УДК 81`33 **DOI: 10.18413/2313-8912-2024-10-2-0-6**

Лапошина А. Н.[1]
Храмченко Т. А.[2]
Лебедева М. Ю.[3]

Отбор многословных выражений на основе корпусных источников и экспертной оценки: обновление языкового содержания РКИ

[1] Государственный институт русского языка им. А.С. Пушкина
ул. Академика Волгина, 6, Москва, 117485, Россия
E-mail: antonina.laposhina@gmail.com
ORCID: 0000-0003-0693-7657

[2] Белорусский государственный университет
пр. Независимости, 4, Минск, 220030, Беларусь
E-mail: hramchenkot@mail.ru
ORCID: 0000-0001-9328-8098

[3] Государственный институт русского языка им. А.С. Пушкина
ул. Академика Волгина, 6, Москва, 117485, Россия
E-mail: m.u.lebedeva@gmail.com
ORCID: 0000-0002-9893-9846

Статья поступила 19 декабря 2023 г.; принята 15 июня 2024 г.;
опубликована 30 июня 2024 г.

Информация об источниках финансирования или грантах: работа выполнена при финансовой поддержке госзадания, проект FZNM-2020-0005 «Трансформация когнитивной и коммуникативной деятельности человека в условиях современной информационной среды» (Лапошина А. Н., Лебедева М. Ю.). Исследование проведено во время участия Храмченко Т. А. в научно-исследовательской стажировке «InteRussia» при финансовой поддержке Фонда Горчакова.

Аннотация: Статья описывает опыт создания с опорой на корпусные данные списка наиболее педагогически ценных многословных выражений для задач преподавания русского языка иностранным учащимся. Современные лингвистические и когнитивные исследования показывают, что наша речь шаблонизирована, во многом состоит из устойчивых сегментов. Этот факт подкрепляется и лингводидактической идеей обучения не изолированным языковым единицам, а их комплексам разной природы. Однако отбор и ранжирование по уровням языкового владения многословных выражений ограничивается трудностью их автоматизированного выделения из корпуса текстов и подсчетом частотности, а также разногласиями в определении границ многословных выражений, их лингвистической природы и терминологии. В данной статье описывается опыт компиляции списка многословных выражений фиксированного типа из разных источников: двух типов существующих уровневых списков РКИ, наиболее частотных n-gram корпуса текстов из учебников РКИ RuFoLa, корпуса интернет-текстов Russian Web, а также списка дискурсивных формул проекта «Прагматикон». В качестве меры определения языкового уровня многословного выражения используется мера максимальной

Delta на основе информации о частотности выражения в корпусе учебных текстов для иностранных учащихся, эффективность которой затем проверяется множественной оценкой экспертов. Получившийся список многословных выражений содержит 1645 вхождений, распределенных по уровням шкалы CEFR от A1 до C1. Полученная версия списка внедрена в систему автоматического анализа сложности текста для изучающих РКИ и может быть полезна широкому кругу профильных специалистов при создании учебного контента. Предложенная мера максимальной Delta показала высокую степень совпадения с оценками уровня экспертами на уровнях A1-B1, что говорит о целесообразности дальнейшего изучения её потенциала для смежных прикладных задач и задач отбора языкового содержания на материале других языков.

Ключевые слова: Многословные выражения; Многословные лексические единицы; Обучение лексике; Формульные последовательности; Лексический подход; Коллокации; Русский язык как иностранный

Информация для цитирования: Лапошина А. Н., Храмченко Т. А., Лебедева М. Ю. Отбор многословных выражений на основе корпусных источников и экспертной оценки: обновление языкового содержания РКИ // Научный результат. Вопросы теоретической и прикладной лингвистики. 2024. Т. 10. № 2. С. 117-137. DOI: 10.18413/2313-8912-2024-10-2-0-6

UDC 81`33 **DOI: 10.18413/2313-8912-2024-10-2-0-6**

Antonina N. Laposhina[1]
Tatsiana A. Khramchanka[2]
Maria Yu. Lebedeva[3]

Multi-word expressions for Russian L2 learners: corpora-based selection with expert verification

[1] Pushkin State Russian Language Institute,
6 Akademika Volgina St., Moscow, 117485, Russia
E-mail: antonina.laposhina@gmail.com
ORCID: 0000-0003-0693-7657

[2] Belarusian State University,
4 Nezavisimosti Avenue, Minsk, 220030, Belarus
E-mail: hramchenkot@mail.ru
ORCID: 0000-0001-9328-8098

[3] Pushkin State Russian Language Institute,
6 Akademika Volgina St., Moscow, 117485, Russia
E-mail: m.u.lebedeva@gmail.com
ORCID: 0000-0002-9893-9846

Received 19 December 2023; accepted 15 June 2024; published 30 June 2024

Acknowledgements: The article was prepared with the financial support of the state assignment of Ministry of Education and Science of the Russian Federation for 2020–2024 (No. FZNM-2020-0005) - Laposhina A.N., Lebedeva M.Yu. The study was conducted during the participation of Khramchanka T. A. in the fellowship programme «InteRussia» of the Gorchakov Fund.

Abstract: The article describes the experience of creating a corpus-based list of the most relevant multi-word expressions for Russian L2 learners, distributed across the levels of the Common European Framework of Reference for Languages (CEFR) from A1 to C1. Modern linguistic and cognitive research shows that our speech is patterned and largely consists of stable segments. This fact is supported by the linguodidactic idea of teaching not isolated language units but their combinations of different nature. However, the selection and ranking of multi-word expressions based on language proficiency levels is constrained by the difficulty of automatically extracting them from a corpus of texts and estimating their frequency, as well as disagreements in defining the boundaries, linguistic nature, and terminology of multi-word expressions. This article describes the experience of compiling a list of the most valuable fixed-type multi-word expressions from various sources: two types of existing CEFR-graded vocabulary lists for Russian L2 learners – lexical minimums for the TORFL (Test of Russian as a Foreign Language) system and Russian KELLY (KEywords for Language Learning for Young and adults alike); the most frequent n-grams from the RuFoLa – Russian L2 textbook corpus and from the Russian Web corpus of internet texts; list of discourse formulas from the «Pragmaticon» project. The CEFR level of each multi-word expression is predicted using the frequency-based Max Delta measure, and its effectiveness is subsequently validated through annotation by multiple experts. The resulting list of multi-word expressions contains 1645 entries from A1 to C1 levels. The proposed version of the list has been implemented into an automated text analysis system for learners of Russian as a Foreign Language and can be useful for a wide range of professionals in the preparation of educational content for foreign language learners. The suggested Max Delta measure has demonstrated a high degree of agreement with expert evaluations within proficiency levels A1-B1. This signifies the importance of further exploring its potential in addressing related practical tasks and in selecting language learning content derived for other languages.

Keywords: Multi-word expressions; Multi-word units; Vocabulary acquisition; Formulaic sequences; Lexical approach; Collocations; Russian as a foreign language
How to cite: Laposhina, A. N., Khramchanka, T. A. and Lebedeva, M.Yu. (2024) Multi-word expressions for Russian L2 learners: corpora-based selection with expert verification, *Research Result. Theoretical and Applied Linguistics*, 10 (2), 117-137. DOI: 10.18413/2313-8912-2024-10-2-0-6

Введение

Масштабные корпусные исследования показывают, что до 50% письменной и устной речи шаблонизировано, т.е. состоит из готовых наборов слов и выражений, сочетание которых вполне предсказуемо – т.н. формул (De Cock, Granger, Leech, & Mcenery, 1998). Классики структурной лингвистики Л. Н. Иорданская и И. А. Мельчук также отмечают, что «люди говорят не словами, а фраземами» (т. е. различными несвободными словосочетаниями) (Иорданская, Мельчук, 2007). Эти положения подкрепляются когнитивными исследованиями, в которых доказывается, что слова или словосочетания обрабатываются одними и теми же когнитивными механизмами (Bybee, 1998; Christiansen & Chater, 2016; Elman, 2009; McClelland, 2010). Предположительно, формульные конструкции хранятся как единое целое в ментальном лексиконе говорящих (Schmitt, 2004). Это находит подтверждение в исследованиях на материале конкретных языков – так, показано, что носители

английского языка с одинаковой скоростью реагируют на шаблонные фразы, состоящие из трех слов, и на частотные трехсловные идиомы (Jolsvai, McCauley&Christiansen, 2013).

