels of study: basic vs. advanced, respectively.

[2] Literature: textbooks for educational organizations / Edited by V. Korovina, Prosveshchenie, Moscow, 2012–2013 [Grades 5–9]; Literature: textbooks for educational organizations / Edited by V. Korovin, Prosveshchenie, Moscow, 2012–2019 [Grades 10–11].

Figure 4. The dynamics of new terms entering the terminology system of the Russian Language educational and methodological complex edited by L. Verbitskaya
Рисунок 4. Динамика пополнения терминосостава в учебно-методическом комплексе «Русский язык» под ред. Л. А. Вербицкой

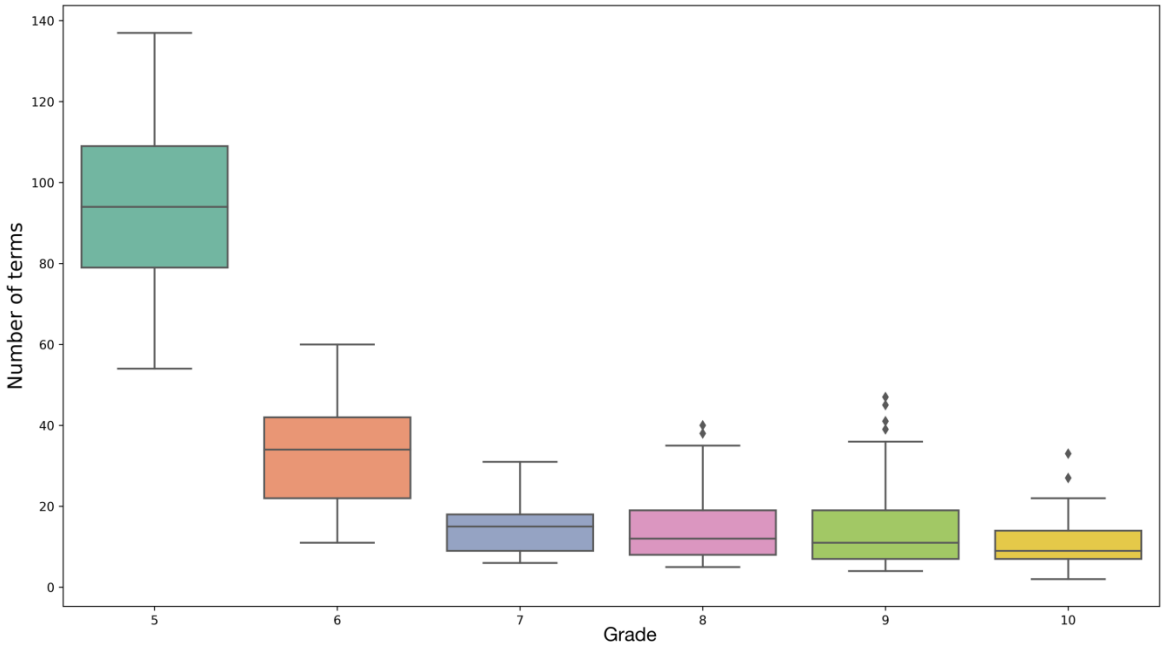


Figure 5. The dynamics of new terms entering the terminology system of the Russian Language educational and methodological complex edited by L. Verbitskaya with the distribution of terms in different parts of textbooks for different grades
Рисунок 5. Динамика пополнения терминосостава в учебно-методическом комплексе «Русский язык» под ред. Л. А. Вербицкой с указанием на терминоупотребление в разных частях учебника

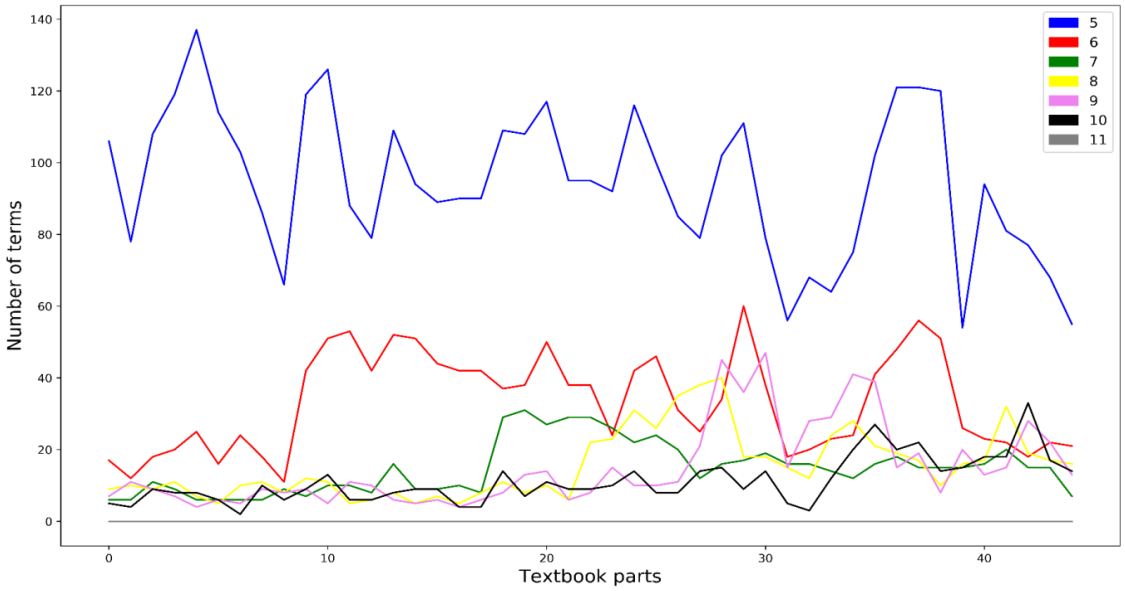
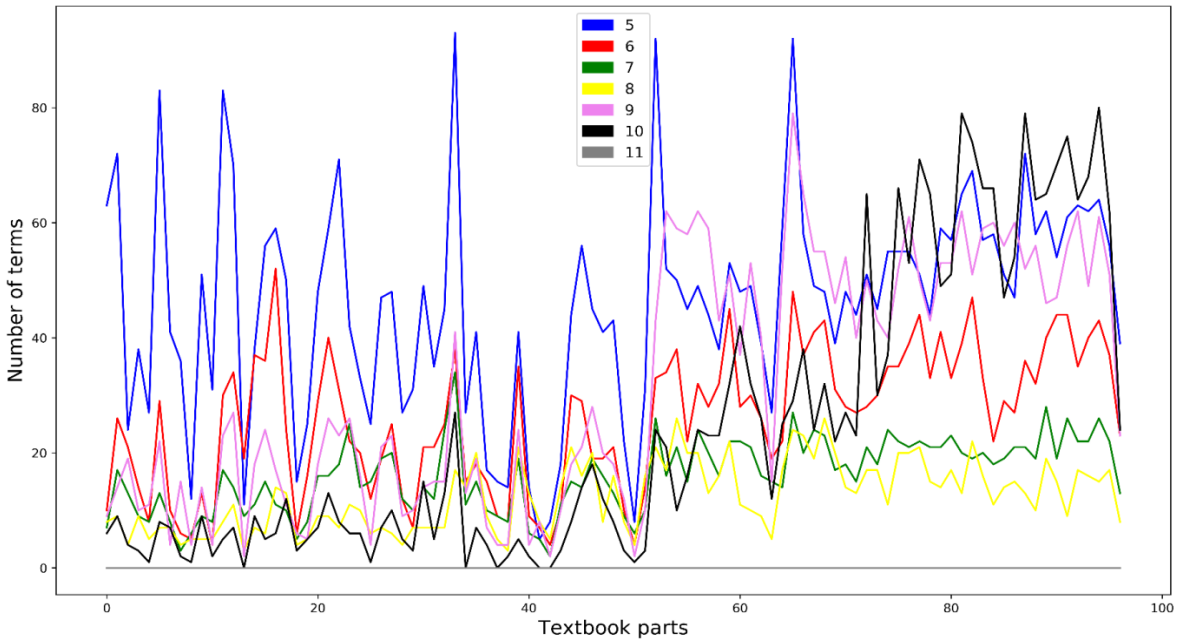


Figure 6. The dynamics of new terms entering the terminology system of the educational and methodological complex in Literature edited by V. Korovina et al. (grades 5-9) and V. Korovin et al. (grades 10-11) with the distribution of terms in different parts of textbooks for different grades
Рисунок 6. Динамика пополнения терминосостава в учебно-методическом комплексе по литературе В. Я. Коровиной и др. (5-9 классы), В. И. Коровина и др. (10–11 классы) с указанием на терминоупотребление в разных частях учебника



Grade-wise clustering of terms makes it possible to analyze how close are the terminological links between specific textbooks within one educational and methodological complex. The analysis algorithm is described below:

1) a set of terminology clusters {T5, T6, T7, T8, T9, T10-11} that include term t of a particular subject of a particular grade was identified;

2) an overlap measure Ti ∩ Tj, i ≠ j was calculated for the terms in each pair of the identified terminology clusters;

3) the identified measures were compiled into a matrix. See an example below with term *морфема* (morpheme) from the educational and methodological complex Russian Language edited by L. Verbitskaya.

Table 1. General terms share in clusters with the term *морфема* (morpheme) in the educational and methodological complex Russian Language edited by L. Verbitskaya
Таблица 1. Доля общих терминов в кластерах, содержащих термин «морфема», в учебно-методическом комплексе по русскому языку под ред. Л. А. Вербицкой

Grades	**5**	**6**	**7**	**8**	**9**	**10**
5	1.00	0.74	0.76	0.76	0.74	0.74
6	0.56	1.00	0.62	0.55	0.61	0.70
7	0.64	0.69	1.00	0.75	0.75	0.71

Grades	**5**	**6**	**7**	**8**	**9**	**10**
8	0.58	0.55	0.67	1.00	0.61	0.59
9	0.70	0.75	0.83	0.75	1.00	0.75
10	0.60	0.74	0.68	0.63	0.65	1.00

The matrix is read row-wise. The figures in cells given within the range [0,1] indicate the share of terms in a cluster that include term t from a textbook Ti and overlap with the terms in a cluster that include term t from a textbook Tj;

4) the matrix was analyzed for the maximum value, i.e., two clusters with the closest terminological links from different textbooks. In the above table this value is 0.83. This is the share of terms in the cluster that includes the term *морфема* (morpheme) in the textbook for grade 9 that overlap with the terms in the cluster that includes the term *морфема* (morpheme) in the textbook for grade 7;

5) pairs of textbooks Ti and Tj with the greatest overlap were logged in the forward_classes glossary if i < j, or in backward_classes glossary if i > j. Thus, forward_classes contained cases of cataphoric repetitions. This means that a terminology system of an earlier grade is part of an extensive terminology system of more advanced stages of training. In its turn, backward_classes contained anaphoric repetitions. They happen when a terminology system of an advanced stage of education repeats the terminology system of an earlier grade;

6) the data obtained for specific terms found in the educational and methodological complex were aggregated by pairs of grades.

An example below illustrates the hierarchy of terminological similarity in the Russian Language educational and methodological complex edited by L. Verbitskaya. The hierarchy is grade-wise. Grades are given in round brackets before the colon. Grades 10 and 11 are designated as 10. The figure after the colon indicates the number of unique terms with the maximum overlap of clusters that contain these unique terms in a given pair of textbooks. The indicators are sorted in descending order.

1) forward_classes – {(5, 6): 48, (7, 8): 46, (7, 10): 35, (5, 10): 26, (6, 10): 22, (7, 9): 21, (5, 8): 19, (8, 9): 19, (9, 10): 13, (5, 9): 9, (8, 10): 8, (5, 7): 7, (6, 9): 6, (6, 8): 4, (6, 7): 1},

2) backward_classes – {(10, 6): 62, (10, 9): 51, (9, 7): 44, (8, 6): 44, (7, 6): 36, (10, 8): 17, (10, 7): 10, (6, 5): 10, (7, 5): 8, (9, 8): 6, (8, 7): 6, (10, 5): 5, (9, 6): 5, (8, 5): 3, (9, 5): 1}.

It should be concluded that the educational and methodological complex in question has the closest thematic links in textbooks for grades 5 and 6 and grades 7 and 8. The textbooks for grades 6 and 7 are the least close thematically. As regards revision and consolidation of new knowledge, the closest thematic links are shown by the textbooks for grades 10 and 6, as well as grades 10 and 9. The least close links are observed between the textbooks for grades 9 and 5.

3.3. Using terminology in different domains

The multifaceted comparative analysis is yet another research opportunity that emerged as an outcome of target corpora vectorization (the textbook corpus and corpus of scholarly articles). Below is the discussion of just one aspect of a possible comparative study. This section compares the functioning of terms in educational and scholarly texts

with their use in non-specific and popular science contexts[1].

The comparative analysis was facilitated by the RusVectōrēs model ruwikiruscorpora-superbigrams_skipgram_300_2_2018. The model was trained with 600 mln. words of the Russian National Corpus and Wikipedia articles in December 2017 (hereinafter the RusVectōrēs model). It is the only model with all possible productive bigrams glued together, regardless of their frequency. The ability of the model to recognize bigrams is a must-have since the terms under study include both one-word and multi-word lexical units.

Next steps included:

1) development and training of the new word embedding model Word2Vec (hereinafter the Corpus model). The Corpus model was trained with the data from the single corpus of school textbooks and scholarly articles. The vector embedding size was 300. This adjustment was necessary as the original model had a size of 32 incompatible with the RusVectōrēs model;

2) selection of vector representations of all the terms from the single corpus that were found in the RusVectōrēs model glossary;

3) distribution of the selected vector representations across the four groups: distance_textbooks_wiki (vector representations of school textbooks terms in the RusVectōrēs model), distance_textbooks (vector representations of school textbooks terms in the Corpus model), distance_articles_wiki (vector representations of terms from scholarly articles in the RusVectōrēs model), distance_articles (vector representations of terms from scholarly articles in the Corpus model);

4) obtention of all possible pairwise combinations of terms assigned to one of the four groups of vector representations and calculation of cosine similarity CS = u * v / (||u|| * ||v||) for each pair, so that CS is within the range [0,1], where 1 denotes vector identity, and 0, vector orthogonality;

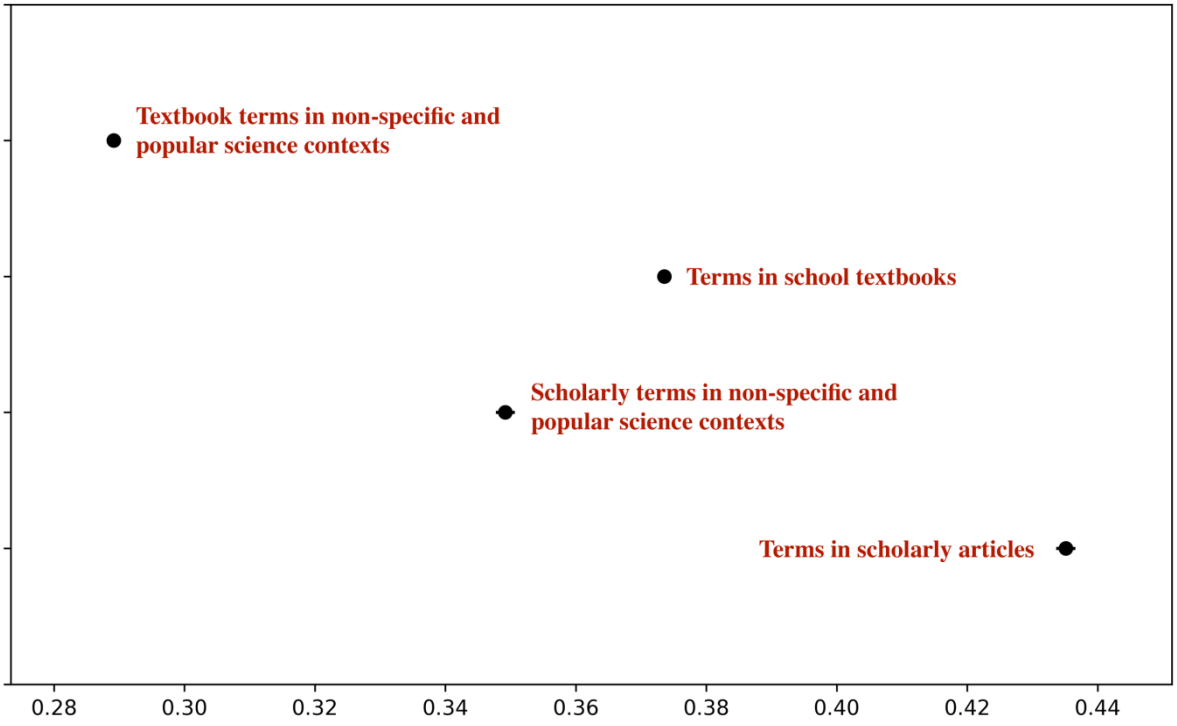
5) processing of all the data obtained for the four groups with one-way analysis of variance (ANOVA) to determine the ratio of systematic (intergroup) variance to random (intragroup) variance. The CSMi measure with i = 1,2,3,4 calculated for each of the four groups, is the arithmetic mean of all pairwise measures of cosine similarity. It gives a general understanding of how closely related the terms from a particular group are in vector space, i.e., how semantically cohesive the group of terms is;

6) analysis of statistically significant differences between the four groups of cosine similarity measures for all the domains under study. Since the analysis of variance does not show which particular groups differ from each other, it was followed by a posteriori comparisons, i.e., pairwise comparisons of the four groups with Tukey's test.

