А. Н. Лапошина, М. Ю. Лебедева

DOI: 10.24412/1811-1629-2022-3-90-99

ФОРМИРОВАНИЕ ЧАСТОТНОГО СЛОВАРЯ-МИНИМУМА РУССКОГО ЯЗЫКА ДЛЯ ДЕТЕЙ-ИНОФОНОВ НА ОСНОВЕ КОРПУСНЫХ ДАННЫХ

ANTONINA N. LAPOSHINA, MARIA IU. LEBEDEVA
DEVELOPING A RUSSIAN FREQUENCY CORE VOCABULARY LIST FOR FOREIGN CHILDREN BASED ON CORPUS DATA

Антонина Николаевна Лапошина

Ведущий эксперт

► antonina.laposhaina@gmail.com

Мария Юрьевна Лебедева

Кандидат филологических наук, ведущий научный сотрудник

► m.u.lebedeva@gmail.com

Государственный институт русского языка им. А. С. Пушкина

117485, Россия, Москва, ул. Академика Волгина, 6

Antonina N. Laposhina, Maria Iu. Lebedeva

Pushkin State Russian Language Institute

6, ul. Akademika Volgina, Moscow, Russia, 117485

Статья посвящена проблеме формирования списка наиболее употребительных слов русского языка для детей-инофонов младшего школьного возраста. Проблема минимизации и оптимизации входного языкового материала крайне актуальна для иностранных учащихся. Лексические списки наиболее употребительной и актуальной лексики способны решить эту проблему, предлагая информацию о лексических единицах, которые наиболее вероятно встретятся данной аудитории студентов и которые, следовательно, целесообразно включать в пособия и другие учебные материалы для этой группы учащихся. Однако в настоящий момент наблюдается дефицит актуальных лексических списков для аудитории младшего школьного возраста, изучающих русский язык как иностранный. Настоящая статья ставит целью заполнение этой лакуны. Предлагаемые списки сформированы на основе объединения информации о частотности слова по нескольким релевантным источникам: коллекции детской литературы, коллекции учебников русского языка как родного и как иностранного. Для расчета единой меры востребованности слова были произведены три типа вычислений: среднее значение ipm, среднее значение ранга Ципфа и авторская формула, учитывающая, помимо частотности, фактор равномерности распределения слова по разным сегментам корпуса. Для проверки качества работы списков мы проанализировали процент покрытия ими коллекции целевых текстовых образцов: художественной литературы для детей, детских журналов, транскриптов мультфильмов. Результаты проверки показали, во-первых, что разработанные списки в среднем охватывают на 9% больше лексики из целевой коллекции, во-вторых, что лучшие результаты показали списки, сформированные по авторской формуле расчета востребованности слова. Разработанный список наиболее употребительных слов, сформированный на основе корпусных данных, может быть использован авторами и редакторами учебного контента для детской иноязычной аудитории младшего школьного возраста, а также как основа для составления учебного словаря-минимума для данной аудитории учащихся.

Ключевые слова: лексический минимум; отбор лексических единиц; частотность; корпусная лингводидактика; русский язык как иностранный.

The article is devoted to the problem of creating a list of the most common Russian words for primary school students learning Russian as a foreign language. The problem of minimizing and optimizing the language input is extremely important in foreign language acquisition studies. Lists of the most commonly used and relevant vocabulary can solve this problem by providing information about lexical units that this group of students is most likely to encounter. At the moment there is a lack of up-to-date lexical lists for an audience of primary school age who study Russian as a foreign language. This article aims to fill this gap. The proposed vocabulary lists are

compiled by combining information about the frequency of a word from several relevant sources: the corpus of children's literature, and the corpus of textbooks of Russian as a native language and as a foreign language. To calculate an aggregated value of a word from these sources, three types of calculations were made: the average value of ipm, the average value of the Zipf rank, and the author's formula, which takes into account, in addition to frequency, the factor of even distribution of the word over different segments of the corpora. The quality of the resulting lists has been confirmed by the text coverage of the collection of target text samples: fiction for children, children's periodicals, and cartoon transcripts. The developed list of the most commonly used words can be used by authors and editors of educational content.

Keywords: vocabulary selection; vocabulary list; Russian as a foreign language; word frequency; corpus-based language teaching.