Эти особенности обработки языка и хранения языковой системы имеют принципиальное значение для такой прикладной области лингвистики, как методика обучения языку. В соответствии с данными лингвистических и психолингвистических исследований, в преподавании языка складывается консенсус о том, что важно обучать не изолированным языковым единицам, а их комплексам (Schmitt, 2004; Wray, 2000; Lewis, 1997; Свирина, 2019).

В теоретической лингвистике такие комбинации квалифицируются по-разному. Для описания этой области языковой системы характерно терминологическое многообразие и широкое поле пересекающихся терминов: *многословные единицы, многокомпонентные единицы, сверхлексемные единицы, полилексемные единицы, коллокации, чанки, полуфраземы, лексикализованные выражения, формульные последовательности, формульные выражения, дискурсивные формулы, лексические паттерны, неоднословные единицы, несвободные выражения и мн.др*.

Однако в прикладном аспекте вопросы лингвистической классификации и терминологии отходят на второй план. Значимым остается вопрос о том, какие именно сочетания из нескольких слов целесообразнее предъявить как одно целое при обучении иностранному языку. В данной статье мы будем оперировать понятием *многословного выражения* (англ. multiword expression, MWE) как максимально широким, общим термином, означающим «последовательность слов, которая действует как единое целое на определенном уровне лингвистического анализа» (Calzolari et al. 2002, перевод с англ. наш). Этот термин способен вместить в себя разные по своей лингвистической природе и степени связанности сочетания, такие как многословные лексемы (*может быть, потому что, банковская карта)*, коллокации (*проливной дождь, оказывать влияние*), речевые формулы (*сколько стоит, на обратном пути*), этикетные формулы (*спокойной ночи, будь здоров*), дискурсивные формулы (*Это еще что! Вот оно как!*), идиомы (*на одной волне, в два счета*) и др. Ещё одну сложность определения термина и предмета исследования составляет определение границ многословных выражений: к ним относятся как фиксированные последовательности из двух и более слов, так и конструкции, состоящие из менее строго определённых элементов с возможными вариативными вставками. Фокус настоящего исследования направлен на многословные выражения фиксированного типа.

Итак, актуальной прикладной исследовательской проблемой является отбор и ранжирование многословных выражений для их включения в содержание обучения языку.

Обзор литературы

Шаблонность речи активно изучается в области преподавания иностранных языков (Schmitt, 2004; Wray, 2000). Так, исследования показывают, что знание многословных единиц у изучающих иностранный язык значительно отстает от их общего словарного запаса (Bahns and Eldaw, 1993). Даже студенты продвинутых уровней продуцируют меньше устойчивых выражений по сравнению с носителями языка, причем как в устной, так и в письменной речи (Paquot & Granger, 2012). Способность использовать шаблоны в языке является показателем свободного владения языком и одним из его аспектов, который отличает детей, овладевающих родным языком, от тех, кто изучает язык как иностранный (Wray, 2002). Ошибки в выборе или лексико-грамматическом оформлении несвободных выражений отмечаются в качестве типичной черты русской речи иностранцев (Ерёмина 2020).

Сочетания из нескольких лексических единиц становятся центральным объектом освоения в лексическом подходе к обучению языку, предложенном Майклом Льюисом (Lewis, 1997). Введенное им понятие лексических чанков включает в себя устойчивые словосочетания, идиомы, фразовые глаголы, коллокации, а также отдельные слова и части предложений.

В исследованиях многословных выражений на материале русского языка, обнаруживается дисбаланс большого количества работ, посвященных лингвистической природе этих выражений, и низкой представленностью работ, посвященных практике преподавания и отбора этих единиц в учебных целях. Так, существует большой пласт лингвистических исследований, связанных с автоматизацией извлечения многословных единиц русского языка и их категоризации (Loukachevitch, Lashevich, 2016; Kopotev, 2013; Janda et al., 2020; Пужаева и др. 2018). Часть исследований сопровождается созданием веб-сервисов с возможностью поиска и анализа многословных выражений. Проект «CoCoCo»[1] предлагает алгоритм и интерфейс для поиска по большим корпусам текстов возможных компонентов многословных единиц различных типов (идиом, коллокаций) к заданному слову (Kopotev, 2015). «Русский Конструктикон[2]» – электронная база конструкций русского языка, сопровождаемых формальными и семантическими признаками, а также разметкой по уровням по шкале CEFR (Janda et al., 2020). База содержит как относительно связанные конструкции, например, *(как) по мне, (так)...*, так и сложные паттерны со множеством переменных и факультативных единиц: *(единственное) (Prep) что (не) VP, так это (не) XP/Cl (*например, *чего дочь не сделала, так это не убралась)*. Каждая конструкция сопровождается информацией об уровне CEFR, однако методика отнесения к тому или иному уровню не уточняется. С проектом «Русский Конструктикон» тесно связан проект «Прагматикон[3]», представляющий базу толкований и примеров для неоднословных прагматических выражений: например, *Все ясно! Как скажешь! Только так!* (Пужаева и др., 2018).

С другой стороны, наблюдается малая представленность многословных единиц в учебных и регулирующих документах по русскому языку как иностранному. На примере дискурсивных выражений русского языка исследователями констатируется малое внимание к этой теме в методике преподавания РКИ (Шляхов, Саакян, 2015). Отмечается также малая представленность таких конструкций в учебниках РКИ и лексических минимумах, несмотря на их высокую частотность в русской речи (Минаева 2017).

Отдельные разделы «Устойчивые выражения» и «Пословицы и поговорки» появляются в современных лексических минимумах Тестирования русскому языку как иностранному (ТРКИ, eng. TORFL) только на уровне B2. На более ранних уровнях многословные выражения самой различной лингвистической природы могут встречаться в общем алфавитном списке как в качестве самостоятельных вхождений в список, таких как *вести себя, сельское хозяйство, может быть, то есть* (см. рисунок 1), так и в виде примеров сочетаемости (см. рисунок 2).

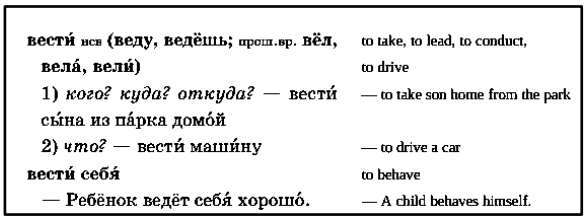
[1] Информационный ресурс «CoCoCo: Collocations, Colligations, Corpora». URL: https://cococo.cosyco.ru/ (дата обращения: 10.11.2023)

[2] Информационный ресурс «Конструктикон». URL: https://constructicon.github.io/russian/ (дата обращения: 10.11.2023)

[3] Информационный ресурс «Прагматикон». URL: https://pragmaticon.ruscorpora.ru (дата обращения: 10.11.2023)

Рисунок 1. Пример оформления многословного выражения в качестве отдельного вхождения в Лексический минимум ТРКИ уровня B1 (Лексический минимум, 2019)

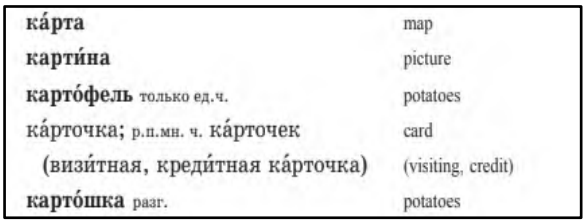
Figure 1. An example of a MWE as a separate entry in the TORFL Lexical Minimum level B1 (Lexical Minimum, 2019)



вести́ нсв **(веду, ведёшь;** прош.вр. **вёл, вела́, вели́)**	to take, to lead, to conduct, to drive
1) *кого? куда? откуда?* — вести́ сы́на из па́рка домо́й	— to take son home from the park
2) *что?* — вести́ маши́ну	— to drive a car
вести́ себя́	to behave
— Ребёнок ведёт себя́ хорошо́.	— A child behaves himself.