To sum up, we found four numerical indicators that describe textual behavior of: (a) terms in school textbooks, (b) terms in scholarly articles, (c) textbook terms in non-specific and popular science contexts, (d) scholarly terms in non-specific and popular science contexts. See Figure 7 for an example of results obtained for the Russian Language / Linguistics domain. As is seen from the diagram, semantic coherence of linguistic terms decreases as we move away from scholarly articles to school textbooks and, then, from textbooks and scholarly articles to non-specific and popular science contexts.

[3] Another important aspect here is the comparison of semantic maps of term use in school textbooks and scholarly articles. This aspect of comparative analysis reveals pronounced differences in term functioning in the two domains, especially as regards the choice of terms in the particular section of a subject domain. For more, refer to (Monakhov et al., 2022).

Figure 7. Comparison of cosine similarity measures as indicators of semantic coherence of terms from the Russian Language / Linguistics subject area in different domains
Рисунок 7. Сопоставление мер косинусной близости, указывающее на степень семантической спаянности терминов области знания «Русский язык / Лингвистика» в различных сферах употребления



The data analysis across different subject areas resulted in the following four groupings by type of established regularities:

0 < CSMdistance_textbooks_wiki < CSMdistance_articles_wiki < CSMdistance_textbooks < CSMdistance_articles < 1 – Art, Geography, Computer Science, Musicology, Physical Education, Russian Language / Linguistics, Social Science;

0 < CSMdistance_articles_wiki < CSMdistance_textbooks_wiki < CSMdistance_articles < CSMdistance_textbooks < 1 – Astronomy, History, Law, Literature;

0 < CSMdistance_articles < CSMdistance_textbooks < CSMdistance_articles_wiki < CSMdistance_textbooks_wiki < 1 – Biology;

0 < CSMdistance_textbooks < CSMdistance_articles < CSMdistance_textbooks_wiki < CSMdistance_articles_wiki < 1 – Chemistry, Mathematics, Physics.

It seems logical that in most domains scholarly terms of the RusVectōrēs model have higher average values of cosine similarity than schoolbook terms. This means that textbook terms have, in general, less potential for independent systematicity that scholarly terms. Another conclusion is more unexpected, while also more illustrative. To make it even more conspicuous, it is feasible to reduce the number of groupings from four to two through the merger of distance_textbooks_wiki with distance_articles_wiki and distance_textbooks with distance_articles. Thus, we get:

0 < CSMdistance_ textbooks+articles_wiki < CSMdistance_textbooks+articles < 1 – Art, Geography, Computer Science, Musicology, Physical Education, Russian Language / Linguistics, Social Science, Astronomy, History, Law, Literature;

0 < CSMdistance_textbooks+articles < CSMdistance_textbooks+articles_wiki < 1 – Biology, Chemistry, Mathematics, Physics.

Thus, we distinguish between key subject areas in exact and natural sciences (Biology, Chemistry, Mathematics, Physics) and other disciplines belonging to arts and humanities. In the RusVectōrēs model the terms from the first group have higher values of average cosine similarity than those in the second group. This indicates that terms from exact and natural sciences retain or even enhance their semantic similarity once they are used beyond scholarly articles or school textbooks. They tend to behave as a relatively cohesive semantic group. It would be fair to say that they resist the pressure of any other communication domain and harshly reject uncommon collocations limiting their functioning to recurrent contexts. In contrast, once they find themselves in a general linguistic context, terms from arts and humanities lose their similarity in contextual behavior and resemble a semantically dispersed cloud. They show more freedom in uncommon contexts and are quicker to transform into common words. These patterns found in the behavior of subject-specific terms used in uncommon contexts objectify the concept of "word familiarity". This concept, traditionally used to assess text complexity, has been often rejected as unreliable due to its subjective nature. Terms of exact and natural sciences tend to retain their semantic proofness as a group even beyond special or educational texts. This prevents their natural assimilation through semantic and communication tools. On the contrary, terms of arts and humanities are faster to become familiar, primarily due to the freedom of use that they show in diverse lexical contexts.

The cases of failure to meet the established pattern (compare, e.g., the data obtained for such subjects as Geography, Computer Science, and Astronomy vs. History, Law, and Literature) require further research that lies beyond the scope of the reported study. To comment just briefly, the reasons for numerical deviations in the overall patterns of term use may vary for different subjects. This may be due to comprehensive nature of a subject. Thus, geography combines elements of the natural sciences and humanities. Another reason is a small amount of information found in textbooks as certain subjects are only taught at a basic level, e.g., a school course in Astronomy. Finally, automatically extracted term lists for a particular subject may show considerable heterogeneity and include a big number of non-terms. In the latter case, the identified deviations from the established patterns are diagnostic in nature. They indicate the necessity of further improvement in automatic term extraction techniques and computer analysis of term functioning.

4. Conclusion

The toolkit used in the reported study to investigate the functioning of terminology is based on the principles of distributional semantics and the Word2Vec algorithms. It takes account of regular use of terms in similar lexical contexts. This creates the conditions to analyze contextual behavior of terms as elements of terminology systems based on semantically coherent groups of lexical units. This, in turn, allows to improve the results of statistical automatic term extraction from target corpora and analyze the behavior of terms in large volumes of text in different knowledge domains. The evidence for the reported study was taken from modern school textbooks. Due to the employed toolkit, we were able to compare the terminological load of textbooks in different subjects, i.e., the systemic use of groups of terms; to describe the structure of high-frequency non-terms; to explore the dynamics of new terms entering terminology systems within a set of educational and

methodological complexes, one of the courses or a specific textbook for a specific school grade; to compare through a range of metrics the regularities of term functioning in school textbooks, scholarly articles and non-specific contexts. The obtained results may be useful to experts in computer-assisted text analysis, general didactics, subject-specific teaching methodology and complexity studies.

Some of the study outcomes are in line with the established intuitive ideas about the use of subject-specific terminology in school textbooks. In some cases, however, these outcomes provide new insights. Thus, a common perception about the complexity and rigor of exact and natural sciences has been proved mathematically. Strikingly, these textual qualities are not simply due to the abundance of terminology, but, rather, due to rigid contextual and semantic coherence of terms that retains and even increases beyond the boundaries of their primary knowledge domain. On the other hand, commonly known for their more general descriptive character, textbooks in the Russian Language and Literature were found to be considerably different as they progressed from lower to upper secondary school. These differences concerned terminological load, systemic use of terms and the dynamics of new terms entering the textbooks.

Such factors as terminological load and the frequency of terms and non-terms contribute to the measure of school textbooks complexity. However, objective lexical indicators of complexity enter into complex and, at times, contradictory relations with both the measure of complexity and didactic principles of textbook efficiency. Thus, the study found lexical and thematic similarity in school textbooks for the Russian Language subject. As lexical diversity is one of the complexity-increasing factors, lexical and thematic similarity reduces both the measure of complexity and the text complexity in general. This facilitates the didactic principle of comprehensibility. At the same time, similarity undermines the motivation for learning and contradicts modern didactic principles that require educational materials to be psychologically appropriate for students in terms of their age and individual characteristics. Irregular and contextually incoherent use of terms in a range of textbooks in arts and humanities decreases their measure of complexity. This, however, contradicts the didactic principles of continuity, consistency and systematicity of learning, which, on the opposite, increases the measure of complexity. Undoubtedly, follow-up research in lexical complexity of school textbooks should include the assessment of a textbook structure as well as the structure of an educational and methodological complex in general. The reason is that conclusions about the balance between complexity and difficulty are impossible to make without accounting for the dynamics of new terms entering textbooks as well as the relationship between the new and already familiar terms.

The reported results are only part of the study outcomes. Ultimately, the study aims to develop the Russian language terminological database relevant to the content of secondary education. Python codes developed specifically for the reported study can be reproduced with any other educational and methodological complex or a term-rich text corpus. All the study-related materials and outcomes including text corpora, term lists, program codes, word embedding models, graphs and semantic maps are available in the open-access scientific repository[1].

References

Brownlee, J. (2017). *Deep Learning for Natural Language Processing: Develop Deep Learning Models for your Natural Language Problems*, Machine Learning Mastery Publ., Vermont, USA. *(In English)*

Cabré, M. T., Estopà, R. and Vivaldi, J. (2001). Automatic Term Detection: a Review of Current Systems, in Bourigault, D., Jacquemin, Ch. and L'Homme, M.-C. (eds.), *Recent Advances in Computational Terminology*,

[1] https://zenodo.org/record/4079198#.X4Mrfy1h29Y; https://zenodo.org/record/5722495#.YZ7FUS2ZPpA

John Benjamins Publ., Amsterdam, Netherlands, 53–87. DOI: 10.1075/nlp.2.04cab *(In English)*

Durda, K. and Buchanan, L. (2008). WINDSORS: Windsor Improved Norms of Distance and Similarity of Representations of Semantics, *Behavior Research Methods*, 40, 705–712. DOI: 10.3758/BRM.40.3.705 *(In English)*

Fisher, D., Frey, N. and Lapp, D. (2016). *Text Complexity: Stretching Readers with Texts and Tasks*, Corwin Press, Thousand Oaks, CA, USA. *(In English)*

Flor, M., Klebanov, B. and Sheehan, K. (2013). Lexical Tightness and Text Complexity, *Proceedings of the 2th Workshop of Natural Language Processing for Improving Textual Accessibility (NLP4ITA)*, Atlanta, USA, 29–38. *(In English)*

Glazkova, A., Egorov, Yu. and Glazkov, M. (2021). A Comparative Study of Feature Types for Age-Based Text Classification, in van der Aalst, W. et al. (eds.), *Analysis of Images, Social Networks and Texts. AIST 2020. Lecture Notes in Computer Science, 12602*, Springer Publ., Cham, Switzerland, 120–134. *(In English)*

Iomdin, B. L. and Morozov, D. A. (2021). Who Can Understand "Dunno"? Automatic Assessment of Text Complexity in Children's Literature, *Russkaya Rech'*, 5, 55–68. DOI: 10.31857/S013161170017239-1 *(In Russian)*

Jones, M. N. and Mewhort, D. J. K. (2007). Representing Word Meaning and Order Information in a Composite Holographic Lexicon, *Psychological Review*, 114, 1–37. DOI: 10.1037/0033-295X.114.1.1 *(In English)*

Kilgarriff, A., Jakubíček, M., Kovář, V. et al. (2014). Finding Terms in Corpora for Many Languages with the Sketch Engine, *Proceedings of the Demonstrations at the 14th Conference the European Chapter of the Association for Computational Linguistics*, Gothenburg, Sweden, 53–56. DOI: 10.3115/v1/E14-2014 *(In English)*

Korkontzelos, I. and Ananiadou, S. (2014). Term Extraction, in Mitkov, R. (ed.), *Oxford Handbook of Computational Linguistics*, Oxford University Press, Oxford, UK, 991–1012. *(In English)*

Kutuzov, A. and Kuzmenko, E. (2017). WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models, in Ignatov, D. et al. (ed.), *Analysis of Images, Social Networks and Texts. AIST 2016. Communications in Computer and Information Science*, *661*, Springer Publ., Cham, Switzerland, 155–161. *(In English)*

Laposhina, A. N., Lebedeva, M. U. and Berlin Khenis, A. (2022). Word Frequency and Text Complexity: An Eye-tracking Study of Young Russian Readers, *Russian Journal of Linguistics*, 26 (2), 493–514. DOI: 10.22363/2687-0088-30084. *(In Russian)*

Laposhina, A. N., Veselovskaya, T. S., Lebedeva, M. U. and Kupreshchenko, O. F. (2019). Lexical Analysis of the Russian Language Textbooks for Primary School: Corpus Study, *Computational Linguistics and Intellectual Technologies: papers from the Annual International Conference "Dialogue"*, Moscow, Russia, 18 (25), 351–363. *(In Russian)*

Leichik, V. M. (2007). *Terminovedenie: predmet, metody, struktura* [Terminology Studies: Subject, Methods, Structure], LKI Publishing House, Moscow, Russia. *(In Russian)*

Levy, O. and Goldberg, Y. (2014). Linguistic Regularities in Sparse and Explicit Word Representations, *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, Baltimore, USA, 171–180. DOI: 10.3115/ v1/W14-1618 *(In English)*

Lukashevich, N. V. and Logachev, Yu. M. (2010). Combining Features for Automatic Term Extraction, *Numerical Methods and Programming*, 11 (4), 108–116. *(In Russian)*

Martynova, E. V., Solnyshkina, M. I., Merzlyakova, A. F. and Gizatulina, D. Yu. (2020). Lexical Parameters of the Academic Text (Based on the Texts of the Academic Corpus of the Russian Language), *Philology and Culture*, 3, 72–80. DOI: 10.26907/2074-0239-2020-61-3-72-80 *(In Russian)*

Mikk, Ya. A. (1981). *Optimizatsiya slozhnosti uchebnogo teksta: V pomoshch' avtoram i redaktoram* [Optimizing the complexity of educational text: To help authors and editors], Prosveshchenie, Moscow, Russia. *(In Russian)*

Mikolov, T., Sutskever, I., Chen, K. et al. (2013a). Distributed Representations of Words and Phrases and their Compositionality, *Advances in Neural Information Processing Systems 26, 27th Annual Conference on Neural Information Processing Systems 2013*, Lake Tahoe, USA, 3136–3144. *(In English)*

Mikolov, T., Yih, W. T and Zweig, G. (2013b). Linguistic Regularities in Continuous Space Word Representations, *Proceedings of the 2013 Conference of the North American Chapter*

of the Association for Computational Linguistics: Human Language Technologies, Atlanta, USA, 746–751. *(In English)*

Mitrofanova, O. A. and Zakharov, V. P. (2009). Automatic Analysis of Terminology in the Russian Text Corpus on Corpus Linguistics, *Computational Linguistics and Intellectual Technologies: papers from the Annual International Conference "Dialogue"*, Bekasosvo, Russia, 8 (15), 321–328. *(In Russian)*

Monakhov, S. I., Turchanenko, V. V. and Cherdakov, D. N. (2022). Terminology in Textbooks and Research Articles: Cluster Analysis of Corpus Data, *Proceedings of 6th International Conference "Informatization of Education and E-learning Methodology: Digital Technologies in Education",* Krasnoyarsk, Russia, 3, 228–233. *(In Russian)*

Morozov, D. A. and Iomdin, B. L. (2019). Criteria of Semantic Complexity of Words, *Computational Linguistics and Intellectual Technologies: papers from the Annual International Conference "Dialogue"*, Moscow, Russia, 18 (25), 119–131. *(In Russian)*

Nokel, M. A., Bolshakova, E. I. and Loukachevitch, N. V. (2012). Combining Multiple Features for Single-word Term Extraction, *Computational Linguistics and Intellectual Technologies: papers from the Annual International Conference "Dialogue"*, Bekasosvo, Russia, 11 (18), 1, 490–501. *(In English)*