Разработка материалов для обучения языку всегда требует многократного и тщательного изучения его современного состояния, необходимого, чтобы взвешенно принимать решение о включении тех или иных лексических единиц в учебные материалы. Вопрос отбора и количества исходного языкового материала для детской аудитории представляется особенно важным: согласно исследованиям, качество "входного материала" (англ. *input*) оказывает на прогресс в освоении языка в детской аудитории большее влияние, нежели возраст начала занятий [Muñoz 2014].

Одним из наиболее распространенных и эффективных подходов к отбору лексики является опора на данные о частотности слова по релевантной коллекции текстов. Основная идея подхода, предполагающего учет частотности слова, в контексте преподавания языка связана с предположением, что частотные слова более вероятно будут встречаться учащимся в аутентичных материалах, поэтому целесообразнее всего изучить их в первую очередь [Nation, Waring 1997]. Основываясь на этом предположении, исследователи направляли усилия на вычисление способности лексики определенной частотности покрывать тексты на разных языках (см., например, [Hsueh-Chao, Nation 2000]. По немногочисленным имеющимся данным на материале русского языка, знание примерно 2500 тысяч наиболее частотных слов русского языка даёт возможность понимать 80% слов любого текста [Фрумкина 1967].

В контексте нашего исследования особенно интересно, что первые частотные лексические списки на русском языке разрабатывались именно для детей. Так, первый опыт создания частотного словаря русского языка на основе сплошной выборки слов из отобранной коллекции текстов был связан с проблемами преподавания русского языка в национальных школах. Частотный словарь Э. А. Штейнфельд был составлен на основе статистических подсчетов встречаемости слов в коллекции объемом 400 тысяч слов, в которую входили образцы оригинальной (А. П. Гайдар, Н. Н. Носов) и переводной (Марк Твен, Андерсен) детской художественной литературы, русской классикой литературы, молодежных газет и журналов и материалов радиопередач для молодежи, и был задуман как основа для составления словаря-минимума для начальной эстонской школы [Штейнфельдт 1963].

Учет частотной информации широко используется для отбора лексических единиц в современной методике РКИ как для общего курса русского языка [Sharoff et al. 2013], так и для языка специальности [Ильина 2013; Сидорова, Шматко 2019] Однако примеров подобных современных словарей для детской аудитории нам найти не удалось.

Стоит однако отметить, что ориентация исключительно на частотность слова при отборе лексики имеет серьёзные риски. В первую очередь, риски обусловлены тем, что в имеющихся корпусах слабо представлена бытовая сфера общения, что приводит к низкой частотности элементарных и ценных с точки зрения преподавания языка лексем: *апельсин, автобус, зубная щетка* и др. [Volodina et al. 2013]. Поэтому традиционно решение о включении того или иного слова в словарь-минимум определяется комплексом параметров, среди которых, помимо частотности, словообразовательная ценность слова, его сочетаемостный потенциал, количество значений, стилистическая нейтральность, и наконец, методическая ценность слова [Андрюшина 2011; Маркина 2011].

Для формализации понятия методической ценности слова некоторые исследователи предлагают использовать информацию о частотности слова в учебной литературе для заданной ау-

дитории. Предполагается, что учет обобщенной информации о частотности слова по большому количеству пособий отражает коллективную комплексную экспертную оценку методической ценности слова для данной аудитории [Volodina et al. 2013]. Иными словами, если слово часто и регулярно появляется в разных пособиях РКИ для детей, значит, это слово обладает методической ценностью для этой категории учащихся.

На наш взгляд, использование любой коллекции текстов в качестве единственного источника информации о востребованности слова несет свои риски. Опора исключительно на накопленный опыт преподавателей РКИ в детской аудитории, отраженный в пособиях для этой группы учащихся, описывает современное состояние области, однако ставит нас в методический тупик, не давая возможности предложить потенциально недооцененную в имеющихся пособиях лексику. Построение словника на основе частотности по корпусу детской литературы приведет к дефициту бытовой элементарной лексики (т. н. *проблема зубной пасты* [Volodina et al. 2013]). Наша гипотеза состоит в том, что составление единого частотного списка, учитывающего информацию о слове из нескольких релевантных источников поможет предложить объективные основания для отбора лексики в список.