Рисунок 2. Пример оформления многословного выражения в качестве варианта сочетаемости в Лексическом минимуме ТРКИ уровня B1 (Лексический минимум, 2019)

Figure 2. An example of MWE as a possible collocate in the TORFL Lexical Minimum level B1 (Lexical Minimum, 2019)



ка́рта	map
карти́на	picture
карто́фель только ед.ч.	potatoes
ка́рточка; р.п.мн. ч. ка́рточек	card
(визи́тная, креди́тная ка́рточка)	(visiting, credit)
карто́шка разг.	potatoes

При формировании градуированных по уровням CEFR списков многословных выражений, неизменно встает вопрос о критериях их отбора, определения их востребованности в иноязычной аудитории и уровня их языковой сложности. Официальные лексические минимумы ТРКИ опираются в основном на экспертную оценку методической ценности выражений: частотность упоминается авторами как один из критериев, но не является основным. Лексические списки международного проекта KELLY, напротив, основываются на частотных данных по большим корпусам текстов и лишь корректируются экспертами (Kilgariff et al., 2014).

Исследовательская группа проекта CEFRLex предлагает учитывать при отнесении к уровню обобщенную экспертную оценку, выраженную частотностью слова или выражения в пособиях для иностранцев, изучающих язык на разных уровнях по шкале CEFR. Этот подход, с одной стороны, базируется на подсчетах частотности, однако сам материал для подсчетов – не большие национальные корпуса, которые чаще всего используются для подобных задач, а корпус текстов из пособий для иностранцев – призван продемонстрировать методическую востребованность слова, выраженную в обобщенном коллективном мнении авторов о включении слова или сочетания в учебники определенного уровня (François et al., 2014; Volodina et al., 2016).

С проблемой подсчета частотности многословных выражений теснейшим образом связана проблема их автоматического поиска в тексте. Несмотря на то, что эта область активно разрабатывается (Parmentier et al., 2019), в том числе и на русскоязычном материале для отдельных типов многословных выражений (Kopotev et al., 2016, Пужаева и др., 2018, Loukachevitch et al., 2016, Инькова, 2015), проблема корректной идентификации и определения наиболее частотных многословных выражений остается одной из сложных задач автоматической обработки текстов на естественном языке.

Таким образом, **цель статьи** состоит в отборе и разметке по уровням языкового владения наиболее актуальных многословных выражений фиксированного типа с опорой на корпусные данные для их включения в содержание обучения русскому языку как иностранному. Выбранная цель ставит перед нами две задачи, решение которых последовательно описано в статье: 1) формирование списка

кандидатов наиболее употребимых многословных выражений на основе существующих лексических баз и большого корпуса текстов 2) градуирование списка кандидатов по шкале уровней владения иностранными языками CEFR.

Материалы и методы

Для формирования первичного списка словосочетаний-кандидатов мы использовали три группы источников: многословные выражения, указанные в существующих лексических списках по РКИ; наиболее частотные сочетания из нескольких слов (n-граммы) по корпусу текстов учебников РКИ и по большому корпусу интернет-текстов Russian Web 2011; список дискурсивных формул проекта «Прагматикон».

К первой группе источников относятся система лексических минимумов ТРКИ (Андрюшина и др., 2019 a, 2019 b, 2020, 2021, 2022) и лексические списки для иностранных учащихся на основе корпусных данных KELLY (Kilgariff, 2014). Лексические минимумы ТРКИ содержат отдельные списки устойчивых словосочетаний только с уровня B2, где объединены идиомы, пословицы и поговорки и другие типы многословных выражений. На более ранних уровнях многословные выражения предлагаются в основном списке в виде примеров сочетаемости к предлагаемым к изучению словам (например, *вести: вести себя*), а также вне алфавитного списка в разделе этикетных формул (*добрый день, очень жаль*). Списки для русского языка KELLY являются частью мультиязычного проекта градуированных списков лексики для изучающих язык как иностранный, созданных на основе данных частотности слова по большому корпусу интернет-текстов. Многословные единицы в этих списках выделены в отдельную группу, MWE: *за рубежом, домашнее животное, всего лишь и др.*). Сравнение количества кандидатов по указанным спискам в зависимости от уровня указано в Таблице 1.

Таблица 1. Количество многословных единиц в существующих лексических списках для изучающих РКИ
Table 1. Number of multi-word units in existing vocabulary lists for Russian L2 students

Источник	Лексические минимумы ТРКИ		Списки Russian KELLY	
Уровень	Кол-во много-словных ед-ц	Общий объём списка	Кол-во многословных ед-ц	Общий объём списка
A1	51	780	27	912
A2	54	1 300	56	1998
B1	64	2 300	75	3 560
B2	138	5 100	146	6 983
C1	594	11 000	48	8 376
C2	-	-	21	8 958
Всего	**901**	**-**	**373**	**-**

Второй группой источников для формирования списка кандидатов в списки наиболее употребимых многословных выражений стали списки наиболее частотных n-грамм по двум корпусам текстов: корпусу текстов из учебников РКИ RuFoLa (Russian as a Foreign Language) (Лапошина, 2020). Он состоит из текстов печатных и электронных учебников по русскому языку как иностранному, сопровождаемых информацией об уровне учебника по шкале CEFR от A1 до C1. Корпус включает 68 источников текстов из изданий начиная с 2005 года, 40 из которых входят в учебные линейки книг («Дорога в Россию», «Точка Ру», «Поехали» и мн. др.) и 28 являются «отдельными» учебниками для конкретного уровня («В мире людей», «Окно в Россию» и мн.др.). Общий размер корпуса составляет около 650 000 токенов. На материале этого корпуса с помощью корпусного менеджера SketchEngine был создан список 1 000 наиболее частотных сочетаний из 2-6 слов (встроенная функция N-grams, поиск по леммам). Аналогичным методом была собрана 1 000 наиболее частотных сочетаний из 2-6 слов, полученные на материале фрагмента корпуса Russian Web 2011 от проекта SketchEngine объемом в 1 млн. словоупотреблений.

Далее для отобранных кандидатов была применена серия частеречных фильтров для удаления незавершенных цепочек (*вчера гулять и*, *купить большой),* географических названий (*красное море, охотный ряд* и т.п.), а также ручное редактирование для удаления дубликатов (например, *обращать внимание, обратить внимание на…*). После редактирования список кандидатов пополнился 803 многословными выражениями, не представленными ранее (*заниматься спортом, в прошлом году, хотеть есть, уметь делать, по интернету, социальная сеть, в высокой степени, прожиточный минимум и др.*).

Наконец, третьим источником кандидатов стал список из 597 дискурсивных формул, функционирующих в качестве ответа на реплику, собранных в рамках проекта «Прагматикон» (*Ух ты! Будь что будет! Не то чтобы. Да ладно!* и др.).

Все собранные словосочетания-кандидаты были объединены в один список на основании лемматизированных форм: с помощью приведения слов к начальной форме удалось объединить различные формы словосочетаний в разных списках (например, *добрый день* и *доброго дня*). Итоговый список словосочетаний-кандидатов содержит 2572 уникальных вхождения.

Ранжирование многословных выражений по языковым уровням на основе частотных данных

Для каждого многословного выражения из объединенного списка кандидатов мы рассчитали его встречаемость в учебниках РКИ по каждому уровню. Для этого была использована нормализованная частотность, ipm (item per million). Далее в расчетах везде использована мера нормализованной частотности.

Для того чтобы предположить уровень сложности кандидата по шкале CEFR, мы применили метрику на основе расчетов значимого начала использования выражения в учебниках для изучающих русский язык как иностранный, предложенную Д. Альфтером и коллегами (Alfter et al., 2016). Значимое начало использования слова или выражения выражается в максимальном значении разницы частотности на исследуемом уровне с предыдущим, уровне максимальной дельты (далее Max Delta). Иными словами, слову или выражению присваивается уровень, в учебниках которого был замечен максимальный рост его частотности по сравнению с учебниками предыдущих уровней. В качестве шкалы уровней сложностей мы приняли стандартную шкалу уровней владения языком CEFR. Так, частотность выражения в учебниках A1 будет

сравниваться с нулем, в учебниках A2 – с учебниками A1, B1 – с A2 и т.д. Дельта D для уровня i рассчитывается по формуле (1), где f_i – это частотность выражения с учебниках исследуемого уровня, и f_{i-1} – частотность выражения в учебниках предыдущего уровня.