Piotrovsky, R. G. and Yastrebova, S. V. (1969). Statistical Term Recognition, in Piotrovskij, R. G. (ed.), *Statistika teksta* [Text statistics], Belorusskij gosudarstvennyj universitet, Minsk, Belarus, 1, 249–259. *(In Russian)*

Rohde, D. L., Gonnerman, L. M. and Plaut, D. C. (2006). An Improved Model of Semantic Similarity Based on Lexical Co-Occurrence, *Communications of the ACM*, 8, 627–633. *(In English)*

Schwanenflugel, P. J. (1991). Why are Abstract Concepts Hard to Understand?, in Schwanenflugel, P. J. (ed.), *The psychology of word meanings*, Lawrence Erlbaum Associates Inc., Hillsdale, USA, 223–250. *(In English)*

Sharoff, S. (2022). What Neural Networks Know about Linguistic Complexity, *Russian Journal of Linguistics*, 26 (2), 371–390. DOI: 10.22363/2687-0088-30178 *(In English)*

Shpakovsky, Yu. F. (2007). Estimation of Perception Difficulty and Optimization of the Educational Text Complexity (on the Material of Texts in Chemistry), Abstract of Ph.D. dissertation, Linguistics, Minsk State Linguistic University, Minsk, Belarus. *(In Russian)*

Solnyshkina, M. I. (2022). Measuring Text Complexity: State of the Art, *Collection of Scientific Papers X Jubilee International Scientific Conference "Teacher. Student. Textbook (in the Context of Global Challenges of Modern Times)"*, Moscow, Russia, 20–24. *(In Russian)*

Solnyshkina, M. I. and Kiselnikov, A. S. (2015). Text Complexity: Study Phases in Russian Linguistics, *Tomsk State University Journal of Philology*, 6 (38), 86–99. DOI: 10.17223/19986645/38/7 *(In Russian)*

Solnyshkina, M. I., McNamara, D. and Zamaletdinov, R. R. (2022). Natural Language Processing and Discourse Complexity Studies, *Russian Journal of Linguistics*, 26 (2), 317–341. DOI: 10.22363/2687-0088-30171 *(In Russian)*

Solovyev, V. D., Ivanov, V. V. and Solnyshkina, M. I. (2018). Assessment of Reading Difficulty Levels in Russian Academic Texts: Approaches and Metrics, *Journal of Intelligent & Fuzzy Systems*, 34 (2), 3049–3058. DOI: 10.3233/JIFS-169489 *(In English)*

Solovyev, V. D., Solnyshkina, M. I. and McNamara, D. (2022). Computational Linguistics and Discourse Complexology: Paradigms and Research Methods, *Russian Journal of Linguistics*, 26 (2), 275–316. DOI: 10.22363/2687-0088-30161 *(In English)*

Stepanova, D. V. (2017). Analiz metodov avtomaticheskogo vydeleniya terminov iz nauchno-tekhnicheskih tekstov [Analysis of Methods for Automatic Terms Extraction from Scientific and Technical Texts], *Aktual'nye problemy sovremennoj prikladnoj lingvistiki* [Current problems of modern applied linguistics], Minskij gosudarstvennyj lingvisticheskij universitet, Minsk, 62–67. *(In Russian)*

Tatarinov, V. A. (2006). *Obshchee terminovedenie: Entsiklopedicheskij slovar'* [Terminology Studies: Encyclopedic Dictionary], Moskovskij Litsej, Moscow, Russia. *(In Russian)*

Turney, P. D. and Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics, *Journal of Artificial Intelligence Research*, 37, 141–188. DOI: 10.1613/jair.2934 *(In English)*

Все авторы прочитали и одобрили окончательный вариант рукописи.
All authors have read and approved the final manuscript.

Конфликты интересов: у авторов нет конфликтов интересов для декларации.
Conflicts of interests: the authors have no conflicts of interest to declare.

Sergei I. Monakhov, Ph.D. in Philology, Research Associate, Friedrich Schiller University Jena, Germany.
Сергей Игоревич Монахов, кандидат филологических наук, научный сотрудник Йенского университета им. Ф. Шиллера, Германия.

Vladimir V. Turchanenko, Junior Researcher, Institute of Russian Literature (Pushkinsky Dom) of the Russian Academy of Sciences, Saint Petersburg, Russia.
Владимир Владимирович Турчаненко, младший научный сотрудник Института русской литературы (Пушкинский Дом) РАН, Россия.

Dmitrii N. Cherdakov, Senior Lecturer, Saint Petersburg University, Russia.
Дмитрий Наилевич Чердаков, старший преподаватель Санкт-Петербургского государственного университета, Россия.

УДК 81`322.2 УДК 373 **DOI: 10.18413/2313-8912-2023-9-1-0-3**

Монахов С. И.[1]
Турчаненко В. В.[2]
Чердаков Д. Н.[3]

Школьный учебный текст в аспекте терминоупотребления: корпусный анализ

[1] Йенский университет им. Ф. Шиллера
Фюрстенграбен, 1, Йена, 07743, Германия
E-mail: sergomon@gmail.com

[2] Институт русской литературы (Пушкинский Дом) РАН
наб. Макарова, 4, Санкт-Петербург, 199034, Россия
E-mail: vladimir.turchanenko@mail.ru

[3] Санкт-Петербургский государственный университет
Университетская наб., 7–9, Санкт-Петербург, 199034, Россия
E-mail: dm.cherdakov@gmail.com

Статья поступила 23 января 2023 г.; принята 13 марта 2023 г.; опубликована 30 марта 2023 г.

Информация об источниках финансирования или грантах: Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 19-29-14032 мк «Изучение терминологических подсистем современных школьных учебников на русском языке с помощью моделей анализа семантики естественных языков Word2Vec и нейронных сетей».

Аннотация. В статье излагаются методы и результаты анализа употребления терминологической лексики в современных школьных учебниках на русском языке. Основным материалом исследования является созданный исследовательский корпус, включающий тексты 207 учебников с 5-го по 11-й класс по 21 школьной дисциплине. Традиционный способ автоматического извлечения терминов, основанный на статистических показателях частотности словоупотребления, предлагается усовершенствовать с помощью создания моделей, обученных по алгоритмам Word2Vec, в основе которых лежат идеи дистрибутивной семантики. Применение этих алгоритмов, выражающее в числовом представлении сочетаемостное поведение слов и соответственно степень их семантической близости, позволило: в существенной мере устрожить результаты автоматического выделения терминов, отграничивая от них высокочастотные нетерминологические единицы; осуществить сопоставительную характеристику состава и употребления терминов в учебниках по разным предметам и разных ступеней обучения; проанализировать динамику пополнения терминологических систем внутри учебно-методических комплексов и охарактеризовать терминологические взаимосвязи между учебниками для отдельных классов. При помощи специально созданного корпуса научных статей по тем дисциплинам, которые соответствуют предметам школьного обучения, были выявлены различия в употреблении терминов в школьной и научной сферах, а также (с использованием дистрибутивно-семантической модели, предоставляемой

ресурсом RusVectōrēs) в сфере общеупотребительной и научно-популярной речи. Для каждого из отмеченных аспектов анализа обнаружены значимые признаки в функционировании терминов, свойственные отдельным школьным дисциплинам или их группам. Полученные результаты оценивались в том числе в свете положений теории сложности текста и принципов дидактики и методики. Отмечены, в частности, случаи противоречия между показателями сложности текста и его предполагаемой трудности, а также неоднозначный характер взаимодействия меры сложности текста с ключевыми дидактическими началами.

Ключевые слова: Термин; Терминология; Школьный учебник; Сложность текста; Частотность слова; Векторное представление; Word2Vec; Нейронная сеть

Информация для цитирования: Монахов С. И., Турчаненко В. В., Чердаков Д. Н. Школьный учебный текст в аспекте терминоупотребления: корпусный анализ // Научный результат. Вопросы теоретической и прикладной лингвистики. 2023. Т. 9. № 1. С. 27-49. DOI: 10.18413/2313-8912-2023-9-1-0-3

UDC 81`322.2 UDC 373 **DOI: 10.18413/2313-8912-2023-9-1-0-3**

Sergei I. Monakhov[1]
Vladimir V. Turchanenko[2]
Dmitrii N. Cherdakov[3]

Terminology use in school textbooks: corpus analysis

[1] Friedrich Schiller University Jena
1 Fuerstengraben, Jena, 07743, Germany
E-mail: sergomon@gmail.com

[2] Institute of Russian Literature (Pushkinsky Dom) of the Russian Academy of Sciences
4 Makarov Emb., Saint Petersburg, 199034, Russia
E-mail: vladimir.turchanenko@mail.ru

[3] St Petersburg University
7-9 Universitetskaya Emb., Saint Petersburg, 199034, Russia
E-mail: dm.cherdakov@gmail.com

Received 23 January 2023; accepted 13 March 2023; published 30 March 2023

Acknowledgements. The reported study was funded by the Russian Foundation for Basic Research, Project number 19-29-14032 mk “Study of terminological subsystems of modern school textbooks in Russian with the help of word embedding models Word2Vec and neural networks”.

Abstract. The article presents the methods and results of the study that investigated the use of terminology in textbooks for secondary schools in Russia. The data were taken from a full-text DIY corpus of 207 textbooks for grades 5-11. The toolkit included models trained with the Word2Vec algorithms driven by the ideas of distributional semantics. The models were used to improve traditional automatic term extraction based on word frequency statistics. Numerical representation of word collocation patterns and their semantic similarity enabled the following: more effective automatic term extraction with a clear dividing line between terminology *per se* and high-frequency common words; comparative analysis of inventory and

functioning of terms in textbooks for different school subjects and grades; analysis of the dynamics of new terms entering educational and methodological complexes and insights into terminological relations between textbooks for different grades. The study included another DIY corpus compiled of scholarly articles across the subjects taught at school. It was used to identify differences in term use in textbooks and scholarly texts as well as in non-specific and popular science contexts. The latter was facilitated by the RusVectōrēs word embedding model. The comprehensive analysis identified some patterns in term functioning relevant for particular school subjects or groups of subjects. The results were evaluated in view of the theory of text complexity, teaching methodology and didactics. The study found some contradictions between the expected and real text complexity. It also showed certain discrepancy between text complexity and basic didactic principles.
Keywords: Term; Terminology; School textbook; Text complexity; Word frequency; Vector representation; Word2Vec; Neural network
How to cite: Monakhov, S. I., Turchanenko, V. V. and Cherdakov, D. N. (2023). Terminology use in school textbooks: corpus analysis, *Research Result. Theoretical and Applied Linguistics*, 9 (1), 27-49. DOI: 10.18413/2313-8912-2023-9-1-0-3

1. Вступление

В богатой истории российского терминоведения изучение функционирования терминов в школьной учебной литературе занимает незначительное место. Достаточно отметить, что в фундаментальном энциклопедическом словаре В. А. Татаринова (Татаринов, 2006), охватывающем все достижения советской и российской науки о терминах, о специфике школьного терминоупотребления не сказано вообще ничего. Между тем терминология в школьных учебниках, измеряемая тысячами единиц и представляющая своеобразную проекцию системы научных понятий, безусловно, заслуживает внимания и в теоретическом терминоведческом аспекте, и в дидактическом плане, и в социокультурном отношении. Цель исследования, некоторые результаты которого изложены ниже, состоит в том, чтобы в известной мере восполнить указанную лакуну, используя современные методы корпусной и компьютерной лингвистики, позволяющие автоматически обрабатывать значительные массивы информации.

Известны различные методики автоматического извлечения терминов из корпусов большого объема (Korkontzelos, Ananiadou, 2014; Степанова, 2017), однако в большинстве случаев для этой цели используется статистический подход, основы которого были намечены еще в 1960-х гг. (Пиотровский, Ястребова, 1969). Этот подход зиждется на установленной закономерности, согласно которой частотность терминологических единиц в специальных текстах существенно выше, чем в текстах общеупотребительной сферы, и алгоритмически представляет собой сопоставление частотности слов в целевом корпусе, откуда необходимо извлечь термины, и референтном корпусе, который, как правило, представляет совокупность неспециальных текстов на данном языке (Kilgarriff et al., 2014). Статистический подход к автоматическому выделению терминов, реализованный в чистом виде, обычно дает не вполне надежные результаты (Cabré et al., 2001), поэтому исследователи ищут различные способы его комбинации с иными методиками (Митрофанова, Захаров, 2009; Лукашевич, Логачев, 2010; Nokel, 2012). На наш взгляд, существенное улучшение результатов автоматического выделения терминов возможно в случае комбинации статистического подхода с алгоритмами Word2Vec (continuous-bag-of-words – CBOW, skip-gram), которые реализуют основную идею дистрибутивной семантики – значение слова может быть выведено из его лексического окружения и математически определено как сумма контекстов, в которых это слово встречается (Rohde et al., 2006; Jones, Mewhort, 2007; Durda, Buchananan, 2008; Turney, Pantel, 2010). Дистрибутивно-семантические модели, обученные по алгоритмам Word2Vec и

использующие для оценки смысловой близости слов их векторные представления (Mikolov et al., 2013a; Mikolov et al., 2013b; Levy, Goldberg, 2014; Brownlee, 2017), широко используются в последнее десятилетие, однако, насколько нам известно, они еще не применялись для изучения терминологического состава учебных текстов. Важнейшее преимущество этого метода перед остальными мы видим в том, что с его помощью можно наблюдать за поведением в тексте семантически связанных терминологических групп и тем самым характеризовать терминоупотребление в аспекте ключевого свойства термина – его принадлежности некоторой терминосистеме (Лейчик, 2007: 98–129).

Результаты стратификации терминологического наполнения школьных учебников, осуществленной методами корпусной и компьютерной лингвистики, могут быть небезынтересны и в аспекте теории сложности текста. Современный этап развития комплексологии отмечен все более широким использованием методов автоматической обработки и анализа текстов; обзоры см. в: (Соловьев и др., 2022; Солнышкина и др., 2022); примеры исследований: (Flor et al, 2013; Иомдин, Морозов, 2021; Glazkova et al., 2021; Sharoff, 2022). Эти методы применяются и для оценки сложности учебных текстов разных типов, которые традиционно являются одним из основных предметов интереса специалистов по комплексологии; см., например, исследования сложности школьных учебных текстов на материале корпусных данных: (Solovyev et al., 2018; Мартынова и др., 2020).

Употребление терминов считается одним из лексических показателей сложности текста (Шпаковский, 2007). При этом природа терминологических единиц такова, что на увеличение меры сложности текста они влияют в силу целого ряда своих характеристик. Термины – это, как правило, малочастотные за пределами специальных текстов слова; традиционно считается, что высокий процент низкочастотных слов увеличивает сложность текста, что в целом подтверждается и исследованиями с применением новейших технологий (Лапошина и др., 2022). Термины, даже референциально отсылая к конкретным объектам действительности, тяготеют к понятийной абстракции (Татаринов, 2006: 231–234) и тем самым увеличивают степень абстрактности текста, следовательно, и меру его сложности (Schwanenflugel, 1991; Fisher et al., 2016). Наконец термины, как правило, считаются семантически сложными словами, степень «знакомства» с которыми у неспециалиста невелика (Микк, 1981: 65). Семантическая сложность слова плохо поддается формальному измерению (Морозов, Иомдин, 2019), однако следует учитывать, что в учебном тексте термины – это именно те слова, значение которых требуется усвоить, что в данном случае оправдывает их характеристику как «незнакомых» слов.

Применительно к учебной книге или учебно-методическому комплексу важно учитывать закрепившееся в комплексологии разграничение абсолютной сложности текста как суммы его объективных характеристик и трудности текста (иначе – относительной сложности текста) как его качества, зависимого от внешних факторов, в частности от познавательных возможностей воспринимающего текст субъекта (Солнышкина, Кисельников, 2015: 86–87; Солнышкина, 2022: 20). Высокая частотность терминов увеличивает меру сложности текста в целом, но их регулярная повторяемость в тексте учебника постепенно снижает его трудность (Микк, 1981: 67). Один и тот же текстовый фрагмент, содержащий подлежащие усвоению термины и размещенный одновременно в начале и в конце учебника, обладает одинаковой мерой сложности, но, очевидно, должен предположительно оцениваться как соответственно более и менее трудный, ибо сумма знаний ученика и степень его знакомства с терминологическими единицами должны были увеличиться в ходе обучения.