Целью настоящей работы является проверка экспериментальной методики формирования лексического списка наиболее востребованной лексики русского языка для детей-инофонов младшего школьного возраста на основе объединения частотной информации из разных источников. Таким образом, предполагается найти баланс между употребительностью слова в целевой коллекции текстов детской литературы и методической ценностью слова с точки зрения преподавания РКИ.

В качестве **материалов исследования** для получения информации о частотности слова из разных источников были использованы несколько коллекций текстов, предназначенных для чтения детьми младшего школьного возраста. Обзор источников представлен в таблице 1. Одним из основных источников стал TIRTEC — корпус текстов учебников русского языка для детей младшего школьного возраста (7–11 лет) с разным уровнем владения русским языком. Наиболее близким к целевой аудитории списков является подкорпус текстов из учебников русского языка для детей-инофонов, TIRTEC-foreign. В него вошли тексты популярных пособий для изучения русского языка как иностранного (комплексы “Сорока”, “Жар-птица”, “Дом”, “Русский язык шаг и шагом” и мн. др)[1]. Учебники для детей-билингвов[2] (TIRTEC-bilingual) и учеников российских школ (TIRTEC-native) также могут содержать актуальную лексику для детей выбранного возраста.

ДетКорпус — это аннотированный корпус русской литературы для детей, включающий бо-

Название	**Состав**	**Объем в словах**
TIRTEC-foreign	учебники русского языка для детей-инофонов 7–11 лет	211 009
TIRTEC-bilingual	учебники русского языка для детей-билингвов, детей с семейным / вторым русским 7–11 лет	555 712
TIRTEC-native	учебники русского языка для 1–4 класса российских школ	908 469
ДетКорпус	корпус литературы для детей	72 380 135
RuFoLa	учебники русского языка как иностранного для взрослых учащихся, уровни A1–B1	126 223

Таблица 1. **Описание корпусов, использованных в качестве источников данных о частотности**

лее 2097 прозаических произведений, написанных на русском языке в период с 1920-х по 2010-е годы и адресованных детям и подросткам. Корпус содержит как художественные тексты различных жанров (реализм, приключения, детектив, ужастик), так и тексты нон-фикшн. Частотность слова по этому корпусу может показать востребованность слова для чтения текстов, адресованных детям.

Корпус Rufola (Russian as a Foreign Language Corpus) содержит тексты из пособий по РКИ для взрослых учащихся и отражает возможную методическую ценность слова с позиции преподавания русского языка как иностранного. Для данной задачи был выбран фрагмент корпуса с текстами уровней A1–B1. Всего по всем предложенным спискам насчитывается 14 253 уникальных лексемы, на материале которых были произведены все последующие расчеты.

Обработка текстов коллекций включала в себя этапы токенизации (деления текста на отдельные слова), лемматизации (приведения каждого слова к его словарной форме), чистки от ударений и спецсимволов, присвоения каждой лексеме частеречного тэга и, наконец, подсчет количества раз, которое слово встретилось в той или иной коллекции. Приведение слов в начальной форме и присвоение им частеречных тэгов было сделано на основе программной библиотеки pymystem3 с небольшими авторскими доработками.

Все формы глагола, включая причастия и деепричастия и видовые пары, были приведены к форме несовершенного вида *(рассказанный, рассказать, расскажи приводится к форме рассказывать)*. Снятие омонимии по грамматическим признакам также происходило автоматически, на основании контекста слова. Так, в списках отдельно значатся *богатый* как прилагательное и как существительное, *лев* как животное и как имя собственное и т. д. Однако разные значения слова внутри одной части речи (*лук*-овощ, *лук*-оружие), крайне затруднительные для автоматизированного различения, не выделялись. Все формы глагола, включая причастия и деепричастия и видовые пары, были приведены к форме несовершенного вида (*рассказанный*, *рассказать, расскажи* приводится к форме *рассказывать*).

Меры сравнения частотности слова

Выбранные коллекции текстов сильно разнятся по объему (см. табл. 1). Чтобы это не приводило к искажению данных о встречаемости слова, была проведена нормализация, уравновешивающая значения частотности из корпусов разного объема.