(1) $D_i = |f_i - f_{i-1}|$

После расчета дельты для учебников всех уровней уровень с максимальным значением дельты принимался как уровень сложности данного выражения. Т.к. эта мера не учитывает абсолютные значения встречаемости выражения в учебниках, мы добавили эмпирически выведенное условие, при котором в случае, если сочетание встретилось на данном уровне больше 20 раз в абсолютных значениях, мы присваиваем слову этот уровень, несмотря на значения коэффициента Max Delta.

Примеры расчета меры Max Delta и сравнение полученных значений с имеющейся информацией о сложности выражения в списках ТРКИ и KELLY представлены в таблице 2.

Таблица 2. Примеры кандидатов в список наиболее употребимых многословных выражений и их встречаемость по разным источникам
Table 2. Examples of candidates for the list of the most commonly used MWEs and their occurrence according to various sources

Многословная единица	**Уровень ТРКИ**	**Уровень KELLY**	**Уровень Max Delta**	**Частотность в учебниках РКИ, IPM**				
				A1	**A2**	**B1**	**B2**	**C1**
выйти/выходить замуж	A2	A2	A2	0	24	48	5	12
под открытым небом	C1	-	A2	0	21	9	11	18
водительские права	-	A1	B1	0	18	36	5	0
испытывать жажду	-	B1	-	0	0	0	0	0
социальная сеть	-	-	B1	0	38	188	39	129
как минимум	-	B2	B2	0	19	0	22	24
на самом деле	C1	B2	B2	0	0	61	106	188
мне пора	-	-	A1	16	6	12	10	12

Часть кандидатов из Таблицы 2 демонстрирует по разным источникам единство мнения об их методической ценности и рекомендованном уровне в терминах CEFR (*выйти замуж*). Есть случаи, когда многословная единица предлагается к включению в список, но с изменением языкового уровня (единица *водительские права* указана в списке KELLY как A1, но расчеты дельты

указывают на уровень B1), есть примеры кандидатов, которые не встречаются ни разу во всей коллекции учебников РКИ и показывают низкую частотность по корпусу Russian Web, а потому предлагаются к исключению из списка (*московская гостиница, испытывать жажду*). Наконец, есть примеры единиц, которых нет в существующих списках, однако они активно используются в учебниках, а потому предлагаются к включению в список (*социальные сети*).

Экспертная оценка полученных списков

Вторым этапом работы стала закрытая экспертная оценка методической ценности кандидатов из списка первого этапа, которая проводилась 6 экспертами в области РКИ для проверки эффективности метода отнесения к уровню с помощью меры Max Delta[4]. Для уменьшения трудоёмкости задачи мы разделили процесс аннотации на две ступени. Первая включала разметку всех предложенных многословных единиц двумя экспертами. Перед экспертами стояла задача аннотации каждого словосочетания по уровню CEFR или рекомендации удалить сочетание из списка. По результатам первого этапа были удалены кандидаты, которые были отмечены тэгом «удалить» обоими экспертами и которые ни разу не появились в корпусе учебников РКИ, среди них были как кандидаты из списков KELLY (*почтовый ящик, выхлопная труба, половое сношение* и др.) и из лексических минимумов (*московская гостиница, госпожа Петрова, идет балет* и др.) и др (195 единиц). Также на первой ступени за сочетаниями, в которых уровень по Max Delta совпал с суждением обоих экспертов, был закреплен этот уровень (663 единицы).

Вторая ступень разметки включала аннотацию 4 другими экспертами оставшихся сочетаний с несовпавшими на 1 этапе суждениями об уровне сочетания. Перед экспертами стояла аналогичная первому этапу задача: поставить или предполагаемый уровень кандидата по шкале CEFR, или специальный символ для обозначения кандидатов, которых эксперт не считает нужным включать в финальный список.

Таблица 3 содержит результаты оценки согласия 4 экспертов второго этапа разметки по дихотомической шкале *оставить кандидата в списке или удалить*.

Невзвешенный коэффициент согласия между экспертами составляет 0,839 со стандартной ошибкой 0,005 и 95% доверительным интервалом (0,829, 0,849). Это указывает на высокий уровень согласия экспертов в вопросе целесообразности нахождения того или иного выражения в списке. Значение p равно 0, что указывает на то, что согласие между оценщиками является статистически значимым.

Таблица 4 содержит результаты оценки согласия экспертов в присвоении конкретного уровня сложности CEFR для каждого сочетания. Эта задача сложнее, поскольку предполагает отнесение каждого сочетания к одному из 5 уровней от A1 до C1. В данной работе применялась методология, в соответствии с которой оценка надежности и согласия экспертов присваивалась только в случае, когда все эксперты выбрали «оставить выражение» и оценили ее по шкале от 1 до 5.

[4] Авторы выражают сердечную благодарность откликнувшимся экспертам за проделанную работу и поддержку проекта, а также Кащенко Е. С. за помощь в подсчете согласия экспертов

Таблица 3. Невзвешенный коэффициент согласия 4 экспертов по вопросу о том, оставить кандидата в списке или нет
Table 3. Unweighted coefficient of agreement of 4 experts on the question of whether to leave a candidate on the list or not

Метод	Коэффициент	Станд. ошибка	95% Д.И.	P-Value
Процент согласия	0.839	0.005	(0.829, 0.849)	0.000e+00

Таблица 4. Невзвешенный коэффициент согласия экспертов при распределении многословных выражений по 5 уровням CEFR (A1-C1)
Table 4. Unweighted rate of expert agreement for the distribution of multiword expressions by 5 CEFR levels (A1-C1)

Метод	Значение коэффициента	Станд. ошибка	95% Д.И.	P-Value
Альфа Криппендорфа	0.316	0.008	(0.301, 0.332)	0.000e+00
Бреннан-Предигер	0.497	0.010	(0.478, 0.516)	0.000e+00
Процент согласия	0.581	0.008	(0.564, 0.597)	0.000e+00

Все эксперты выбрали одинаковый уровень сложности сочетания в 58,1% случаев, что указывает на умеренный уровень согласия. Коэффициенты Альфа Криппендорфа и Бреннан-Предигер также указывают на умеренный и высокий уровни согласия соответственно. Значение p для обоих коэффициентов меньше 0,05, что указывает на статистическую значимость наблюдаемых уровней согласия. Коэффициент Бреннана-Предигера равен 0,497 со стандартной ошибкой 0,010 и 95% доверительным интервалом (0,478, 0,516), что также указывает на высокий уровень достоверности этой оценки согласия.

Обобщенный уровень сложности выражения был рассчитан на основании комбинации полученных экспертных данных (4 оценки) и меры Max Delta (1 оценка). Выражение оставалось в списке, если получало 4 и более оценок. Выражению присваивался уровень на основании медианного значения 4 оценок экспертов и меры Max Delta: Например, сочетание *образ жизни*, двумя экспертами отмеченный уровнем A2, двумя – B1, и мерой Max Delta – B1, получает финальную экспертную оценку B1.

Результаты

Финальный список многословных единиц (RFL-LIST MWE) содержит 1 645 сочетаний, распределенных по уровням шкалы CEFR от A1 до C1, как показано в Таблице 5.