В аспекте дидактики сложность и трудность учебного текста преломляются в дихотомии дидактических принципов научности и доступности: содержание школьного учебника должно так или иначе соответствовать современным научным представлениям, отражая в том числе и их терминологическую составляющую (и в этом плане оно неизбежно трудно), но при этом

оно должно быть посильным для усвоения. Отклонение от реализации этих принципов в любую сторону снижает дидактическую эффективность учебника.

2. Материал и методы

2.1. Создание целевых корпусов

В основе проведенного исследования лежит ряд процедур, в частности создание исследовательских (иначе – целевых) корпусов, векторизация корпусов, кластеризация полученных данных.

Первоочередной задачей было создание целевого корпуса текстов школьных учебников. Для создания корпуса были использованы тексты 207 учебников с 5-го по 11-й класс, выпущенных издательством «Просвещение» (от издательства было получено официальное разрешение использовать тексты в исследовательских целях). Все учебники на момент включения их текстов в корпус (2020 г.) входили в формируемый Министерством просвещения федеральный перечень изданий, рекомендованных к использованию в школах. Тексты сканировались, распознавались, проходили обработку (удалялись небуквенные символы, знаки препинания, унифицировался регистр букв и др.); с помощью программных средств словоформы приводились к начальной форме, осуществлялась частеречная разметка слов. В конечном счете общий объем корпуса составил чуть более 13 965 000 словоупотреблений. Корпус был разбит на подкорпусы в соответствии со школьными дисциплинами (всего – 21): алгебра (18 учебников; всего 1 144 089 словоупотреблений), астрономия (2; 89 574), биология (21; 1 125 648), всеобщая история и история России (15; 973 498), география (8; 512 173), геометрия (8; 370 054), естествознание (2; 158 665), изобразительное искусство (8; 283 608), информатика (6; 284 683), литература (18; 3 939 054), математика (10; 525 035), математический анализ (14; 1 134 786), мировая художественная культура (2; 33 130), музыка (4; 76 241), обществознание (12; 505 822), право (2; 171 349), русский язык (18; 1 131 575), технология (4; 163 574), физика (15; 1 098 625), физическая культура (7; 301 371), химия (13; 543 283). Кроме того, каждый предметный подкорпус был разделен на составные части в соответствии с годами обучения.

Еще одним специально созданным исследовательским корпусом, необходимым для изучения школьного терминоупотребления в сопоставительном аспекте (см. ниже – п. 3.2), стал корпус современных научных статей на русском языке. Он формировался согласно следующим принципам. Из наиболее цитируемых, при этом не узкоспециальных научных журналов различных областей знания извлекались статьи, опубликованные в 2016–2021 гг. Для каждой области знания отбирались от двух до пяти журналов, при этом доля статей из каждого журнала определялась мерой его цитируемости. В корпус включались тексты научных статей и научных сообщений из основных разделов журналов; учитывались основой текст статьи, ее название и аннотация. Например, из 100 с лишним журналов географической тематики, индексируемых в Российском индексе научного цитирования, было отобрано три не узкоспециальных журнала с наибольшим количеством цитирований – «География и природные ресурсы», «Вестник Московского университета. Серия 5: География» и «Известия Российской академии наук. Серия географическая». Поскольку эти журналы имеют приблизительно одинаковый индекс средней цитируемости (8,51, 9,72 и 9,00 соответственно), было принято решение использовать для формирования корпуса все три журнала примерно в равном соотношении: в корпус вошли тексты соответственно 50, 46 и 40 статей (всего 136 статей). В соответствии с областями знания корпус научных статей был разделен на подкорпусы, которые, за отдельными исключениями, соответствуют подкорпусам корпуса школьных учебников (в корпусе научных статей не выделялись школьные подкорпусы «Естествознание» и «Технология»; школьным подкорпусам «Математика», «Алгебра», «Геометрия», «Математический анализ» соответствовал единый подкорпус научных статей «Математика», точно так же школьным подкорпусам «Изобразительное искусство» и «Мировая художественная культура» соответствовал единый подкорпус научных статей «Искусство»). Тексты проходили необходимую обработку, аналогичную тем

операциям, которые осуществлялись в отношении текстов школьных учебников. Объем каждого подкорпуса в корпусе научных статей составил не менее 75 % от объема аналогичного подкорпуса корпуса школьных текстов, например: география – (а) в подкорпусе корпуса научных статей около 434 000 словоупотреблений, (б) в соответствующем подкорпусе корпуса школьных учебников около 512 000 словоупотреблений; биология – (а) около 853 000 словоупотреблений и (б) около 1 126 000 словоупотреблений; история – (а) около 902 000 словоупотреблений и (б) около 973 000 словоупотреблений. Общий объем корпуса научных статей составил около 10 795 500 словоупотреблений.

Подготовленные корпусы были загружены на платформу Sketch Engine (https://www.sketchengine.eu), на которой осуществлялось автоматическое извлечение кандидатов в термины согласно охарактеризованной выше процедуре сопоставления относительной частотности слов в целевом и референтном корпусах. В качестве референтного корпуса выступал Russian Web 2011 Sample (ruTenTen11), доступный в Sketch Engine и содержащий более 900 миллионов словоупотреблений из русскоязычных интернет-текстов.

2.2. Автоматическое извлечение терминов и последующая векторизация данных

Алгоритмы извлечения однословных и неоднословных кандидатов в термины различались. Для каждой однословной лексической единицы, употребленной в соответствующем подкорпусе не менее трех раз, высчитывалась метрика keyness score по формуле: ((Lt * 1,000,000 / Ct) +1) / ((Lr * 1,000,000 / Cr) +1), где Lt – частота употребления единицы в целевом корпусе, Ct – общее количество токенов в целевом корпусе, Lr – частота употребления единицы в референтном корпусе, Cr – общее количество токенов в референтном корпусе. Однословная единица получала статус кандидата в термины, если значение метрики keyness score превышало 1; ср., например, значения метрики keyness score для слов в корпусе школьных учебников в подкорпусе «Алгебра» (7–9 классы): «многочлен» – 743.2, «множитель» – 380.4, «парабола» – 322.3; в подкорпусе «Русский язык (5 класс): «существительное» – 479.4, «падеж» – 231.4, «антоним» – 170.4; в подкорпусе «Биология» (9 класс): «фотосинтез» – 562.3, «фенотип» – 166.1, «цитоплазма» – 11.2; в корпусе научных статей в подкорпусе «Химия»: «макромолекула» – 306.5, «адсорбция» – 103.042, «полимеризация» – 67.9; в подкорпусе «Астрономия»: «полуось» – 197.4, «галактика» – 94.6, «цефеиды» – 31.3.

Вычленение неоднословных кандидатов в термины происходило в два этапа. Первоначально из всех возможных сочетаний лексем, встречающихся в соответствующем подкорпусе не менее трех раз, были выделены сочетания, характеризующиеся положительным значением метрики Log-Dice score, рассчитываемой по следующей формуле: 14 + log(2(|X∩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. Затем для выделенных сочетаний была рассчитана метрика keyness score согласно приведенной выше формуле. Ср., например, значения метрики keyness score для коллокаций в корпусе школьных учебников в подкорпусе «Алгебра» (7–9 классы): «график функции» – 725.9, «натуральное число» – 200.9, «линейная функция» – 97.6; в подкорпусе «Русский язык» (5 класс): «часть речи» – 428.2, «единственное число» – 222.4, «прошедшее время» – 76.1; в подкорпусе «Биология» (9 класс): «бесполое размножение» – 240.4, «пищевая цепь» – 190.9, «генная инженерия» – 67.0; в корпусе научных статей в подкорпусе «Химия»: «элементный анализ» – 145.0, «реакционная масса» – 75.4, «буферный раствор» – 65.7; в подкорпусе «Астрономия»: «красное смещение» – 120.4, «дыра Локмана» – 81.0, «солнечный ветер» – 51.2.

Полученные списки однословных и неоднословных кандидатов в термины были упорядочены по убыванию значения метрики keyness score; для дальнейшей обработки использовались первые 1000 единиц, извлеченных из корпуса школьных учебников, и первые 2000 единиц, извлеченных из корпуса научных статей.

Одной из основных проблем при применении вышеописанной процедуры

является отграничение в кругу автоматически выделенных слов и сочетаний собственно терминологической лексики от нетерминологических лексических единиц, уподобленных терминам по поведению в текстах учебников, то есть низкочастотных в референтном корпусе, но высокочастотных в целевом корпусе. Ниже такие слова будут условно называться лжетерминами. Подчеркнем условность этого наименования, ибо при автоматическом отграничении терминов от лжетерминов в состав последних по разным причинам могут попасть и собственно термины.

Для оптимизации полученных результатов использовались алгоритмы Word2Vec, на основе которых была проведена векторизация целевых корпусов и созданы и обучены дистрибутивно-семантические модели (word embedding models), которые позволяли определить меру сходства в синтагматических характеристиках автоматически выделенных единиц в каждом из созданных подкорпусов. Для каждого подкорпуса было получено две модели – для однословных единиц и для неоднословных единиц (биграмм и триграмм). Обучение моделей происходило в следующем порядке: 1) определялась частотность каждого слова в корпусе; 2) слова сортировались по частоте, редкие слова удалялись; 3) для снижения вычислительной сложности алгоритма словарь кодировался с помощью дерева Хаффмана (Huffman Binary Tree); 4) для каждого слова в корпусе строился вектор, элементы которого представляют собой обозначения количества случаев, когда данное слово оказывается в одном окне контекстов с другими наиболее частотными словами данного корпуса (параметр окна́ контекстов – заданная максимальная дистанция между текущим и предсказываемым словом в предложении); 5) построенные векторы подавались на вход нейросети прямого распространения (feedforward neural network), которая обучается предсказывать либо контекст по заданному слову, либо слово по заданному контексту.

Векторное представление позволяет оценивать степень семантической близости каждой пары слов на основе косинусной меры их векторов. Для каждой парной комбинации была высчитана мера косинусной близости CS = u * v / (||u|| * ||v||), так что CS находится в пределах [0,1], где 1 обозначает идентичность векторов (что указывает на идентичность контекстов, в которых встречаются слова, а следовательно, на их предельную семантическую близость), а 0 – на их ортогональность (то есть отсутствие общих контекстов, а значит и общих сем). Ср., например, показатели косинусной близости в школьном подкорпусе «Русский язык»: для пары слов «суффикс» и «окончание» – 0.80, для пары слов «суффикс» и «груша» – 0.18.

Результаты векторизации использовались для усовершенствования итогов автоматического выделения терминов двумя разными способами – в зависимости от того, к школьному или научному целевому корпусу они применялись. Это вызвано разным характером текстов, составивших соответствующий целевой корпус: в случае с корпусом научных статей следует предполагать более однородный лексический состав текстов и их относительно сходную композицию.

В отношении единиц, автоматически выделенных из школьного корпуса, был избран способ кластеризации семантических карт. С помощью алгоритма стохастического вложения соседей с t-распределением (t-SNE) были построены соответствующие подкорпусам карты взаимного расположения терминологических кандидатов в обученных дистрибутивно-семантических моделях; из векторного пространства высокой размерности эти карты с целью визуализации полученных результатов были спроецированы в двухмерную плоскость. См. примеры подобных карт на рис. 1 и 2.

Допускалась высокая вероятность того, что собственно термины образуют на картах черно́ты – кластеры, объединенные небольшой косинусной дистанцией; лжетермины, напротив, предположительно рассеяны на остальном пространстве карты. С помощью метода вычисления k-средних (k-means) была осуществлена кластеризация точек на плоскости по их координатам и маркировка каждого из полученных кластеров как содержащего термины или как содержащего лжетермины. В каждой семантической карте выделялось 20 кластеров, по которым распределялись все имеющиеся точки.

Рисунок 1. Семантическая карта распределения кандидатов в термины. Подкорпус «Алгебра» (9 класс)
Figure 1. Semantic map representing the distribution of term candidates. Subcorpus Algebra (grade 9)

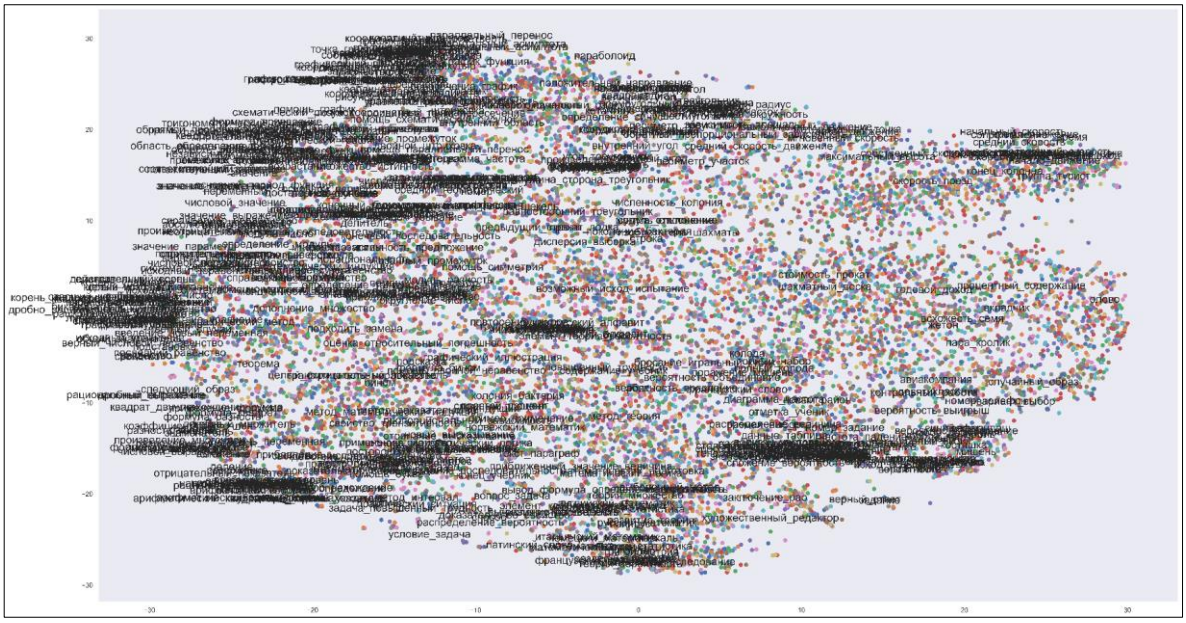
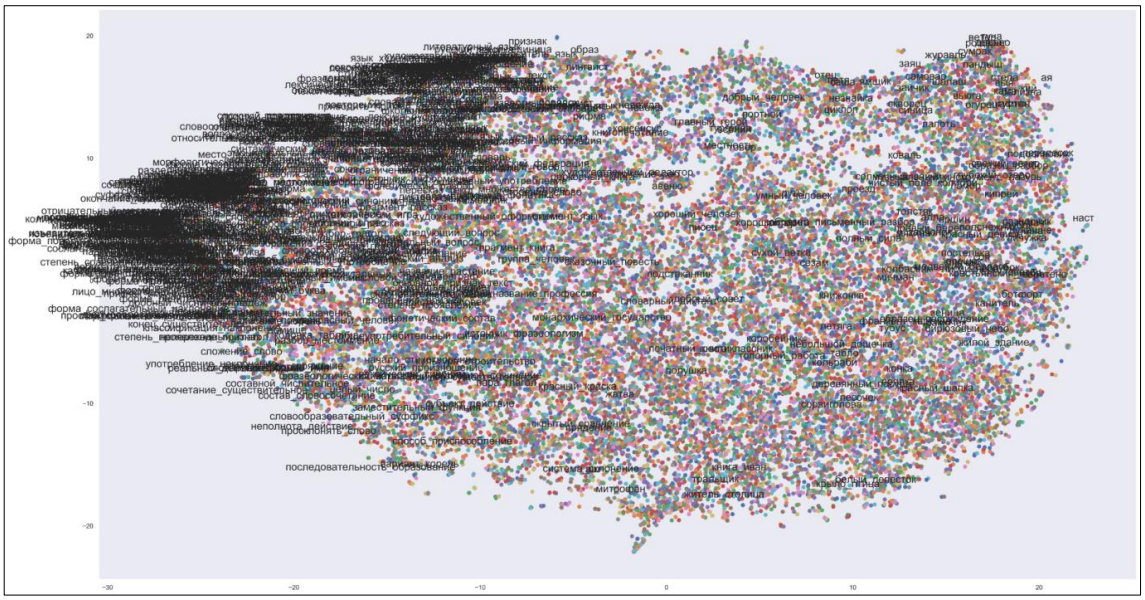


Рисунок 2. Семантическая карта распределения кандидатов в термины. Подкорпус «Русский язык» (6 класс)
Figure 2. Semantic map representing the distribution of term candidates. Subcorpus Russian (grade 6)



При маркировке кластеров принимались во внимание следующие факторы: 1) удельная доля слов, встречающихся внутри кластера и в качестве самостоятельных единиц, и в составе биграмм и триграмм (предполагается, что для кластеров, содержащих термины, характерна более высокая степень повторяемости слов); 2) удельная доля неоднословных единиц внутри кластера (предполагается, что в кластерах, содержащих термины, количество неоднословных единиц больше, так как уровень точности при автоматическом выделении терминологических сочетаний выше, чем при выделении однословных

терминов); 3) удельная доля кандидатов в термины внутри кластера, которые соответствуют терминам, наличествующим в Федеральных государственных образовательных стандартах (предполагается, что для кластеров, содержащих термины, характерно большее число подобных совпадений). С учетом этих факторов высчитывалась единая метрика, значение которой для каждого кластера варьируется от 1 (с высокой вероятностью кластер содержит термины) до 7200 (с высокой вероятностью кластер содержит лжетермины). Ср., например, фрагменты полученных списков для подкорпуса «Русский язык» (9 класс): значение метрики «1» – «русский_язык», «синтаксис», «фонетика», «орфография», «история_язык», «слово», «неологизм», «морфема», «этимология», «старославянский_язык», «значение», «древнерусский_язык», «современный_русский_язык», «славянский_язык», значение метрики «3600» – «ветер», «дерево», «осина», «оса», «колокольчик», «ель», «соловей», «ямщик», «рябина», «пирог», «туман», «гром», «туча», «роса»; для подкорпуса «География» (7 класс): значение метрики «4.6» – «воздушный_масса», «форма_рельеф», «котловина», «высотный_поясность», «высотный_пояс», «землетрясение», «кристаллический_фундамент», «платформа», «муссон», значение метрики «16.8» – «причина_образование», «французский_язык», «карта_приложение», «сочетание_фактор», «карта_евразия», «деление_земля», «бразильский_карнавал», «благосостояние_население», «главный_занятие».