Одной из базовых нормализованных метрик является *относительная частотность слова, ipm* (instances per million), которая вычисляется как отношение количества раз, которое слово встретилось в корпусе к размеру корпуса в миллионах слов [Kilgariff 2012]. Например, слово *мама* встретилось в учебниках для детей-инофонов 864 раза, а в коллекции литературы для детей — 95 241 раз. Однако очевидно, что наблюдаемая разница в цифрах напрямую зависит от разницы в объеме этих коллекций текстов. Подсчет меры ipm делает эти величины сравнимыми между собой: 864 / 0.211 = 4090 ipm для детей-инофонов и 95241 / 72 = 1323 ipm для детской литературы. Таким образом, эта мера позволяет увидеть, что в действительности *мама* появляется в учебниках для детей-иностранцев в 3 раза чаще.

При сборе информации из нескольких источников важным показателем является не только собственно частотность, но и “универсальность” слова, его стилистическая и тематическая нейтральность. Подобную информацию способна проиллюстрировать **мера R** (range), которая отражает количество сегментов корпуса, в которых встретилось слово. Она вычисляется как отношение количества корпусов, в которых хотя бы раз встретилось это слово, к общему количеству корпусов в коллекции и умноженное на 100 [Ляшевская, Шаров 2009]. Чем в большем количестве частей коллекции появляется слово, тем больше значение меры R. Коэффициент R 100 означает, что слово встречается хотя бы один раз во всех учебниках коллекции: например, *текст, девочка, ухо, добрый*. Высокая частотность слова в сочетании с низким коэффициентом R помогает подсветить уникальные слова, возможно, составляющие концепцию авторов учебника: например, имена персонажей или излюбленные темы автора.

Мера DP (degree of dispersion) более детально характеризует степень равномерности распределения встречаемости слова в разных фрагментах корпуса [Gries 2008]. Причем здесь полезен расчет равномерности распределения частоты как среди разных учебников внутри одного подкорпуса, так и по целым подкорпусам. Например, слово *дуб* встречается в 50% учебников РКИ для детей и имеет коэффициент R равный 50. Однако в одних учебниках оно встречается от 1 до 7 раз, а в одной линейке пособий — 40 раз. Это приводит к DP равному 0.67 и сигнализирует о неравномерно распределенной встречаемости слова.

Таблица 2 иллюстрирует степень разнородности данных о частотности слов в зависимости от коллекции текстов, на материале которых ведутся подсчеты. Так, виден пласт лексики, характерной для детских учебников русского языка — она связана в первую очередь с процессом обучению языку (*пословица, слово, буква, записать*) и лексикой природной тематики (*дуб, береза, заяц*). Некоторые лексемы, мотивированные специфическим грамматическим материалом (*доходить-дойти*), значительно выделяются по частотности в учебниках РКИ для детей. Существует также пласт лексики, характерный для детской литературы, скудно представленный в учебниках (*однако*). В текстах учебников РКИ для взрослых значительно выше показатели частотности лексем, связанных с политической и финансовой сферой (*государство, зарплата, экономика*).

Создание сводного списка

Проблемы методики создания сводных списков из нескольких источников уже ставились исследователями на материале русского языка. Так, например, в системе лексических минимумов под ред. В. В. Морковкина для создания сводного списка самых употребительных слов русского языка на материале 8 частотных списков использовался усредненный “индекс употребительности слова” (номер слова в списке по убыванию абсолютной частотности) [Богачева и др. 2003]. Другим решением могло бы стать создание сводного списка на основании среднего значения относительной частотности ipm [Francois, Fairon 2012] или её модифицированной версии, где вместо значений ipm используется более узкая шкала рангов частотности Ципфа от 0 до 8 [Blinova 2020]. Однако после анализа материала нам стала очевидна необходимость учитывать при отборе лексики еще и параметр равномерности распределения частотности по разным учебникам РКИ, чтобы избежать очевидных авторских предпочтений

Лемма	**Относительная частотность леммы, ipm**				
	TIRTEC-foreign	**TIRTEC_bilingual**	**TIRTEC-native**	**Деткорпус**	**RuFoLa**
в	254797	30647	31623	22378	34277
папа	3194	1391	278	689	1091
пословица	1467	4384	521	7	0
плохой	194	173	56	177	311
доходить	5356	59	33	136	88
дуб	298	204	255	39	0
банан	355	61.26	20	12	40
государство	9	64.86	19	28	119
однако	4.5	29	54	259	72