Таблица 5. Распределение полученного списка наиболее употребимых многословных единиц RFL-LIST MWE по уровням CEFR
Table 5. Distribution of the resulting RFL-LIST MWE by CEFR levels

Уровень CEFR	Количество MWE	Примеры
A1	114	может быть; день рождения; очень приятно; идти пешком; мобильный телефон; электронная почта; идти в душ
A2	199	за границей; ложиться спать; образ жизни; водить машину; пешеходный переход; не мочь жить без; счастливого пути; ничего страшного
B1	201	выйти на пенсию; давать возможность; средство массовой информации; точка зрения; под открытым небом; честно говоря
B2	528	сельское хозяйство; пресная вода; глава семьи; научный сотрудник; человек с ограниченными возможностями; на самом деле, в приниципе
C1	603	вступать в силу, гражданский брак; канатная дорога; головной мозг; задом наперед; само собой разумеется, в конечном счёте; едва ли
Всего	**1645**	**-**

Список включает в себя многословные выражения, показавшие свою актуальность в ходе частотного анализа и экспертной разметки, самой разной лингвистической природы: многокомпонентные лексемы (*мобильный телефон, потому что*), этикетные формулы (*счастливого пути*), коллокации (*образ жизни, пресная вода*), вводные конструкции (*в конечном счете, честно говоря*), дискурсивные выражения (*к тому же, по сути, ну и ну*). Объемы списка для каждого уровня занимают от 10 до 17% от общего объема лексики, предложенной для изучения на данном уровне.

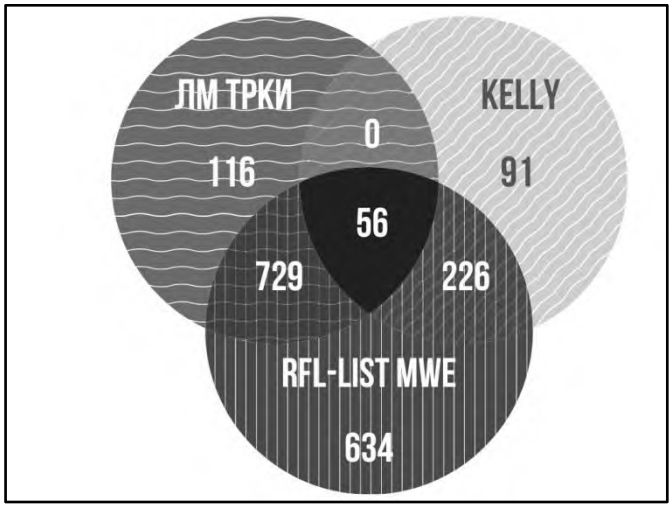
Всего список содержит около 200 комбинаций частей речи компонентов многословных выражений, однако большая часть списка (52%) покрывается 10 самыми частотными сочетаниями частей речи, проиллюстрированными в таблице 6.

Степень пересечения полученного списка с уже известными лексическими списками для РКИ визуализирована на рисунке 3. Всего 56 многословных единиц (3% списка) встречаются по всем трем источникам. Около 44% списка пересекается с Лексическим минимумом ТРКИ (из них большую часть составляют фразеологизмы и поговорки), и около 14% – со списком KELLY. При этом 38% списка (634 единицы) представляют собой многословные единицы, не предлагавшиеся в ранее созданных списках. Часть кандидатов из ранее созданных списков не попали в RFL-LIST MWE по результатам экспертной оценки. Показательно, что нет ни одного выражения, которое бы встречалось в обоих ранее созданных списках, но отсутствовало бы в RFL-LIST MWE. Это говорит о приоритете полноты списка над минимизацией его объема на данном этапе.

Таблица 6. 10 самых частотных комбинаций частей речи в составе многословных выражений полученного списка RFL-LIST MWE
Table 6. The top-10 frequent POS-combinations in multi-word expressions in the RFL-LIST MWE

Комбинация частеречных тэгов	**Кол-во сочетаний в списке**	**Пример**
прилагательное+существительное	290	банковская карта
глагол+существительное	158	проводить время
глагол+предлог+существительное	119	выйти на пенсию
предлог+прилагательное+существительное	87	под открытым небом
предлог+существительное	78	по сути, в итоге, в шоке
существительное+существительное	68	образ жизни
предлог+прилагательное+существительное	50	на всякий случай
наречие+глагол	38	честно говоря
местоимение+существительное	27	всей душой
наречие+частица	25	вряд ли, сразу же

Рисунок 3. Распределение общих и уникальных вхождений полученного списка многословных выражений RFL-LIST MWE, лексических минимумов ТРКИ и списков проекта KELLY
Figure 3. Distribution of common and unique occurrences of the resulting list of multi-word expressions RFL-LIST MWE, official lexical minima of TORFL, and Russian KELLY list



Практическое применение списка многословных единиц. Полученный список имеет несколько вариантов практического приложения в области преподавания русского языка как иностранного. Первым вектором применения полученного списка является непосредственное использование информации о наиболее употребимых многословных конструкциях определенного типа и уровня языкового владения для разработки учебных материалов и упражнений, а также дополнения лексических минимумов. Например, приведем полученный список наиболее употребимых дискурсивных формул, оцененных экспертами уровнем B2: *тем более, ни разу, в самом деле, дай бог, не исключено, как знать, как знать, честное слово, ничего подобного, что за вопрос, ну и ну, нет так нет, не факт, что теперь, ну и что, вот как, ну как, а как же, как сказать, а вдруг, надо же, с ума сойти, а что такое, вот видишь, вот так вот, не совсем так, трудно сказать, я же говорил (говорила), не спрашивай, да ладно, не говори, не смешно, не удивительно, хорошо бы, да так, как хочешь, какая разница, ни за что, сам не знаю, это точно, вроде как, в смысле, без сомнения, еще бы, еще как, понятное дело, какой смысл, никаких сомнений, так надо, не вопрос, похоже на то.*

Вторым вектором использования полученного списка является его внедрение в систему автоматизированного анализа текста для оптимизации работы алгоритма автоматического определения сложности текста сервиса «Текстометр»[5]. В рамках анализа введенного пользовательского текста алгоритм проходит несколько шагов автоматической обработки текста. Первичная предобработка текста включает в себя очистку от лишних символов и знаков ударения, приведение слов текста к нижнему регистру. Поиск по подмножеству точных форм осуществляется для поиска конструкций в фиксированных грамматических формах, например: *спокойной ночи, более того*. После морфологического анализа текста и приведения слов к начальной форме становится доступен поиск по подмножеству лемматизированных версий многословных конструкций для поиска изменяемых по падежам, числам, лицам и склонениям конструкций (*свободный время, образ жизнь, друг друг*). Наконец, последний шаг включает анализ лексики, не вошедшей в многословные конструкции, по алгоритмам анализа односложных лексем.

Результатом работы данной части алгоритма является детальный анализ пользовательского текста и демонстрация однословных и многословных конструкций текста, выходящих за рамки словников каждого из уровней по шкале CEFR. Схема предлагаемой лексической информации для диалога (1) приведена в Таблице 7.

(1) – Ну ладно, я пойду, мне пора домой, тогда до завтра! Во сколько я могу тебе позвонить завтра и обсудить нашу презентацию?

– Давай созвонимся завтра в 10, тебе удобно?

– Да, договорились. Не забудь зонтик, там льёт как из ведра. Тебе заказать такси?

– Нет, спасибо, я на машине. Всего хорошего!

Описанный в таблице 7 функционал позволяет, во-первых, получить информацию о словах и многословных выражениях, остающихся за пределами целевого уровня владения русским языком, для прогнозирования возможных трудностей и составления плана работы с лексикой текста. Особенно актуальной представляется проверка текста сервисом на наличие прагматических выражений при работе с текстами, передающими устную речь: диалогами, интервью, расшифровками аудиоподкастов и др. Полученную информацию преподаватель может использовать исходя из целей урока и уровня языковой подготовки обучающихся: мотивировать замену сложной конструкции, запланировать притекстовую работу с данными единицами.

[5] Информационный ресурс «Текстометр». URL: https://textometr.ru/ (дата обращения: 10.11.2023)

Таблица 7. Прототип результата системы автоматического анализа текста с многословными выражениями
Table 7. Prototype result of an automated text analysis system with multi-word expressions

Название признака	**Значение признака**
Лексический список A1 покрывает	78%
Не входит в лексический список A1	ну ладно обсуждать презентация созваниваться удобно договариваться лить как из ведра заказывать такси всего хорошего
Лексический список A2 покрывает	83%
Не входит в лексический список A2	ну ладно обсуждать презентация созваниваться договариваться лить как из ведра
Лексический список B1 покрывает	89%
Не входит в лексический список B1	презентация созваниваться лить как из ведра
Лексический список B2 покрывает	89%
Не входит в лексический список B2	презентация созваниваться лить как из ведра
Лексический список C1 покрывает	100%
Анализ многословных конструкций текста	
Речевые формулы в тексте	мне пора во сколько заказывать такси
Дискурсивные формулы в тексте	ну ладно
Этикетные формулы в тексте	до завтра всего хорошего
Идиомы в тексте	лить как из ведра

Дискуссия

Рассмотрим, во-первых, полученные результаты с точки зрения эффективности использования частотных данных по корпусу учебников РКИ и применения меры максимальной Delta в задаче ранжирования многословных выражений по уровням языкового владения. Соответствие предположений об уровне по уровню максимальной Delta с получившемся в результате экспертной оценки уровнем выражения представлена в Таблице 8.