Для усовершенствования результатов автоматического выделения терминов из корпуса научных статей был использован другой способ. Автоматически выделенные кандидаты в термины были иерархически упорядочены по мере их семантического удаления от условного центра лексической системы общеупотребительного языка. За этот центр было принято высчитанное среднее значение O всех векторных представлений, содержащихся в обученной на материале Национального корпуса русского языка дистрибутивно-семантической модели, предоставляемой сервисом RusVectōrēs (Kutuzov, Kuzmenko, 2017). (1) Векторное представление каждого кандидата в термины Ci в исходном списке {LCj}, так что Ci ∈ {LCj}, сопоставлялось с O, а именно высчитывалось косинусное расстояние между векторами θ(Ci) = cos(Ci, O); (2) кандидату с наибольшим косинусным расстоянием θ присваивался индекс 1 как наиболее вероятному термину, после чего он удалялся из списка, так что Ci => K1 и K1 ∈ {KC} ∉ {LCj+1}, если θ(Ci) = argmax(θi … θn); (3) шаги 1–2 повторялись для Ci+1 ∈ {LCj+1} до тех пор, пока список не опустевал, так что {LCn} = ∅ и {KC} = {LCj}. Наконец среди множества иерархически упорядоченных индексов {i … n} в списке {KC} выбирался такой индекс k (i < k < n), чтобы подмножество кандидатов {Kk … Kn} можно было исключить из списка {KC} как составленное из наименее вероятных терминов. На данном этапе исследования выбор соответствующей точки отсечки для каждой дисциплины производился экспертным решением. Ср., например, 15 первых и последних кандидатов в термины из выстроенного подобным образом списка для предметной области «Русский язык»: кандидаты с 15 первыми индексами – «экспликация», «предикативность», «именование», «модус», «денотат», «интенция», «лексема», «актант», «пресуппозиция», «дескрипция», «модальность», «семантика», «референция», «предикат», «пропозиция»; кандидаты с 15 последними индексами – «мальчик», «варвара», «петя», «наци», «бенефициант», «господин», «скотина», «парная», «зеница», «макар», «жучок», «обида», «скука», «тополь», «червяк».

После применения к изначально выделенному материалу описанных операций количество однословных и неоднословных единиц, отнесенных к терминам, составило для корпуса школьных учебников 26 328; распределение по подкорпусам: алгебра – 1 526, астрономия – 456, биология – 2 324, всеобщая история и история России – 2 491, география – 1 635, геометрия – 570, естествознание – 198, изобразительное искусство – 808, информатика – 682, литература – 2 306, математика – 903, математический анализ – 635, мировая художественная культура – 215, музыка – 46, обществознание – 2 286, право – 404, русский

язык – 2 633, технология – 406, физика – 2 836, физическая культура – 1 161, химия – 1 807. Для корпуса научных статей количество однословных и неоднословных единиц, отнесенных к терминам, составило 15 247; распределение по подкорпусам: астрономия – 1 060, биология – 1 157, география – 1 112, информатика – 896, искусствознание – 1 182, история – 891, литературоведение – 1 101, математика – 753, музыковедение – 955, обществознание – 1 116, право – 1 169, русский язык / лингвистика – 945, физика – 999, физическая культура – 892, химия – 1 019.

3. Результаты и обсуждение

3.1. Термины и высокочастотная нетерминологическая лексика

Векторизация данных позволяет улучшить результаты автоматического извлечения терминов и создает основу как для интерпретации терминоупотребления в школьных учебных текстах, так и для дальнейших операций, направленных на изучение его особенностей.

В первую очередь следует обратить внимание на факт автоматического выделения в текстах школьных учебников существенного числа лжетерминов – единиц, относительная частотность которых в целевом корпусе превышает общеязыковые показатели, но отсеянных в результате векторизации. Тематическая характеристика лжетерминов, как и их количество, различается в предметных подкорпусах. Состав лексико-семантических групп, в которые входят лжетермины в учебниках по разным дисциплинам, может стать предметом обсуждения в рамках методики школьного обучения и общей дидактики.

Для ряда дисциплин денотативная соотнесенность лжетерминов очевидным образом определяется тем аспектом реальности, который является предметом описания в соответствующих школьных учебниках; ср., например, фрагмент обширного ряда наименований растительных организмов, высокочастотных в целевом подкорпусе по биологии: «абрикос, акация, арахис, астра, бегония, белена, бузина, вишня, георгин, горох, дуб, дурман, ель, земляника, ива, капуста, картофель, кипарис, кислица, клевер, кукуруза, ландыш, лещина, липа, лиственница, люпин, люцерна, малина, можжевельник, нарцисс, одуванчик, ольха, орешник, орхидея, осина, осока, пальма, папоротник, пеларгония, пихта, подорожник, подсолнечник, пшеница, пырей, редис, редька, репа, рыжик, рябина, саксаул, сирень, слива, сосна, томат, тополь, тюльпан, фасоль, фиалка, фикус, хлопчатник, хризантема, цикорий, шиповник, эвкалипт, яблоня, ясень, ячмень».

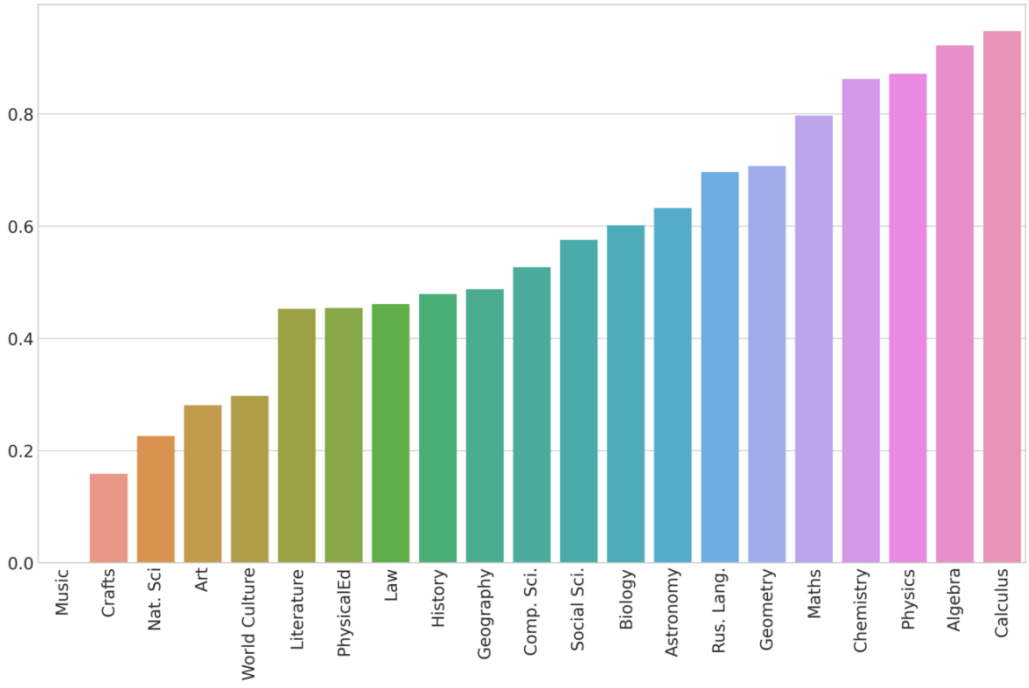
В то же время в иных предметных подкорпусах наличие некоторых ярко выделяющихся групп лжетерминов объясняется скорее сложившейся методической традицией. Так, в подкорпусе «Литература» (10–11 класс) привлекает внимание автоматически выделяемая группа абстрактных существительных с соответствующими суффиксами, относящихся к стилистическим пластам высокой или сугубо книжной лексики: «безверие, возмездие, высокий_предназначение, дарование, духовный_возрождение, жертвенность, искание, искренность, личный_достоинство, мечтание, мироздание, мироощущение, миросозерцание, нравственный_чувство, обличение, поэтический_вдохновение, предчувствие, преображение, прозрение, простодушие, раздумье, самопожертвование, скитание, сострадание, сочувствие, страдание, счастие, тщеславие, человечность, чужбина». Очевидно, что эти слова автоматически попадают в списки кандидатов в термины, поскольку регулярно используются в учебниках для интерпретации художественного смысла произведений или особенностей биографий авторов.

Еще более показательна картина в подкорпусе «Русский язык» на всех ступенях обучения с 5 по 9 класс. В составе лжетерминов здесь в большом количестве обнаруживаются слова и сочетания, являющиеся элементами пейзажных описаний или характеристики природных явлений; ср., например, фрагмент подобного ряда: «аист, былинка, верба, ветер, воробей, восход_солнце, вьюга, глубокий_озеро, голубой_небо, гроза, гром, дождик, долгий_зима, дубрава, дымок, ель, жаворонок, журавль, заяц, зимний_утро, зяблик, ива, изморозь, изморось, иней, камыш, крапива, кукушка, лазурь, ландыш, лесной_озеро, лесной_поляна, лесок, липа, лиса, листопад, метель, наст, начало_осень, облачко, овраг, озимь, орешник, оса, осина, осока, перелесок, песок, пичужка, подосиновик, подснежник,

поздний_осень, пороша, пригорок, проталина, ракита, родной_природа, роса, роща, рябина, свежий_ветер, синий_небо, синица, сирень, скворец, снегирь, снежный_буря, сова, соловей, старый_дуб, стужа, сумрак, туман, туча, фиалка, холодный_ветер, чаща, чибис, шалаш, шмель». Установлено, что доминирование темы природы характерно и для лексического наполнения учебников по русскому языку для младших классов (Лексический состав…, 2019). Следует отметить, что именно для пособий по русскому языку формирование тематически сбалансированной, эмоционально актуальной и стилистически разнообразной речевой среды учебника имеет принципиальное значение, поскольку непосредственным образом связано с дидактическими целями учебного предмета. В связи с этим и вопреки силе методической традиции можно полагать, что ощущаемая интуитивно, но теперь и математически подтвержденная тематическая однородность языкового материала, предлагаемого для лингвистического анализа в учебниках по русскому языку, дидактически малоэффективна.

Представляет интерес количественное соотношение терминов и лжетерминов в кругу высокочастотной лексики в различных предметных подкорпусах. По этому показателю школьные дисциплины образуют градацию с относительно отчетливым противопоставлением учебников по точным и естественным наукам учебникам по гуманитарным предметам. Первые отмечены преобладанием собственно терминологической лексики, в то время как для вторых характерна значительная доля лжетерминов (см. рис. 3). Следует отметить не вполне ожидаемое положение на этой шкале подкорпуса «Русский язык», по степени терминологической насыщенности сближающегося с подкорпусами точных и естественных дисциплин, что указывает на весьма обширный терминологический аппарат, используемый в учебниках по русскому языку (ср. место на этой шкале предмета «Литература», который, казалось бы, образует вместе с предметом «Русский язык» единство в отношении словесности как предмета изучения).

Рисунок 3. Доля терминов от общего числа высокочастотной лексики в подкорпусах корпуса школьных учебников
Figure 3. Share of terms in the total number of high-frequency words in the textbook subcorpora



Интерпретация распределения, приведенного на рис. 3, не вполне тривиальна. Согласно использованным алгоритмам выделение терминологической лексики среди высокочастотных лексических единиц в учебнике обеспечивается (а) регулярным употреблением терминов в тексте, причем (б) таким употреблением, которое в синтагматическом плане сходно с употреблением терминологических единиц той же тематической группы. Несоблюдение первого из этих требований может привести к невыделению единицы в составе высокочастотной лексики в целевом корпусе, несоблюдение второго – к зачислению термина в группу лжетерминов.

Показательно в этом отношении парадоксальное на первый взгляд положение на данной шкале подкорпуса «Естествознание» (код Natsci), который формировался на основе учебников базового уровня для старшей школы, объединяющих сведения по физике, астрономии, химии, биологии. Малая доля терминов в этом подкорпусе, отграниченных от лжетерминов в ходе кластеризации семантических карт, объясняется не только облегченным терминологическим составом текста, но и самим характером употребления терминов: вынужденно беглый обзор проблематики соответствующих научных областей приводит к резкому ослаблению частотности терминологических единиц и отсутствию подобия в контекстуальном поведении терминологических групп.

Другим примером может служить подкорпус «Музыка». При кластеризации соответствующих семантических карт количество выделенных терминов оказывается здесь очень мало по сравнению с другими предметными подкорпусами, что свидетельствует о том, что группы терминов, автоматически выделенные при помощи метрики keyness score (в частности, названия музыкальных жанров – «кантата», «симфония», «сюита» и др.), не обнаруживают сходства в своем текстовом поведении и в конечном счете не попадают в число терминов.

Подобные ситуации возможны и в какой-либо части предметного подкорпуса, включающего лексику конкретного учебно-методического комплекса. Так, в учебно-методическом комплексе «Русский язык. 5–11 классы», созданном под общей редакцией Л. А. Вербицкой (М.; СПб.: Просвещение, 2019), в учебнике для 5 класса встречается сочетание «орфографический_правило», которое не определяется как термин именно в этом учебнике, но распознается в качестве такового в учебниках для 6 и 7 классов, где оно встречается с достаточной частотой и, что более важно, является частью терминологического кластера, включающего в себя такие единицы, как «орфограмма», «правописание_гласный», «правописание_приставка», «правописание_слово», «правописание_суффикс», «ударение», «условие_выбор_буква» и др.

Безусловно, высокие показатели присутствия терминологической лексики среди высокочастотных лексических единиц в учебнике указывают на значительную степень сложности соответствующих учебных текстов. С другой стороны, абсолютная сложность учебного текста, вызванная обилием терминов, в известном смысле компенсируется в плане трудности его усвоения благодаря регулярному и контекстуально систематизированному употреблению терминологических групп, образующих некоторое смысловое целое. И напротив – употребление терминов, не обнаруживающее должной меры частотности и/или контекстуального сходства с употреблением близких по смыслу единиц, может вызывать трудности при усвоении текста, хотя и уменьшает степень его абсолютной сложности.