Таблица 2. **Примеры значений относительной частотности ipm в зависимости от корпуса**

и особенностей сюжетной канвы учебника. Для этого была разработана величина Lex Value, которая рассчитывается по формуле (1),

(1) Lex value = ZIPF_rki * 2 + MEAN_ZIPF + (1 — DP_rki)

где Lex value — это коэффициент ценности слова, ZIPF_rki — это частотность слова в коллекции текстов пособий РКИ для детей, приведенная к ранговому значению, MEAN_ZIPF — это среднее значение ранга частотности слова по остальным коллекциям (TIRTEC-bilingual, TIRTEC-native, RuFoLa и Деткорус) и DP_rki — это степень равномерности распределения частотности слова внутри всех пособий коллекции для детей-инофонов.

В результате были созданы три варианта сводных списка исходя из методики подсчета единой частотной ценности слова: среднее значение IPM по всем пяти источникам, средний ранг Ципфа по всем пяти источникам и предложенная нами мера Lex value. В случае одинакового значения коэффициента у нескольких слов, их позиция в списке определялась по суммарной абсолютной частотности слова.

Покрытие новыми списками целевых текстов

Основным критерием качества лексического списка является процент покрытия ими релевантной коллекции текстов [Маркина 2011; Сидорова, Шматко 2019; Sharoff et al. 2013]. Этот параметр показывает, насколько лексика из списков отражает реальный языковой материал, с которым сталкиваются учащиеся, и следовательно, насколько эффективно подготавливает к общению на русском языке. Для проверки работы полученных списков мы отобрали два пособия из одной серии УМК «Сорока» разных уровней сложности, лексику из разделов «Чтение» и «Лексико-грамматический тест» тренировочных тестов по русскому языку для иностранных школьников Государственного института русского языка им. А. С. Пушкина. Примерами художествен-

Коллекция текстов	**Покрытие текста лексическими списками, %**			
	Способ получения списка объемом 1 000 слов			**ЛМ ТРКИ А1**
	Среднее значение ipm	**Среднее значение коэффициента Ципфа**	**Lex_value**	
Сорока-1	86	86	89	84
Сорока-3	76	77	82	76
Тест РКИ-дети А1	85	86	87	85
Тест РКИ-дети А2	83	82	85	77
Денискины рассказы	72	72	72	60
Гарри Поттер	64	62	61	51
Мурзилка. Письма читателей	62	62	61	55
Мурзилка. Статьи	54	55	53	42
Три кота, м/с	75	75	74	60
Смешарики, м/с	74	74	73	60
Среднее итоговое значение	**73.1**	**73.1**	**73.7**	**65**

Таблица 3. **Покрытие образцов релевантных для детской иноязычной аудитории текстов полученными лексическими списками**

ной литературы является фрагменты по 3 000 слов из книги В. В. Драгунского «Денискины рассказы» и Дж. Роулинг «Гарри Поттер и философский камень». Детскую публицистику мы представили двумя коллекциями текстов из журнала «Мурзилка» 2019–2021 гг., первая из которых включает письма читателей, а вторая — статьи из журнала научно-популярной тематики. Образцы дискурса мультсериалов представлены м/с «Три кота» и «Смешарики» объемом по 30 транскриптов случайно отобранных серий.

Покрытие текста списком представляет собой процент слов текста, присутствующих в лексическом списке. Имена собственные и географические названия были удалены из текста перед подсчетами. Все три варианта полученных списков были ограничены 1 000 слов. В качестве дополнительного источника сравнения мы также подсчитали покрытие текста существующим лексическим минимумом системы ТРКИ уровня А1, ориентированного на взрослую аудиторию. Данный список для корректной работы автоматической подсчета лексики был расширен дериватами