Таблица 8. Матрица соответствия значений меры Max Delta и оценки выражений экспертами
Table 8. Confusion matrix of the Max Delta and expert annotation values

Уровень после экспертной оценки	**Уровень по мере максимальной Delta**					
	A1	**A2**	**B1**	**B2**	**C1**	**ВНЕ СПИСКА**
A1	57.56%	4.78%		0.45%		0.10%
A2	21.51%	57.42%	12.75%	1.82%	0.70%	1.46%
B1	1.16%	13.88%	46.98%	13.64%	10.18%	3.51%
B2		1.44%	10.74%	56.36%	39.65%	25.85%
C1		0.96%	0.67%	5.91%	34.04%	47.32%
ВНЕ СПИСКА	19.77%	21.53%	28.86%	21.82%	15.44%	21.76%
Всего	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00 %**	**100.00%**

Данные таблицы 8 позволяют увидеть, что для выражений уровней от A1 до B2 максимальная Delta совпадает с уровнем выражения после экспертной оценки в примерно в половине случаев: от 47% до 58%. При этом большая часть несовпадений меры и оценок экспертов относится на +- один уровень CEFR, что является не самой критичной ошибкой. Интересно, что для начальных уровней A1 и A2 более характерна тенденция экспертов присваивать выражениям более высокий уровень, чем Delta, а на уровнях B1-C1 эксперты, наоборот, в случаях несовпадения с дельтой чаще присваивают выражениям более низкий уровень сложности. Для определения выражений уровня C1 мера максимальной дельты показала значительное несовпадение с оценками экспертов. Наконец, самая проблемная зона несовпадения мнений экспертов и меры максимальной Дельты являются случаи, когда выражение крайне редко встречается в учебниках РКИ или не встречается вовсе, из-за чего выражение оказывается в группе «Вне вписка», при этом по результатам обобщенной экспертной оценке 47% таких выражений маркируются уровнем C1, 25% – уровнем B2 и только в 22% случаев эксперты соглашаются с максимальной Delta о нецелесообразности представления выражения в списке.

На основании этих данных можно сделать общий вывод о том, что мера максимальной Delta, полученная на основании встречаемости выражения в учебниках РКИ, может использоваться в качестве предварительной оценки уровня выражения на уровнях A1-B1 с высокой и средней степенью точности и полноты результата. На более высоких уровнях, где растет разнообразие и количество предлагаемых лексических единиц

эффективность меры максимальной Delta пока нельзя назвать удовлетворительной. Причина этого явления может заключаться в том, что на высоких уровнях среди списка кандидатов велика доля фразеологических оборотов, пословиц и поговорок, а также формул, свойственных официально-деловому стилю речи. Изучение этих единиц зачастую происходит с помощью специализированных пособий, поэтому их встречаемость в текстах общего курса русского языка, составляющих корпус RuFoLa может быть низкой или нулевой. Кроме того, несовпадение представлений экспертов о методической ценности выражения и его максимальной Delta в ряде случаев может сигнализировать о недостаточной представленности актуальных многословных выражений в современных учебниках русского языка.

Во-вторых, представленная версия списка позволила приступить к решению практической проблемы вычленения и оценки уровня многословных конструкций при автоматическом анализе сложности текста, однако на данном этапе разработки имеет ряд ограничений. Так, актуальная версия списка многословных выражений и алгоритм их вычленения из текста функционирует только для точных форм, точных лемматизированных форм, а также единичных случаев вариации личных местоимений (у <меня, тебя, него, неё, нас, вас, их> аллергия). Таким образом, более сложные случаи вариативности употребления конструкций (лить <прямо, сегодня, совсем и т.д.> как из ведра, мне <уже, совсем, всё же и т.д.> пора) остаются за пределами настоящего этапа разработки.

Наконец, дальнейшего уточнения требует классификация полученных многословных выражений. Основная трудность здесь заключается в том, что список сочетает в себе крайне разнородные лексические единицы: сложные существительные, дискурсивные формулы, идиомы, этикетные формулы, лингвистическая природа и терминологическая принадлежность многих из них является предметом актуальных научных дискуссий. Возможным выходом здесь может стать ориентация на функцию или ситуацию употребления выражения, а не его лингвистическую природу.

Выводы

В данной статье описан опыт комбинированного подхода к формированию градуированного списка многословных выражений для изучающих русский язык как иностранный. Метод включал формирование первичной базы путем компиляции кандидатов из существующих лексических ресурсов для изучающих РКИ и частотных n-грамм из нескольких корпусов. Оценка уровня выражения осуществлялась на основании его частотности в корпусе учебников РКИ и проверялась с помощью множественной экспертной оценки. Исследование показало, что использованная мера максимальной Delta показывает высокую и среднюю точность и полноту на кандидатах уровней A1-B1, для остальных уровней наблюдалось значительное расхождение меры и обобщенной экспертной оценки педагогической ценности выражения, что говорит о необходимости дальнейшей разработки методики выделения формальных признаков наиболее актуальных многословных выражений.

Прикладным результатом исследования стала первая версия градуированного списка многословных выражений русского языка, наиболее актуальных для представления в иноязычной аудитории общим объемом 1645 единиц. Предложенная версия списка внедрена в систему автоматического анализа сложности текста для изучающих РКИ и может быть полезна широкому кругу профильных специалистов при подготовке учебного контента.

Список использованных источников

Андрюшина Н. П. Лексический минимум по русскому языку как иностран-

ному. Второй сертификационный уровень. Общее владение / Андрюшина Н. П. и др. 9-е изд., испр. и доп. СПб.: Златоуст, 2021. 168 с.

Андрюшина Н. П. Лексический минимум по русскому языку как иностранному. Первый сертификационный уровень. Общее владение / Андрюшина Н. П. и др. 10-е изд. СПб.: Златоуст, 2019. 200 с.

Андрюшина Н. П. Лексический минимум по русскому языку как иностранному. Третий сертификационный уровень. Общее владение / Андрюшина Н. П. и др. 2-е изд. СПб.: Златоуст, 2019. 200 с.

Андрюшина Н. П., Козлова Т. В. Лексический минимум по русскому языку как иностранному. Базовый уровень. Общее владение. 7-е изд. СПб.: Златоуст, 2022. 116 с.

Андрюшина Н. П., Козлова Т. В. Лексический минимум по русскому языку как иностранному. Элементарный уровень. Общее владение. 6-е изд., испр. и доп. СПб.: Златоуст, 2020. 80 с.