3.2. Терминологические связи

В свете сказанного большое значение имеют терминологические связи и между разными частями одного учебника, и между разными учебниками в составе одного учебно-методического комплекса. Эти связи реализуют основополагающие для любого текста принципы проспекции и ретроспекции, которые в методике традиционно обозначаются как «опережающее обучение» и «повторение и закрепление пройденного». В первую очередь интересно оценить динамику накопления терминов и удельный вес того вклада, который учебники каждой ступени обучения вносят в формирование терминосистемы соответствующей учебной дисциплины. Алгоритмически решение этой задачи строилось так: 1) каждый учебно-методический комплекс был разбит на отрезки, соответствующие длине одного тематического раздела, в среднем по 1000 словоупотреблений каждый; 2) в каждом из этих отрезков было сосчитано количество

терминов и терминологических сочетаний, относящихся к ранее выделенным терминосистемам разных ступеней обучения; 3) для каждого класса было подсчитано общее количество новых терминов, которые соответствующий учебник вносит в общую терминосистему учебно-методического комплекса; 4) для каждого тематического отрезка учебно-методического комплекса было подсчитано количество терминов из терминосистем всех представленных ступеней обучения.

Ниже приведен пример динамики пополнения терминосостава в учебно-методическом комплексе «Русский язык» под ред. Л. А. Вербицкой: на рис. 4 показана доля новых терминов на каждой ступени обучения, на рис. 5 – изображена та же динамика с указанием на распределение терминоупотребления в разных частях учебника для каждого класса (учебники для 10–11 классов обозначены совместно цифрой 10).

Рисунок 4. Динамика пополнения терминосостава в учебно-методическом комплексе «Русский язык» под ред. Л. А. Вербицкой
Figure 4. The dynamics of new terms entering the terminology system of the teaching methodological package Russian Language edited by L. Verbitskaya

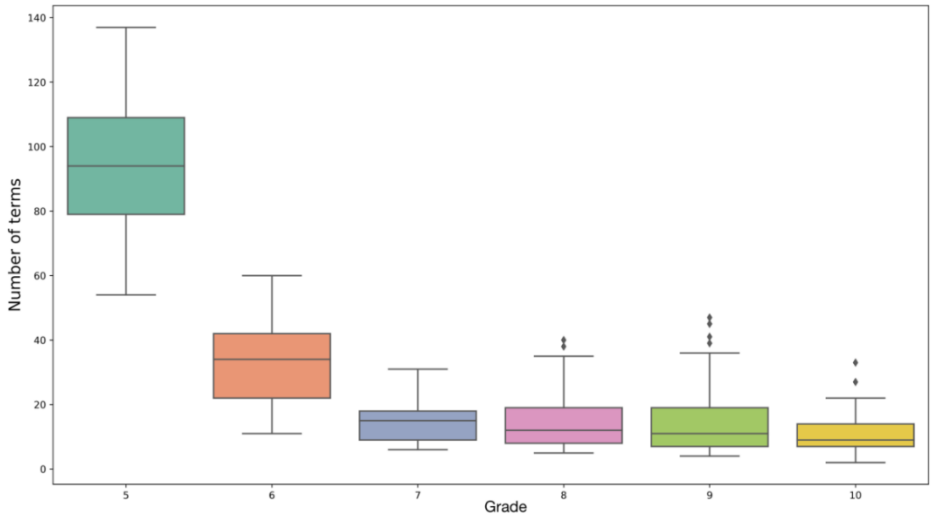
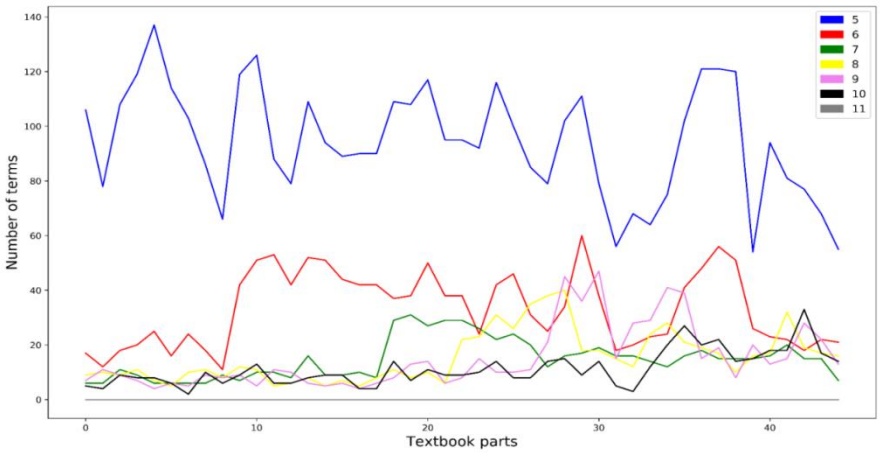


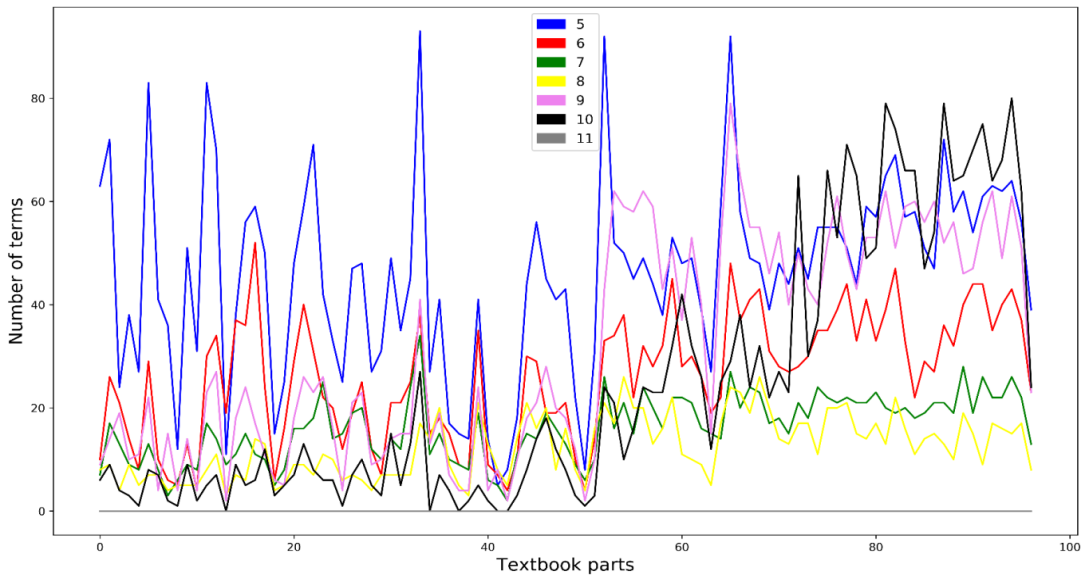
Рисунок 5. Динамика пополнения терминосостава в учебно-методическом комплексе «Русский язык» под ред. Л. А. Вербицкой с указанием на терминоупотребление в разных частях учебника
Figure 5. The dynamics of new terms entering the terminology system of the teaching methodological package Russian Language edited by L. Verbitskaya with the distribution of terms in different parts of textbooks for different grades



Как видно, учебник для 5 класса с большим отрывом лидирует по количеству терминов, представленных в каждом из выделенных отрезков учебно-методического комплекса. Иными словами, в основных чертах терминологическая картина данной области знания складывается уже на раннем этапе обучения, возвращение к ней происходит на всех последующих ступенях обучения. Учебники же для старших классов почти не вводят новой терминологии, что объясняется их нацеленностью в первую очередь на систематизацию и повторение.

Сопоставляя с этими результатами аналогичную визуализацию пополнения терминосостава в учебно-методическом комплексе по литературе В. Я. Коровиной и др. (5–9 классы), В. И. Коровина и др. (10–11 классы) (М.: Просвещение, 2013), можно отметить, что учебники по литературе для старшей школы формируют терминосистему, в значительной степени новую в отношении той, что была сформирована на предыдущих ступенях обучения (см. рис. 6). Различия между учебниками для старшей школы по русскому языку и литературе в отношении их вклада в пополнение терминосостава объясняются в числе прочего тем, первые являются учебниками базового уровня, тогда как вторые – профильного.

Рисунок 6. Динамика пополнения терминосостава в учебно-методическом комплексе по литературе В. Я. Коровиной и др. (5–9 классы), В. И. Коровина и др. (10–11 классы) с указанием на терминоупотребление в разных частях учебника
Figure 6. The dynamics of new terms entering the terminology system of the teaching methodological package Literature edited by V. Korovina et al. (grades 5–9) and V. Korovin et al. (grades 10–11) with the distribution of terms in different parts of textbooks for different grades



Распределение терминов по терминологическим кластерам разных ступеней обучения позволяет проанализировать также, насколько тесно связаны между собой отдельные учебники в рамках одного учебно-методического комплекса. Алгоритмически этот анализ был выстроен следующим образом:

1) для каждого термина учебной дисциплины t был определен набор терминологических кластеров {T5, T6, T7, T8, T9, T10-11}, к которым данный термин относится на соответствующей ступени обучения;

2) для каждой пары определенных терминологических кластеров была высчитана мера пересечения относящихся к ним терминов $Ti \cap Tj$, $i \neq j$;

3) полученные для каждого термина меры были сведены в общую матрицу следующего вида (на примере термина «морфема» в учебно-методическом комплексе по русскому языку под ред. Л. А. Вербицкой):

Таблица 1. Доля общих терминов в кластерах, содержащих термин «морфема», в учебно-методическом комплексе по русскому языку под ред. Л. А. Вербицкой
Table 1. General terms share in clusters with the term *морфема* (morpheme) in the educational and methodological complex Russian Language edited by L. Verbitskaya

Классы	**5**	**6**	**7**	**8**	**9**	**10**
5	1,00	0,74	0,76	0,76	0,74	0,74
6	0,56	1,00	0,62	0,55	0,61	0,70
7	0,64	0,69	1,00	0,75	0,75	0,71
8	0,58	0,55	0,67	1,00	0,61	0,59
9	0,70	0,75	0,83	0,75	1,00	0,75
10	0,60	0,74	0,68	0,63	0,65	1,00

Матрица читается по строкам; цифры в ячейках, варьирующиеся в пределах [0,1], обозначают долю терминов в кластере, к которому отнесен термин t в учебнике Ti, совпадающих с терминами в кластере, к которому отнесен термин t в учебнике Tj;

4) из получившейся матрицы выбиралось максимальное значение – два максимально тесно связанных кластера в разных учебниках (в приведенной выше таблице это 0.83 – доля терминов в кластере, к которому отнесен термин «морфема», в учебнике для 9 класса, совпадающих с терминами в кластере, к которому отнесен термин «морфема», в учебнике для 7 класса);

5) пара учебников Ti и Tj, характеризующихся максимальной степенью пересечения записывалась в словарь forward_classes, если i < j, и в словарь backward_classes, если i > j; таким образом, словарь forward_classes содержал случаи катафорических повторов (терминосистема, представленная на более ранней ступени обучения, входит в состав расширенной терминосистемы, представленной на более поздней ступени обучения), тогда как словарь backward_classes содержал случаи анафорических повторов (терминосистема, представленная на более поздней ступени обучения, воспроизводит терминосистему, представленную на более ранней ступени обучения);

6) данные, полученные для отдельных терминов, входящих в состав учебно-методического комплекса, суммировались по парам классов.

Например, в учебно-методическом комплексе по русскому языку под ред. Л. А. Вербицкой наблюдается следующая иерархия соответствий классов (в круглых скобках до двоеточия даются порядковые номера ступеней обучения, при этом 10 и 11 классы обозначены совместно цифрой 10; цифра после двоеточия представляет собой количество уникальных терминов, показавших для данной пары учебников максимальную степень пересечения содержащих их кластеров; показатели упорядочены от большего к меньшему):

1) forward_classes – {(5, 6): 48, (7, 8): 46, (7, 10): 35, (5, 10): 26, (6, 10): 22, (7, 9): 21, (5, 8): 19, (8, 9): 19, (9, 10): 13, (5, 9): 9, (8, 10): 8, (5, 7): 7, (6, 9): 6, (6, 8): 4, (6, 7): 1},

2) backward_classes – {(10, 6): 62, (10, 9): 51, (9, 7): 44, (8, 6): 44, (7, 6): 36, (10, 8): 17, (10, 7): 10, (6, 5): 10, (7, 5): 8, (9, 8): 6, (8, 7): 6, (10, 5): 5, (9, 6): 5, (8, 5): 3, (9, 5): 1}.

Следует вывод, что в данном учебно-методическом комплексе наиболее тесно связаны между собой – в плане тематического развития – учебники для 5 и 6 классов, с одной стороны, и учебники для 7 и 8 классов, с другой стороны (наименее связаны учебники для 6 и 7 классов). В плане же повторения пройденного наиболее тесно связаны между собой учебники для 10 и 6 классов, с одной стороны, и учебники для 10 и 9 классов, с другой стороны (наименее связаны учебники для 9 и 5 классов).

3.3. Терминоупотребление в различных сферах

Еще одним результатом векторизации целевых корпусов (корпуса школьных

учебников и корпуса научных статей) стала возможность проведения разноаспектного сопоставительного анализа употребления терминов в различных сферах. Ниже будет охарактеризован только один из аспектов подобного сопоставления, а именно сопоставление функционирования терминов в сфере школьной и научной литературы с терминоупотреблением в сфере общеупотребительного языка, в том числе в научно-популярных текстах[1].

Для решения поставленной задачи использовалась предоставляемая ресурсом RusVectōrēs модель ruwikiruscorpora-superbigrams_skipgram_300_2_2018, обученная на 600 миллионах словоупотреблений из Национального корпуса русского языка и Википедии за декабрь 2017 г. (далее – модель RusVectōrēs). Это единственная модель, в которой склеены все возможные биграммы продуктивных типов, независимо от их частотных характеристик. Способность модели распознавать биграммы является необходимым качеством, так как исследуемый нами терминологический состав включает как однословные, так и неоднословные терминологические единицы.