№	слово	№	слово	№	слово	№	слово
1	мама	26	урок	51	телефон	76	телевизор
2	день	27	язык	52	учитель	77	молоко
3	друг	28	сестра	53	страна	78	музей
4	дом	29	спасибо	54	неделя	79	билет
5	год	30	дедушка	55	чай	80	театр
6	папа	31	семья	56	хлеб	81	рубль
7	школа	32	рука	57	стул	82	парк
8	город	33	раз	58	сад	83	**мяч**
9	бабушка	34	место	59	цветок	84	карандаш
10	ребенок	35	вода	60	лето	85	компьютер
11	человек	36	комната	61	рыба	86	яблоко
12	время	37	минута	62	игра	87	рождение
13	**ребята**	38	ночь	63	магазин	88	сок
14	слово	39	улица	64	подруга	89	футбол
15	девочка	40	работа	65	**кот**	90	фрукт
16	мальчик	41	класс	66	врач	91	глаз
17	стол	42	**лес**	67	цвет	92	голова
18	окно	43	дядя	68	ручка	93	дело
19	машина	44	**двор**	69	кошка	94	дверь
20	час	45	вопрос	70	фотография	95	нога
21	вечер	46	дерево	71	ученик	96	отец
22	утро	47	тетя	72	подарок	97	земля
23	брат	48	море	73	привет	98	свет
24	книга	49	сын	74	праздник	99	дорога
25	собака	50	гость	75	автобус	100	деньги

Таблица 4. **100 самых употребительных существительных в коллекции текстов для младших школьников**

[Лапошина 2021] и составляет 970 слов, т. е. сравним по объему с полученными списками.

Становится очевидна разница между «взрослым» ЛМ ТРКИ и разработанными детскими списками: вне зависимости от методики создания сводного списка, детские списки показывают больший процент покрытия целевых текстов (табл. 3). Это ещё раз доказывает некорректность определения уровня лексической сложности текста для данной аудитории с помощью лексических минимумов ТРКИ и необходимость дальнейших методических разработок для детской аудитории.

Самый большой процент покрытия текстов ожидаемо демонстрирует учебные и контрольные материалы по РКИ, представленные пособием «Сорока» для детей-инофонов и тренировочными тестами для данной группы учащихся. Материалы журнала «Мурзилка» содержат наименьшее количество знакомых слов из списков.

Интересно, что предложенный нами способ объединения списка Lex value показывает лучший результат по всем коллекциям, связанным с преподаванием РКИ. При подсчетах же на аутентичных текстах чуть лучший результат показыва-

№	слово	№	слово	№	слово	№	слово
1	быть	26	приходить	51	ждать	76	забывать
2	сказать	27	посмотреть	52	лежать	77	**вставать**
3	говорить	28	рассказывать	53	показать	78	**входить**
4	хотеть	29	помогать	54	слушать	79	**удивляться**
5	идти	30	ходить	55	спать	80	услышать
6	давать	31	работать	56	находиться	81	**принимать**
7	видеть	32	писать	57	**искать**	82	**считать**
8	смотреть	33	ехать	58	**приезжать**	83	**приносить**
9	делать	34	поехать	59	написать	84	брать
10	жить	35	заниматься	60	**начинаться**	85	повторять
11	любить	36	нравиться	61	звонить	86	**уметь**
12	читать	37	учиться	62	отдыхать	87	встречать
13	играть	38	есть	63	учить	88	**упасть**
14	звать	39	купить	64	**болеть**	89	**уезжать**
15	мочь	40	прочитать	65	ездить	90	**проводить**
16	знать	41	пить	66	подарить	91	**собирать**
17	**стать**	42	гулять	67	готовить	92	прийти
18	спрашивать	43	понимать	68	рисовать	93	позвонить
19	пойти	44	**выходить**	69	покупать	94	**встречаться**
20	стоять	45	начинать	70	**кататься**	95	разговаривать
21	думать	46	сделать	71	**оставаться**	96	стоить
22	отвечать	47	**уходить**	72	узнать	97	**висеть**
23	сидеть	48	**находить**	73	**садиться**	98	петь
24	увидеть	49	решать	74	**собираться**	99	**зайти**
25	взять	50	**проходить**	75	**слышать**	100	**выбирать**

Таблица 5. **100 самых употребительных глаголов в коллекции текстов для младших школьников**

ют классические меры, усредненное значение IPM или средний ранг Ципфа. Таким образом, можно сделать осторожный вывод, что полученный с помощью меры Lex value список оптимальнее отражает лексику с учетом специфики преподавания иностранного языка. Впрочем, такая малая разница требует дальнейшего уточнения и анализа.