Corpus Materials

Andryushina, N. P. (2019). *Leksicheskiy minimum po russkomu yazyku kak inostrannomu. Pervy sertifikatsionniy uroven. Obshhee vladenie* [Lexical minimum in Russian as a foreign language. First certification level. General proficiency], Zlatoust, St. Petersburg, Russia. *(In Russian)*

Andryushina, N. P. (2019). *Leksicheskiy minimum po russkomu yazyku kak inostrannomu. Tretiy sertifikatsionniy uroven. Obshhee vladenie* [Lexical minimum in Russian as a foreign language. Third certification level. General proficiency], Zlatoust, St. Petersburg, Russia. *(In Russian)*

Andryushina, N. P. (2021). *Leksicheskiy minimum po russkomu yazyku kak inostrannomu. Vtoroy sertifikatsionniy uroven. Obshhee vladenie* [Lexical minimum in Russian as a foreign language. Second certification level. General proficiency], Zlatoust, St. Petersburg, Russia. *(In Russian)*

Andryushina, N. P. and Kozlova, T. V. (2020). *Leksicheskiy minimum po russkomu yazyku kak inostrannomu. Elementarniy uroven. Obshhee vladenie* [Lexical minimum in Russian as a foreign language. Elementary level. General proficiency], Zlatoust, St. Petersburg, Russia. *(In Russian)*

Andryushina, N. P. and Kozlova, T. V. (2022). *Leksicheskiy minimum po russkomu yazyku kak inostrannomu. Bazoviy uroven. Obshhee vladenie* [Lexical minimum in Russian as a foreign language. Basic level. General proficiency], Zlatoust, St. Petersburg, Russia. *(In Russian)*

Список литературы

Ерёмина О. С. Русские несвободные выражения в речи иностранцев: корпусный подход // Русский язык за рубежом. 2020. № 6 (283). С. 29-35. DOI: 10.37632/PI.2020.283.6.004

Инькова О. Ю. К вопросу о лемматизации многокомпонентных единиц // Захаров В. П. и др. (ред.). Труды международной конференции «Корпусная лингвистика 2015», СанктПетербург, 22–26 июня 2015 года. СПб.: СПбГУ, 2015. С. 1–10.

Иорданская Л. Н., Мельчук И. А. Смысл и сочетаемость в словаре. М.: Языки славянских культур, 2007. 672 с.

Лапошина А. Н. Корпус текстов учебников РКИ как инструмент анализа учебных материалов // Русский язык за рубежом. 2020. № 6 (283). С. 22–28. DOI: 10.37632/PI.2020.283.6.003

Минаева Е. В. Дискурсивные слова в современной разговорной речи и в учебниках РКИ // Международный аспирантский вестник. 2017. № 2. С. 74–79.

Пужаева С. Ю. Автоматическое извлечение дискурсивных формул из текстов на русском языке / Пужаева С. Ю., Герасименко Е. А., Захарова Е. С., Рахилина Е. В. // Вестн. Новосиб. гос. ун-та. Серия: Лингвистика и межкультурная коммуникация. 2018. Т. 16. № 2. С. 5–18. DOI: 10.25205/1818-7935-2018-16-2-5-18

Свирина Л. О. Формульный язык и уровень иноязычной коммуникативной компетенции // Филология и культура. 2019. №1 (55). С. 97–101.

Шляхов В. И., Саакян Л. Н. Текст в коммуникативном пространстве. М.: Ленанд, 2015. 236 с.

Alfter D. From distributions to labels: A lexical proficiency analysis using learner corpora / Alfter D., Bizzoni Y., Agebjorn A., Volodina E., Pilan I. // Proceedings of the joint workshop on NLP4CALL and NLP for Language Acquisition at SLTC, 2016. № 130. Pp. 1–7.

Bahns J., Eldaw, M. Should We Teach EFL Students Collocations? // System. 1993. Volume 21. № 1. Pp.101–114.

Bybee J. The emergent lexicon // Chicago Linguistic Society. 1998. № 34. Pp. 421–435.

Calzolari N. Towards best practice for multiword expressions in computational lexicons / Nicoletta C., Fillmore C., Grishman R., Ide N., Lenci A., Macleod C., Zampolli A. In Proceedings of LREC 2002. 2002. Pp. 1934–1940.

Christiansen M. H., Chater, N. The Now-or-Never bottleneck: A fundamental constraint on language // Behavioral & Brain Sciences. 2016. Volume 39. Pp. 62–102. https://doi.org/10.1017/S0140525X1500031X

De Cock S. An automated approach to the phrasicon of EFL learners / De Cock S., Granger S., Leech G., Mcenery T. // Learner English on computer. London & New York: Routledge, 1998. Pp. 67–79. https://doi.org/10.4324/9781315841342

Volodina E. SweLL on the rise: Swedish Learner Language corpus for European Reference Level studies / Volodina E., Pilán I., Enström I., Llozhi L., Lundkvist P., Sundberg G., Sandell M. // Proceedings of LREC 2016. Pp. 206–212.

Elman J. L. On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon // Cognitive Science. 2009. № 33. Pp. 547–582. https://doi.org/10.1111/j.1551-6709.2009.01023.x

François T. FLELex: a graded lexical resource for French foreign learners / François T., Gala N., Watrin P., Fairon C. // In the 9th International Conference on Language Resources and Evaluation (LREC 2014). 2014. Pp. 3766–3773.

Janda L. How to build a constructicon in five years: The Russian Example / Janda L., Endresen A., Zhukova V., Mordashova D., Rakhilina E. // The Wealth and Breadth of Construction-Based Research (a thematic issue of Belgian Journal of Linguistics 34). 2020. Pp. 162–175.

Jolsvai H., McCauley S. M., Christiansen M. H. Meaning overrides frequency in idiomatic and compositional multiword chunks // Proceedings of the 35th Annual Conference of the Cognitive Science Society, Austin. 2013. Pp. 692–697.

Kilgarriff A. Corpus-Based Vocabulary lists for Language Learners for Nine Languages / Kilgarriff A., Charalabopoulou F., Gavrilidou M., Johannessen J., Saussan K., Kokkinakis S., Lew R., Sharoff S., Vadlapudi R., Volodina E. // Language Resources and Evaluation Journal. 2014. № 48. Pp. 121–163. https://doi.org/10.1007/s10579-013-9251-2

Kopotev M. CoCoCo: Online Extraction of Russian Multiword Expressions / Kopotev M., Escoter L., Kormacheva D., Pierce M., Pivovarova L., Yangarber R. // The 5th Workshop on Balto-Slavic Natural Language Processing 2015, Hissar. 2015. Pp. 43–45.

Kopotev M., Pivovarova L., Kormacheva D. Constructional generalization over Russian collocations // Mémoires de la Société néophilologique de Helsinki. 2016. Volume Tome C (Collocations Cross-Linguistically). Pp. 121–140.

Kopotev M. Automatic detection of stable grammatical features in n-grams / Kopotev M., Pivovarova L., Kochetkova N., Yangarber R. // Proceedings of the 9th Workshop on Multiword Expressions, Atlanta. 2013. Pp. 73–81.

Lewis M. Implementing the Lexical Approach: Putting Theory into Practice. Hove, England: Language Teaching Publications, 1997. 223 p.

Loukachevitch N., Lashevich G. Multiword expressions in Russian Thesauri RuThes and RuWordNet // Proceedings of the AINL FRUCT 2016 Conference, Saint Petersburg. 2016. Pp. 66–71.

McClelland J. L. Emergence in cognitive science // Topics in Cognitive Science. 2010. Volume 2. №4. Pp. 751–770. https://doi.org/10.1111/j.1756-8765.2010.01116.x

Paquot M., Granger S. Formulaic Language in Learner Corpora // Annual Review of Applied Linguistics. 2012. Volume 32. Pp. 130–149. https://doi.org/10.1017/S0267190512000098

Parmentier Y., Waszczuk J. Representation and parsing of multiword expressions: Current trends (Phraseology and Multiword Expressions 3). Berlin: Language Science Press, 2019. 326 p.

Schmitt N. Formulaic Sequences: Acquisition, processing and use. Amsterdam: John Benjamins Publishing Company, 2004. 304 p.

Wray A. Formulaic sequences in second language teaching: Principles and practice // Applied Linguistics. 2000. Volume 21. № 4. Pp. 463–489. https://doi.org/10.1093/applin/21.4.463

Wray A. Formulaic language and the lexicon. Cambridge, UK: Cambridge University Press, 2002. 348 p.