Дальнейшие действия были таковы:

1) была создана и обучена на материале объединенного корпуса школьных учебников и научной периодики новая дистрибутивно-семантическая модель Word2Vec, с размерностью векторов эмбеддинга равной 300 (далее – модель Корпуса). Это было необходимо, поскольку первоначально созданная модель имела размерность 32 и была несопоставима с моделью RusVectōrēs;

2) были отобраны векторные представления всех терминов объединенного корпуса, которые встречаются в словаре модели RusVectōrēs;

3) эти векторные представления были разделены на четыре группы: distance_textbooks_wiki (векторные представления терминов школьных учебников в модели RusVectōrēs), distance_textbooks (векторные представления терминов школьных учебников в модели Корпуса), distance_articles_wiki (векторные представления терминов научной периодики в модели RusVectōrēs), distance_articles (векторные представления терминов научной периодики в модели Корпуса);

4) внутри каждой из четырех групп были получены все возможные парные комбинации отнесенных к этой группе векторных представлений терминов и для каждой парной комбинации высчитана мера косинусной близости CS = u * v / (||u|| * ||v||), так что CS находится в пределах [0,1], где 1 обозначает идентичность векторов, а 0 – их ортогональность;

5) полученные для всех четырех групп данные были подвергнуты однофакторному дисперсионному анализу (ANOVA), чтобы определить соотношение систематической (межгрупповой) дисперсии к случайной (внутригрупповой) дисперсии в измеряемых данных. Высчитанная для каждой из четырех групп мера CSMi, i = 1,2,3,4 – среднее арифметическое всех парных мер косинусной близости – дает общее представление о том, насколько термины данной группы в среднем близки друг к другу в векторном пространстве, то есть в какой степени они образуют семантически спаянную группу;

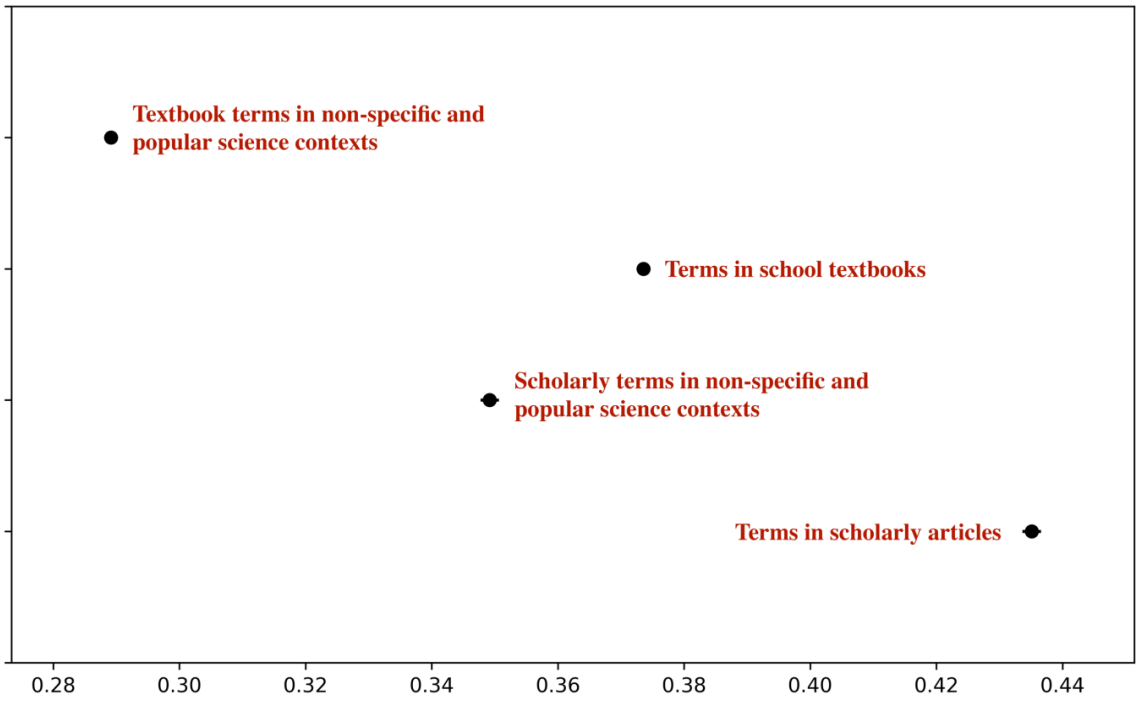
6) для всех анализируемых областей знания были обнаружены статистически значимые отличия четыре групп мер косинусной близости. Поскольку дисперсионный анализ сам по себе не дает возможности ответить на вопрос, какие именно группы отличаются друг от друга, по его завершении были проведены апостериорные сравнения – попарные сравнения изучаемых групп с помощью критерия Тьюки.

Таким образом, при обобщении данных получаем четыре числовых показателя, характеризующих текстовое поведение: (а) терминов в школьных учебниках, (б) терминов в научной периодике, (в) терминов школьных учебников в общеупотребительной, научно-популярной сферах, (г) терминов научной периодики в общеупотребительной, научно-популярной сферах. Пример результатов, полученных для области знания «Русский язык / Лингвистика»,

[1] Другим важным аспектом такого сопоставления является сравнение семантических карт термино-употребления в школьных учебниках и в текстах научной периодики, которое демонстрирует существенные отличия между этими двумя сферами бытования терминов, прежде всего в аспекте актуализации терминов того или иного раздела некоторой области знания, см. об этом (Монахов и др., 2022).

дан на рис. 7: здесь видно, как степень семантической спаянности лингвистических терминов падает в направлении, во-первых, от специальных научных текстов к учебной литературе, а во-вторых, от учебной и научной сферы употребления языка к общеупотребительной и научно-популярной.

Рисунок 7. Сопоставление мер косинусной близости, указывающее на степень семантической спаянности терминов области знания «Русский язык / Лингвистика» в различных сферах употребления
Figure 7. Comparison of cosine similarity measures as indicators of semantic coherence of terms from the subject area Russian Language / Linguistics in different domains



Анализ данных по всем областям знания позволяет объединить их в четыре группы – в зависимости от обнаруженной закономерности:

0 < CSMdistance_textbooks_wiki < CSMdistance_articles_wiki < CSMdistance_textbooks < CSMdistance_articles < 1 – «Искусствознание», «География», «Информатика», «Музыковедение», «Физическая культура», «Русский язык / Лингвистика», «Обществознание»;

0 < CSMdistance_articles_wiki < CSMdistance_textbooks_wiki < CSMdistance_articles < CSMdistance_textbooks < 1 – «Астрономия», «История», «Право», «Литература»;

0 < CSMdistance_articles < CSMdistance_textbooks < CSMdistance_articles_wiki < CSMdistance_textbooks_wiki < 1 – «Биология»;

0 < CSMdistance_textbooks < CSMdistance_articles < CSMdistance_textbooks_wiki < CSMdistance_articles_wiki < 1 – «Химия», «Математика», «Физика».

Закономерным кажется то, что в модели RusVectōrēs термины научной периодики в большинстве сфер знания имеют более высокие показатели средней косинусной близости, чем термины школьных учебников, что говорит о том, что актуальная школьная терминология в целом обладает меньшим потенциалом автономной системности, чем собственно научная. Другой вывод является более неожиданным и в то же время более показательным. Чтобы обозначить его в более явном виде, целесообразно сократить число

уровней с четырех до двух, объединив distance_textbooks_wiki и distance_articles_wiki, с одной стороны, и distance_textbooks и distance_articles, с другой. Тогда получаем:

0 < CSMdistance_textbooks+articles_wiki < CSMdistance_textbooks+articles < 1 – «Искусствознание», «География», «Информатика», «Музыковедение», «Физическая культура», «Русский язык / Лингвистика», «Обществознание», «Астрономия», «История», «Право», «Литература»;

0 < CSMdistance_textbooks+articles < CSMdistance_textbooks+articles_wiki < 1 – «Биология», «Химия», «Математика», «Физика».

Таким образом, выделяются ключевые сферы знания, относимые к точным и естественным наукам (биология, химия, физика, математика), которые противопоставлены остальным дисциплинам, в частности наукам общественно-гуманитарного цикла. Термины первой группы в модели RusVectōrēs характеризуются более высокими показателями средней косинусной близости, чем термины второй группы. Это свидетельствует о том, что термины точных и естественных наук при употреблении вне сферы учебной и научной литературы сохраняют и даже упрочивают свою семантическую близость, выступают как относительно цельная смысловая группа. Можно было бы сказать, что они сопротивляются давлению иной коммуникативной среды и жестко отсекают нетипичные для них сочетаемостные связи, замыкаясь в кругу регулярно повторяющихся контекстов. Термины общественных и гуманитарных наук, напротив, попадая в сферу общеупотребительного языка, теряют сходство в своём текстовом поведении и предстают как семантически распылённое облако. Они свободнее употребляются в контекстах, нетипичных для их терминологической природы, и быстрее детерминологизируются. Эти особенности текстового поведения терминов двух указанных групп в чуждой коммуникативной сфере в некотором роде объективируют понятие «знакомости» слова, которое с давних пор употребляется для оценки степени сложности текста, но, как правило, отвергается как ненадежный критерий в силу его субъективности. Термины точных и естественных наук, даже при употреблении за пределами специальных или учебных текстов, сохраняют свою групповую смысловую герметичность, что препятствует их стихийному семантическому и коммуникативному освоению. Термины общественных и гуманитарных наук, напротив, быстрее становятся «знакомыми», прежде всего в силу их более свободного употребления в лексически разнообразных контекстах.

Несоответствия отмеченным закономерностям (см., например, полученные данные для дисциплин «География», «Информатика», «Астрономия», а с другой стороны – для дисциплин «История», «Право», «Литература») требуют специального рассмотрения, от которого мы в данном случае отвлекаемся, ограничившись следующим замечанием. Причины, которые вызывают числовые отклонения от общей картины терминоупотребления, могут быть разными для разных сфер знания. Они могут быть связаны с комплексной природой дисциплины (например, география сочетает в себе черты естественных и общественных наук), или с малым объемом информации, обусловленным слабой представленностью дисциплины в школьном преподавании (это касается, например, астрономии), или, наконец, с существенной разнородностью автоматически выделенного терминологического состава определенной дисциплины, включающего большое число нетерминологических единиц. В последнем случае выявленные нарушения обнаруживаемых закономерностей носят диагностический характер и указывают на необходимость дальнейшего совершенствования методик автоматического вычленения терминов и машинного анализа характеристик их употребления.

4. Заключение

Использованные методы исследования терминоупотребления, основанные на принципах дистрибутивной семантики и алгоритмах Word2Vec, учитывают неслучайное, регулярное употребление терминов в сходных лексических контекстах. Тем самым появляется возможность анализировать текстовое поведение терминов как элементов терминосистем, образованных семантически связанными группами лексических единиц. В свою очередь это позволяет не только

усовершенствовать результаты автоматического вычленения терминов из целевых корпусов, опирающегося на традиционный статистический подход, но и проследить на больших текстовых объемах за функционированием терминов в разных сферах знания. В отношении современных школьных учебников становится возможным, в частности, сопоставить терминологическую насыщенность учебных пособий по разным предметам – имея в виду как раз систематическое употребление терминологических групп; охарактеризовать состав высокочастотной нетерминологической лексики; рассмотреть динамику пополнения терминосистем как в пределах группы учебно-методических комплексов или одного из них, так и внутри конкретной учебной книги определенной ступени обучения; сопоставить в разных аспектах закономерности терминоупотребления в школьной, научной и неспециальной областях.

Полученные результаты могут представлять интерес для специалистов по автоматическому анализу текстов, по общей дидактике и методике отдельной области знания, по комплексологии.

Отметим, что эти результаты в одном случае могут вполне коррелировать с устоявшимися интуитивными представлениями о характере учебных текстов по той или иной дисциплине, но в другом – существенно уточнять их. Так, получает математическое обоснование общераспространенное представление о «сложности» и «строгости» точных и естественных наук. Подчеркнем, что в аспекте терминоупотребления эти качества обеспечиваются не просто обилием слов специальной семантики, но их жесткой контекстуальной и семантической сплоченностью, степень которой сохраняется и даже возрастает при употреблении вне основной сферы их бытования. С другой стороны, учебники по русскому языку и литературе, хотя и имеют, как считается, во многом общий предмет описания, существенно отличаются и степенью терминологической насыщенности, и мерой системности терминоупотребления, и особенностями пополнения терминологического состава при переходе от основной школы к старшей.

Обширность терминологического поля, частотность терминологической или нетерминологической лексики, безусловно, коррелируют с мерой сложности школьных учебных текстов. Однако объективные лексические показатели сложности вступают здесь в очень непростые и порой противоречивые отношения как с мерой трудности, так и с принципами дидактической эффективности школьного учебника. Например, выявленное лексико-тематическое однообразие языкового материала в школьных учебниках по русскому языку снижает меру сложности, как и меру трудности текста (лексическое разнообразие – один из факторов, повышающих меру сложности), и так или иначе способствует реализации дидактического принципа доступности. Между тем этот же фактор отрицательно влияет на формирование мотивации к обучению и противоречит дидактическим требованиям к современному учебнику обеспечить психологическое соответствие учебного материала возрастным и индивидуальным особенностям школьников. Нерегулярный и контекстуально разрозненный характер терминоупотребления во многих учебниках общественно-гуманитарного цикла снижает меру сложности текста, но при этом противоречит дидактическим принципам преемственности, последовательности и систематичности обучения, что отнюдь не способствует снижению меры трудности. Безусловно, дальнейшие исследования лексической сложности школьного учебного текста должны опираться и на оценку композиции как конкретной учебной книги, так и учебно-методического комплекса в целом, ибо характеристика соотношения сложности и трудности невозможна без учета динамики обновления терминологического состава и взаимодействия впервые вводимых и уже известных из предыдущего изложения терминов.

Изложенные результаты являются частью итогов проводимого исследования, направленного в конечном счете на формирование базы данных русской терминологической лексики, соответствующей содержанию среднего образования. Разработанные программные коды на языке Python позволяют

воспроизвести все описанные алгоритмы на материале любого учебно-методического комплекса или иного терминологически насыщенного корпуса текстов. Все материалы и результаты исследования, включая корпусы текстов, таблицы терминов, программные коды, дистрибутивно-семантические модели, графики и семантические карты, помещены в научное хранилище открытого доступа[1].

Список литературы

Иомдин Б. Л., Морозов Д. А. Кто поймет «Незнайку»? Автоматическое определение сложности текстов для детей // Русская речь. 2021. № 5. С. 55–68. DOI: 10.31857/S013161170017239-1

Лапошина А. Н., Лебедева М. Ю., Берлин Хенис А. А. Влияние частотности слов текста на его сложность: экспериментальное исследование читателей младшего школьного возраста методом айтрекинга // Russian Journal of Linguistics. 2022. Т. 26. № 2. С. 493–514. DOI: 10.22363/2687-0088-30084

Лейчик В. М. Терминоведение: предмет, методы, структура. М.: ЛКИ, 2007. 256 с.

Лексический состав текстов учебников русского языка для младшей школы: корпусное исследование / Лапошина А. Н., Веселовская Т. С., Лебедева М. Ю., Купрещенко О. Ф // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог» (Москва, 29 мая – 1 июня 2019 г.). Вып. 18 (25). М., 2019. С. 351–363.

Лукашевич Н. В., Логачев Ю. М. Комбинирование признаков для автоматического извлечения терминов // Вычислительные методы и программирование. 2010. Т. 11. Вып. 4. С. 108–116.

Мартынова Е. В., Солнышкина М. И., Мерзлякова А. Ф., Гизатулина Д. Ю. Лексические параметры учебного текста (на материале текстов учебного корпуса русского языка) // Филология и культура. 2020. № 3 (61). С. 72–80.

Микк Я. А. Оптимизация сложности учебного текста: В помощь авторам и редакторам. М.: Просвещение, 1981. 119 с.

Митрофанова О. А., Захаров В. П. Автоматизированный анализ терминологии в русскоязычном корпусе текстов по корпусной лингвистике // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог» (Бекасово, 27–31 мая 2009 г.). Вып. 8 (15). М., 2009. С. 321–328.

Монахов С. И., Турчаненко В. В., Чердаков Д. Н. Школьная и научная терминология: корпусное исследование и кластеризация // Информатизация образования и методика электронного обучения: цифровые технологии в образовании. Материалы VI Международной научной конференции. Красноярск, 2022. Ч. 3. С. 228–233.

Морозов Д. А., Иомдин Б. Л. Критерии семантической сложности слова // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог» (Москва, 29 мая – 1 июня 2019 г.). Вып. 18 (25). М., 2019. Дополнительный том. С. 119–131.

Пиотровский Р. Г., Ястребова С. В. Статистическое опознание термина // Статистика текста / гл. ред. Р. Г. Пиотровский. Т. 1. Минск: Белорусский государственный университет, 1969. С. 249–259.

Солнышкина М. И. Определение уровня лексической сложности текстов: современное состояние проблемы // Сборник научных трудов X Юбилейной международной научно-практической конференции «Учитель. Ученик. Учебник (в контексте глобальных вызовов современности)», 19–20 ноября 2021. М., 2022. С. 20–24.

Солнышкина М. И., Макнамара Д., Замалетдинов Р. Р. Обработка естественного языка и изучение сложности дискурса // Russian Journal of Linguistics. 2022. Т. 26. № 2. С. 317–341. DOI: 10.22363/2687-0088-30171

Солнышкина М. И., Кисельников А. С. Сложность текста: этапы изучения в отечественном прикладном языкознании // Вестник Томского государственного университета. Филология. 2015. № 6 (38). С. 86–99. DOI: 10.17223/19986645/38/7

Соловьев В. Д., Солнышкина М. И., Макнамара Д. С. Компьютерная лингвистика и дискурсивная комплексология: парадигмы и методы исследований // Russian Journal of Linguistics. 2022. Т. 26. № 2. С. 275–316. DOI: 10.22363/2687-0088-30161

Степанова Д. В. Анализ методов автоматического выделения терминов из научно-технических текстов // Актуальные проблемы современной прикладной лингвистики. Минск: Минский государственный лингвистический университет, 2017. С. 62–67.

Татаринов В. А. Общее терминоведение: Энциклопедический словарь. М.: Московский Лицей, 2006. 528 с.

Шпаковский Ю. Ф. Оценка трудности восприятия и оптимизация сложности учебного текста (на материале текстов по химии): Автореф. … канд. филол. наук. Минск, 2007. 21 с.