В результате подсчетов в качестве совокупной меры была выбрана методика формулы Lex value, показавшая наивысший процент покрытия текстов. Полученный список представляет собой 5 700 лексических единиц, расставленных в порядке их употребительности в целевых коллекциях текстов. Каждое слово списка снабжено информацией о его месте в списке, части речи и информацией по его частотности в разных сегментах корпуса. Для иллюстрации приведем в статье список 100 самых употребительных существительных (табл. 4) и глаголов (табл. 5). Жирным шрифтом отмечены единицы, которые не входят в традиционный лексический минимум ТРКИ уровня А1.

Данные таблиц 4 и 5 достаточно ярко демонстрируют наиболее актуальные для данной аудитории темы, связанные с семьей, учебой и общением со сверстниками: обозначения членов семьи, учебная и школьная лексика, дружба и игры, природа и животные. Даже на уровне максимально употребительных слов русского языка становится видна разница со «взрослыми» лексическими минимумами ТРКИ, в которые не входят такие слова как *ребята, кот, лес, двор*. В случае с самыми частотными глаголами разница становится заметно сильнее: почти 30% лексики из получившегося списка 100 глаголов отсутствует в лексическом минимуме уровня А1. Среди них можно выделить как лексику, тематически ценную для преподавания в детской аудитории (*кататься, упасть, уметь*), так и группы глаголов движения (*уходить, выходить, проходить, приезжать, уезжать*). Это говорит о том, что данная лексика активно отрабатывается в пособиях и часто встречается в детской литературе. Однако очевидно, что её включение или невключение в конкретное пособие требует от автора комплексного методического решения, связанного с этапами освоения грамматики русского языка.

Заключение

Предложенный список наиболее употребительных слов, сформированный на основе корпусных данных, может рассматриваться в качестве ориентира при составлении пособий и другого учебного контента для детской иноязычной аудитории младшего школьного возраста, а также как основа для составления учебного словаря-минимума для данной аудитории учащихся.

ПРИМЕЧАНИЯ

[1] Полный библиографический список пособий, включенных в корпус, а также все описанные ниже лексические списки доступны на странице проекта: https://digitalpushkin.tilda.ws/tirtec

[2] Авторы признают размытость этого термина и обозначают им широкую группу пособий для детей со вторым/семейным/неродным/эритажным русским, изучающим русский язык вне языковой среды.

ЛИТЕРАТУРА

Андрюшина 2011 — Андрюшина Н. П. Лексические минимумы по русскому языку как иностранному: проблема отбора лексических и фразеологических единиц. *Проблемы истории, филологии, культуры*. 2001, 3 (33): 648–652.

Богачёва и др. 2003 — Богачева Г. Ф., Луцкая Н. М., Морковкин В. В., Попова З. П. *Система лексических минимумов современного русского языка: 10 лексических списков от 500 до 5000 самых важных русских слов*. М.: АСТ; Астрель, 2003. 768 с.

Ильина 2013 — Ильина О. А. Лексический минимум по языку специальности «Робототехника» как основы формирования лингвокоммуникативной компетенции иностранных магистрантов. *Гуманитарный вестник*. 2013, 2 (4): 1–16.

Лапошина 2021 — Лапошина А. Н. Что значит «не входит в лексический минимум»? Подсчет процента незнакомой лексики в тексте по РКИ с учетом доступных словообразовательных моделей. *Преподаватель XXI век*. 2021, (4 (2)): 473–483.

Ляшевская, Шаров 2009 — Ляшевская О. Н., Шаров С. А. *Частотный словарь современного русского языка (на материалах Национального корпуса русского языка)*. М.: Азбуковник, 2009.

Маркина 2011 — Маркина Е. И. Лингводидактические основы разработки лексических минимумов по русскому языку как иностранному (для разных уровней и профилей обучения). Дис. ... канд. пед. наук. М., 2011. 235 с.

Сидорова, Шматко 2019 — Сидорова М. Ю., Шматко А. С. От «Лексического минимума» к «Лексико-грамматической основе»: новый подход к представлению языка предметной области. *Мир русского слова*. 2019, (3): 83–91.

Фрумкина 1967 — Фрумкина Р. М. Словарь-минимум и понимание текста. *Русский язык за рубежом*. 1967, (2): 15–21.

Штейнфельдт 1963 — Штейнфельдт Э. А. *Частотный словарь современного русского литературного языка: 2500 наиболее употребительных слов: Пособие для преподавателей рус. яз.* Таллин, 1963. 316 с.