References

Eremina, O. S. (2020). Russian non-free expressions in the speech of foreigners: a corpus approach, *Russkiy yazyk za rubezhom*, 6 (283), 29-35. DOI: 10.37632/PI.2020.283.6.004 *(In Russian)*

Inkova, O. Ju. (2015). On the question of lemmatization of multi-component units, *Proceedings of the international conference* «*Corpus Linguistics – 2015*», June 22–26, 2015, St. Petersburg, Russia, 1-10. *(In Russian)*

Iordanskaya, L. N. and Melchuk, I. A. (2007). *Smysl i sochetaemost v slovare* [Meaning and combinability in the dictionary], M.: Yazyki slavjanskih kul'tur, Moscow, Russia. *(In Russian)*

Laposhina, A. N. (2020). A corpus of Russian textbook materials for foreign students as an instrument of an educational content analysis, *Russkiy yazyk za rubezhom*, 6 (283), 22-28. DOI: 10.37632/PI.2020.283.6.003 *(In Russian)*

Minaeva, E. V. (2017). Discursive words in modern colloquial speech and in RCT textbooks, *Mezhdunarodniy aspirantskiy vestnik*, 2, 74-79. *(In Russian)*

Puzhaeva, S. Ju., Gerasimenko, E. A., Zakharova, E. S., Rakhilina, E. V. (2018). Automatic extraction of discourse formulas from Russian language texts, *Vestn. Novosib. gos. un-ta. Seriya: Lingvistika i mezhkulturnaya kommunikatsiya*, 16 (2), 5-18. DOI: 10.25205/1818-7935-2018-16-2-5-18 *(In Russian)*

Svirina, L. O. (2019). Formal language and level of foreign language communicative competence, *Filologiya i kultura*, 1 (55), 97-101. *(In Russian)*

Shlyakhov, V. I. and Saakyan, L. N. (2015). *Tekst v kommunikativnom prostranstve* [Text in communicative space], Moscow, Russia. *(In Russian)*

Alfter, D., Bizzoni, Y., Agebjorn, A., Volodina, E. and Pilan, I. (2016). From distributions to labels: A lexical proficiency analysis using learner corpora, *Proceedings of the joint workshop on NLP4CALL and NLP for Language Acquisition at SLTC*, Umeå, Sweden, November 2016, 130, 1-7. *(In English)*

Bahns, J. and Eldaw, M. (1993). Should We Teach EFL Students Collocations? *System*, 21(1), 101-114. (In English)

Bybee, J. (1998). The emergent lexicon, *Chicago Linguistic Society*, 34, 421-435. *(In English)*

Calzolari, N., Fillmore, C., Grishman, R., Ide, N., Lenci, A., Macleod, C. and Zampolli, A. (2002). Towards best practice for multiword expressions in computational lexicons, *Proceedings of LREC 2002*, Las Palmas, Canary Islands – Spain, May 2002, 1934-1940. *(In English)*

Christiansen, M. H. and Chater, N. (2016). The Now-or-Never bottleneck: A fundamental constraint on language, *Behavioral & Brain Sciences*, 39, 62-102. https://doi.org/10.1017/S0140525X1500031X *(In English)*

De Cock, S., Granger, S., Leech, G. and Mcenery, T. (1998). An automated approach to the phrasicon of EFL learners, *Learner English on computer*, 67-79. https://doi.org/10.4324/9781315841342 *(In English)*

Volodina, E., Pilán, I., Enström, I., Llozhi, L., Lundkvist, P., Sundberg, G. and Sandell, M. (2016). SweLL on the rise: Swedish Learner Language corpus for European Reference Level studies, *Proceedings of LREC 2016*, Portorož, Slovenia, May 2016, 206-212. *(In English)*

Elman, J. L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon, *Cognitive Science*, 33, 547-582. https://doi.org/10.1111/j.1551-6709.2009.01023.x *(In English)*

François, T., Gala, N., Watrin, P. and Fairon, C. (2014). FLELex: a graded lexical resource for French foreign learners, *The 9th International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Iceland, 26-31 May 2014, 3766–3773. *(In English)*

Janda, L., Endresen, A., Zhukova, V., Mordashova, D. and Rakhilina, E. (2020). How to build a constructicon in five years: The Russian Example, *The Wealth and Breadth of Construction-Based Research (a thematic issue of Belgian Journal of Linguistics)*, 34, 162-175. *(In English)*

Jolsvai, H., McCauley, S. M. and Christiansen, M. H. (2013). Meaning overrides frequency in idiomatic and compositional multiword chunks, Proceedings of the 35th Annual Conference of the Cognitive Science Society, Austin, Tx, January 2013, 692–697. *(In English)*

Kilgarriff, A., Charalabopoulou, F., Gavrilidou, M., Johannessen, J., Saussan, K.,

Kokkinakis, S., Lew, R., Sharoff, S., Vadlapudi, R. and Volodina, E. (2014). Corpus-Based Vocabulary lists for Language Learners for Nine Languages, *Language Resources and Evaluation Journal*, 48, 121-163. https://doi.org/10.1007/s10579-013-9251-2 *(In English)*

Kopotev, M., Escoter, L., Kormacheva, D., Pierce, M., Pivovarova, L. and Yangarber, R. (2015). CoCoCo: Online Extraction of Russian Multiword Expressions, *The 5th Workshop on Balto-Slavic Natural Language Processing 2015*, Hissar, Bulgaria, September 2015, 43-45. *(In English)*

Kopotev, M., Pivovarova, L. and Kormacheva, D. (2016). Constructional generalization over Russian collocations, *Mémoires de la Société néophilologique de Helsinki*, Tome C (Collocations Cross-Linguistically), 121-140. *(In English)*

Kopotev, M., Pivovarova, L., Kochetkova, N. and Yangarber, R. (2013). Automatic detection of stable grammatical features in n-grams, *Proceedings of the 9th Workshop on Multiword Expressions*, Atlanta, GA, June 2013, 73-81. *(In English)*

Lewis, M. (1997). *Implementing the Lexical Approach: Putting Theory into Practice*, Language Teaching Publications, Hove, England. *(In English)*

Loukachevitch, N. and Lashevich, G. (2016). Multiword expressions in Russian Thesauri RuThes and RuWordNet, *Proceedings of the AINL FRUCT 2016 Conference*, Saint Petersburg, Russia, December 2016, 66-71. *(In English)*

McClelland, J. L. (2010). Emergence in cognitive science, *Topics in Cognitive Science*, 2 (4), 751-770. https://doi.org/10.1111/j.1756-8765.2010.01116.x *(In English)*

Paquot, M. and Granger, S. (2012). Formulaic Language in Learner Corpora, *Annual Review of Applied Linguistics*, 32, 130–149. https://doi.org/10.1017/S0267190512000098 *(In English)*

Parmentier, Y. and Waszczuk, J. (2019). *Representation and parsing of multiword expressions: Current trends (Phraseology and Multiword Expressions 3)*, Language Science Press, Berlin, Germany, 326. *(In English)*

Schmitt, N. (2004). *Formulaic Sequences: Acquisition, processing and use*, John Benjamins Publishing Company, Amsterdam, Netherlands, 304. *(In English)*

Wray, A. (2000). Formulaic sequences in second language teaching: Principles and practice, *Applied Linguistics*, 21(4), 463-489. https://doi.org/10.1093/applin/21.4.463 *(In English)*

Wray, A. (2002). *Formulaic language and the lexicon*, Cambridge University Press, Cambridge, UK. *(In English)*

Все авторы прочитали и одобрили окончательный вариант рукописи.

All authors have read and approved the final manuscript.

Конфликты интересов: у авторов нет конфликта интересов для декларации.

Conflicts of Interest: the authors have no conflict of interest to declare.

Лапошина Антонина Николаевна, кандидат педагогических наук, научный сотрудник лаборатории когнитивных и лингвистических исследований ФГБОУ ВО «Государственный институт русского языка им. А.С. Пушкина», Москва, Россия.
Antonina N. Laposhina, PhD in Education, Research Fellow, Pushkin State Russian Language Institute, Moscow, Russia.

Храмченко Татьяна Александровна, старший преподаватель кафедры теории и методики преподавания РКИ Белорусского государственного университета, Минск, Беларусь.
Tatsiana A. Khramchanka, Senior Lecturer, Belarusian State University, Minsk, Belarus.

Лебедева Мария Юрьевна, кандидат филологических наук, зав. лабораторией когнитивных и лингвистических исследований, ФГБОУ ВО «Государственный институт русского языка им. А.С. Пушкина», Москва, Россия.
Maria Yu. Lebedeva, PhD in Philology, Head of the Language and Cognition Laboratory, Pushkin State Russian Language Institute, Moscow, Russia.