[2] https://zenodo.org/record/4079198#.X4Mrfy1h29Y; https://zenodo.org/record/5722495#.YZ7FUS2ZPpA

Brownlee J. Deep Learning for Natural Language Processing: Develop Deep Learning Models for your Natural Language Problems. Vermont: Machine Learning Mastery Publ., 2017. 414 p.

Cabré M. T., Estopà R., Vivaldi J. Automatic Term Detection: A Review of Current Systems // Recent Advances in Computational Terminology / Bourigault D., Jacquemin Ch., L'Homme M.-C. (eds.). Amsterdam: John Benjamins Publ., 2001. Pp. 53–87. DOI: 10.1075/nlp.2.04cab

Durda K., Buchanan L. WINDSORS: Windsor Improved Norms of Distance and Similarity of Representations of Semantics // Behavior Research Methods. 2008. Vol. 40. Pp. 705–712. DOI: 10.3758/BRM.40.3.705

Fisher D., Frey N., Lapp D. Text Complexity: Stretching Readers with Texts and Tasks. Thousand Oaks, CA: Corwin Press, 2016. 216 p.

Flor M., Klebanov B., Sheehan K. Lexical Tightness and Text Complexity // Proceedings of the 2th Workshop of Natural Language Processing for Improving Textual Accessibility (NLP4ITA). Atlanta, 2013. Pp. 29–38.

Glazkova A., Egorov Yu., Glazkov M. A Comparative Study of Feature Types for Age-Based Text Classification // Analysis of Images, Social Networks and Texts. AIST 2020. Lecture Notes in Computer Science. Vol. 12602 / van der Aalst W. et al. (eds.). Cham: Springer Publ., 2021. Pp. 120–134.

Jones M. N., Mewhort D. J. K. Representing Word Meaning and Order Information in a Composite Holographic Lexicon // Psychological Review. 2007. Vol. 114. Pp. 1–37. DOI: 10.1037/0033-295X.114.1.1

Kilgarriff A., Jakubíček M., Kovář V. et al. Finding Terms in Corpora for Many Languages with the Sketch Engine // Proceedings of the Demonstrations at the 14th Conference the European Chapter of the Association for Computational Linguistics, 26–30 April, 2014. Gothenburg, 2014. Pp. 53–56. DOI: 10.3115/v1/E14-2014.

Korkontzelos I., Ananiadou S. Term Extraction // Oxford Handbook of Computational Linguistics / Mitkov R. (ed.). Oxford: Oxford University Press, 2014. Pp. 991–1012.

Kutuzov A., Kuzmenko E. WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models // Analysis of Images, Social Networks and Texts. AIST 2016. Communications in Computer and Information Science. Vol. 661 / Ignatov D. et al. (eds.). Cham: Springer Publ., 2017. Pp. 155–161.

Levy O., Goldberg Y. Linguistic Regularities in Sparse and Explicit Word Representations // Proceedings of the Eighteenth Conference on Computational Natural Language Learning / Morante R., Yih S. W-t. (eds.). Proceedings of the Eighteenth Conference on Computational Natural Language Learning. Stroudsburg: Association for Computational Linguistic Publ., 2014. Pp. 171–180. DOI: 10.3115/v1/W14-1618

Mikolov T., Sutskever I., Chen K. et al. Distributed Representations of Words and Phrases and their Compositionality // Advances in Neural Information Processing Systems 26. 27th Annual Conference on Neural Information Processing Systems 2013. Red Hook: Curran Associates Publ., 2013. Pp. 3136–3144.

Mikolov T., Yih W. T., Zweig G. Linguistic Regularities in Continuous Space Word Representations // Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: Association for Computational Linguistics, 2013. Pp. 746–751.

Nokel M. A., Bolshakova E. I., Loukachevitch N. V. Combining Multiple Features for Single-word Term Extraction // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог» (Бекасово, 30 мая – 3 июня 2012 г.). Вып. 11 (18). М., 2012. С. 490–501.

Rohde D. L., Gonnerman L. M., Plaut D. C. An Improved Model of Semantic Similarity Based on Lexical Co-Occurrence // Communications of the ACM. 2006. Vol. 8. Pp. 627–633.

Schwanenflugel P. J. Why are Abstract Concepts Hard to Understand? // The Psychology of Word Meanings / Schwanenflugel P. J. (ed.). Hillsdale: Lawrence Erlbaum Associates Inc., 1991. Pp. 223–250.

Sharoff S. What Neural Networks Know about Linguistic Complexity // Russian Journal of Linguistics. 2022. Т. 26. № 2. С. 371–390. DOI: 10.22363/2687-0088-30178

Solovyev V. D., Ivanov V. V., Solnyshkina M. I. Assessment of Reading Difficulty Levels in Russian Academic Texts: Approaches and Metrics // Journal of Intelligent & Fuzzy Systems. 2018. Vol. 34 (2). Pp. 3049–3058. DOI:10.3233/JIFS-169489

Turney P. D., Pantel, P. From Frequency to Meaning: Vector Space Models of Semantics // Journal of Artificial Intelligence Research. 2010. Vol. 37. Pp. 141–188. DOI: 10.1613/jair.2934

References

Brownlee, J. (2017). *Deep Learning for Natural Language Processing: Develop Deep Learning Models for your Natural Language Problems*, Machine Learning Mastery Publ., Vermont, USA. *(In English)*

Cabré, M. T., Estopà, R. and Vivaldi, J. (2001). Automatic Term Detection: a Review of Current Systems, in Bourigault, D., Jacquemin, Ch. and L'Homme, M.-C. (eds.), *Recent Advances in*

Computational Terminology, John Benjamins Publ., Amsterdam, Netherlands, 53–87. DOI: 10.1075/nlp.2.04cab *(In English)*

Durda, K. and Buchanan, L. (2008). WINDSORS: Windsor Improved Norms of Distance and Similarity of Representations of Semantics, *Behavior Research Methods*, 40, 705–712. DOI: 10.3758/BRM.40.3.705 *(In English)*

Fisher, D., Frey, N. and Lapp, D. (2016). *Text Complexity: Stretching Readers with Texts and Tasks*, Corwin Press, Thousand Oaks, CA, USA. *(In English)*

Flor, M., Klebanov, B. and Sheehan, K. (2013). Lexical Tightness and Text Complexity, *Proceedings of the 2th Workshop of Natural Language Processing for Improving Textual Accessibility (NLP4ITA)*, Atlanta, USA, 29–38. *(In English)*

Glazkova, A., Egorov, Yu. and Glazkov, M. (2021). A Comparative Study of Feature Types for Age-Based Text Classification, in van der Aalst, W. et al. (eds.), *Analysis of Images, Social Networks and Texts. AIST 2020. Lecture Notes in Computer Science, 12602*, Springer Publ., Cham, Switzerland, 120–134. *(In English)*

Iomdin, B. L. and Morozov, D. A. (2021). Who Can Understand "Dunno"? Automatic Assessment of Text Complexity in Children's Literature, *Russkaya Rech'*, 5, 55–68. DOI: 10.31857/S013161170017239-1 *(In Russian)*

Jones, M. N. and Mewhort, D. J. K. (2007). Representing Word Meaning and Order Information in a Composite Holographic Lexicon, *Psychological Review*, 114, 1–37. DOI: 10.1037/0033-295X.114.1.1 *(In English)*

Kilgarriff, A., Jakubíček, M., Kovář, V. et al. (2014). Finding Terms in Corpora for Many Languages with the Sketch Engine, *Proceedings of the Demonstrations at the 14th Conference the European Chapter of the Association for Computational Linguistics*, Gothenburg, Sweden, 53–56. DOI: 10.3115/v1/E14-2014 *(In English)*

Korkontzelos, I. and Ananiadou, S. (2014). Term Extraction, in Mitkov, R. (ed.), *Oxford Handbook of Computational Linguistics*, Oxford University Press, Oxford, UK, 991–1012. *(In English)*

Kutuzov, A. and Kuzmenko, E. (2017). WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models, in Ignatov, D. et al. (ed.), *Analysis of Images, Social Networks and Texts. AIST 2016. Communications in Computer and Information Science*, 661, Springer Publ., Cham, Switzerland, 155–161. *(In English)*

Laposhina, A. N., Lebedeva, M. U. and Berlin Khenis, A. (2022). Word Frequency and Text Complexity: An Eye-tracking Study of Young Russian Readers, *Russian Journal of Linguistics*, 26 (2), 493–514. DOI: 10.22363/2687-0088-30084. *(In Russian)*

Laposhina, A. N., Veselovskaya, T. S., Lebedeva, M. U. and Kupreshchenko, O. F. (2019). Lexical Analysis of the Russian Language Textbooks for Primary School: Corpus Study, *Computational Linguistics and Intellectual Technologies: papers from the Annual International Conference "Dialogue"*, Moscow, Russia, 18 (25), 351–363. *(In Russian)*

Lejchik, V. M. (2007). *Terminovedenie: predmet, metody, struktura* [Terminology Studies: Subject, Methods, Structure], LKI Publishing House, Moscow, Russia. *(In Russian)*

Levy, O. and Goldberg, Y. (2014). Linguistic Regularities in Sparse and Explicit Word Representations, *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, Baltimore, USA, 171–180. DOI: 10.3115/v1/W14-1618 *(In English)*

Lukashevich, N. V. and Logachev, Yu. M. (2010). Combining Features for Automatic Term Extraction, *Numerical Methods and Programming*, 11 (4), 108–116. *(In Russian)*

Martynova, E. V., Solnyshkina, M. I., Merzlyakova, A. F. and Gizatulina, D. Yu. (2020). Lexical Parameters of the Academic Text (Based on the Texts of the Academic Corpus of the Russian Language), *Philology and Culture*, 3, 72–80. DOI: 10.26907/2074-0239-2020-61-3-72-80 *(In Russian)*

Mikk, Ya. A. (1981). *Optimizaciya slozhnosti uchebnogo teksta: V pomoshch' avtoram i redaktoram* [Optimizing the complexity of educational text: To help authors and editors], Prosveshchenie, Moscow, Russia. *(In Russian)*

Mikolov, T., Sutskever, I., Chen, K. et al. (2013a). Distributed Representations of Words and Phrases and their Compositionality, *Advances in Neural Information Processing Systems 26, 27th Annual Conference on Neural Information Processing Systems 2013*, Lake Tahoe, USA, 3136–3144. *(In English)*

Mikolov, T., Yih, W. T and Zweig, G. (2013b). Linguistic Regularities in Continuous Space Word Representations, *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, USA, 746–751. *(In English)*

Mitrofanova, O. A. and Zakharov, V. P. (2009). Automatic Analysis of Terminology in the Russian Text Corpus on Corpus Linguistics, *Computational Linguistics and Intellectual Technologies: papers from the Annual International Conference "Dialogue"*, Bekasosvo, Russia, 8 (15), 321–328. *(In Russian)*

Monakhov, S. I., Turchanenko, V. V. and Cherdakov, D. N. (2022). Terminology in Textbooks and Research Articles: Cluster Analysis of Corpus Data, *Proceedings of 6th International Conference*

"Informatization of Education and E-learning Methodology: Digital Technologies in Education", Krasnoyarsk, Russia, 3, 228–233. *(In Russian)*

Morozov, D. A. and Iomdin, B. L. (2019). Criteria of Semantic Complexity of Words, *Computational Linguistics and Intellectual Technologies: papers from the Annual International Conference "Dialogue"*, Moscow, Russia, 18 (25), 119–131. *(In Russian)*

Nokel, M. A., Bolshakova, E. I. and Loukachevitch, N. V. (2012). Combining Multiple Features for Single-word Term Extraction, *Computational Linguistics and Intellectual Technologies: papers from the Annual International Conference "Dialogue"*, Bekasosvo, Russia, 11 (18), 1, 490–501. *(In English)*

Piotrovskij, R. G. and Yastrebova, S. V. (1969). Statistical Term Recognition, in Piotrovskij, R. G. (ed.), *Statistika teksta* [Text statistics], Belorusskij gosudarstvennyj universitet, Minsk, Belarus, 1, 249–259. *(In Russian)*

Rohde, D. L., Gonnerman, L. M. and Plaut, D. C. (2006). An Improved Model of Semantic Similarity Based on Lexical Co-Occurrence, *Communications of the ACM*, 8, 627–633. *(In English)*

Schwanenflugel, P. J. (1991). Why are Abstract Concepts Hard to Understand?, in Schwanenflugel, P. J. (ed.), *The psychology of word meanings*, Lawrence Erlbaum Associates Inc., Hillsdale, USA, 223–250. *(In English)*

Sharoff, S. (2022). What Neural Networks Know about Linguistic Complexity, *Russian Journal of Linguistics*, 26 (2), 371–390. DOI: 10.22363/2687-0088-30178 *(In English)*

Shpakovskij, Yu. F. (2007). Estimation of Perception Difficulty and Optimization of the Educational Text Complexity (on the Material of Texts in Chemistry), Abstract of Ph.D. dissertation, Linguistics, Minsk State Linguistic University, Minsk, Belarus. *(In Russian)*

Solnyshkina, M. I. (2022). Measuring Text Complexity: State of the Art, *Collection of Scientific Papers X Jubilee International Scientific Conference "Teacher. Student. Textbook (in the Context of Global Challenges of Modern Times)"*, Moscow, Russia, 20–24. *(In Russian)*

Solnyshkina, M. I. and Kiselnikov, A. S. (2015). Text Complexity: Study Phases in Russian Linguistics, *Tomsk State University Journal of Philology*, 6 (38), 86–99. DOI: 10.17223/19986645/38/7 *(In Russian)*

Solnyshkina, M. I., McNamara, D. and Zamaletdinov, R. R. (2022). Natural Language Processing and Discourse Complexity Studies, *Russian Journal of Linguistics*, 26 (2), 317–341. DOI: 10.22363/2687-0088-30171 *(In Russian)*

Solovyev, V. D., Ivanov, V. V. and Solnyshkina, M. I. (2018). Assessment of Reading Difficulty Levels in Russian Academic Texts: Approaches and Metrics, *Journal of Intelligent & Fuzzy Systems*, 34 (2), 3049–3058. DOI: 10.3233/JIFS-169489 *(In English)*

Solovyev, V. D., Solnyshkina, M. I. and McNamara, D. (2022). Computational Linguistics and Discourse Complexology: Paradigms and Research Methods, *Russian Journal of Linguistics*, 26 (2), 275–316. DOI: 10.22363/2687-0088-30161 *(In English)*

Stepanova, D. V. (2017). Analiz metodov avtomaticheskogo vydeleniya terminov iz nauchno-tekhnicheskih tekstov [Analysis of Methods for Automatic Terms Extraction from Scientific and Technical Texts], *Aktual'nye problemy sovremennoj prikladnoj lingvistiki* [Current problems of modern applied linguistics], Minskij gosudarstvennyj lingvisticheskij universitet, Minsk, 62–67. *(In Russian)*

Tatarinov, V. A. (2006). *Obshchee terminovedenie: Entsiklopedicheskij slovar'* [Terminology Studies: Encyclopedic Dictionary], Moskovskij Litsej, Moscow, Russia. *(In Russian)*

Turney, P. D. and Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics, *Journal of Artificial Intelligence Research*, 37, 141–188. DOI: 10.1613/jair.2934 *(In English)*

Все авторы прочитали и одобрили окончательный вариант рукописи.

All authors have read and approved the final manuscript.

Конфликты интересов: у авторов нет конфликтов интересов для декларации.

Conflicts of interests: the authors have no conflicts of interest to declare.

Сергей Игоревич Монахов, кандидат филологических наук, научный сотрудник Йенского университета им. Ф. Шиллера, Германия.
Sergei I. Monakhov, candidate of philology, research associate, Friedrich Schiller University Jena, Germany.

Владимир Владимирович Турчаненко, младший научный сотрудник Института русской литературы (Пушкинский Дом) РАН, Россия.
Vladimir V. Turchanenko, junior researcher, Institute of Russian Literature (Pushkinskij Dom) of the Russian Academy of Sciences, Saint Petersburg, Russia.

Дмитрий Наилевич Чердаков, старший преподаватель Санкт-Петербургского государственного университета, Россия.
Dmitrii N. Cherdakov, senior lecturer, Saint Petersburg University, Russia.