Blinova et al. 2020 — Blinova O. V., Tarasov N. A., Modina V. V., Blekanov I. S. Modeling Lemma Frequency Bands for Lexical Complexity Assessment of Russian Texts. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2020"*. 2020, 19 (26): 76–92.

Francois, Fairon 2012 — Francois T., Fairon C. An 'AI readability' formula for French as a foreign language. *Proceedings of the EMNLP and CoNLL 2012, Jeju Island, Korea, 12–14 July 2012*. P. 466–477.

Gries 2008 — Gries S. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*. 2008, (13): 403–437.

Kilgarriff 2012 — Kilgarriff, A. Getting to know your corpus. In *International conference on text, speech and dialogue.* Berlin, Heidelberg: Springer. 2012. P. 3–15.

Muñoz 2014 — Muñoz C. Contrasting Effects of Starting Age and Input on the Oral Performance of Foreign Language Learners. *Applied Linguistics*. 2014, (35): 463–482.

Nation, Waring 1997 — Nation P., Waring R. Vocabulary Size, Text Coverage and Word Lists. In *Vocabulary: Description, Acquisition, and Pedagogy*. Editors Schmitt N., McCarthy M. Cambridge: Cambridge University Press, 1997. P. 6–19.

Sharoff et al. 2013 — Sharoff S. A.; Umanskaya E.; Wilson J. *A Frequency Dictionary of Russian: Core vocabulary for learners.* New York: Routedge, 2013.

Volodina et al. 2013 — Volodina E., Pijetlovic D., Pilán I., Johansson S. Towards a gold standard for Swedish CEFR-based ICALL. *Proceedings of the second workshop on NLP for computer-assisted language learning at NODALIDA*. 2013. P. 48–65.

REFERENCES

Андрюшина 2011 — Andrjushina N. P. Lexical Minima in Russian as a Foreign Language: The Problem of Selecting Lexical and Phraseological Units. *Problemy istorii, filologii, kultury.* 2011, 3 (33): 648–652. (In Russian)

Богачева и др. 2003 — Bogacheva G. F., Luckaya N. M., Morkovkin V. V., Popova Z. P. *The System of Lexical Minima of the Modern Russian Language: 10 Lexical Lists from 500 to 5000 of the Most Important Russian Words*. Moscow: AST Publ.; Astrel Publ., 2003, 768 p. (In Russian)

Ильина 2013 — Ilina O. A. Lexical Minimum for Language of the Specialty "Robotics" as the Basis for the Formation of Linguocommunicative Competence of Foreign Master Students. *Gumanitarnyj vestnik*. 2013, 2 (4): 1–16. (In Russian)

Лапошина 2021 — Laposhina A. N. «Is Out of the Vocabulary List», What Does It Mean? Calculating the Percentage of Unknown Words in a Text for Foreign Students Considering Their Derivatives. *Prepodavatel XXI vek.* 2021, (4 (2)): 473–483. (In Russian)

Ляшевская, Шаров 2009 — Lyashevskaya O. N., Sharov S. A. *Modern Russian Frequency Dictionary (based on the data from the Russian National Corpus)*. Moscow: Azbukovnik Publ., 2009. (In Russian)

Маркина 2011 — Markina E. I. *Linguodidactic Basic for the Lexical Minima for Russian as a Foreign Language (For Different Levels and Profiles of Education)*. Dis. ... PhD in Pedagogy. Moscow, 2011. 235 p. (In Russian)

Сидорова, Шматко 2019 — Sidorova M. Iu., Shmatko A. S. From "Minimized Word List" to "Lexico-Grammatical Base": A New Approach Towards Representation of the Language of aScientific Discipline. *Mir russkogo slova*. 2019, (3): 83–91. (In Russian)

Фрумкина 1967 — Frumkina R. M. Dictionary-minimum and text understanding. *Russki jazik za rubezhom*. 1967, (2): 15–21. (In Russian)

Штейнфельдт 1963 — Shteifeldt E. A. *Frequency Dictionary of a Modern Russian Literary Language: 2500 Most Common Words*. Tallin, 1963. 316 p. (In Russian)

Blinova et al. 2020 — Blinova O. V., Tarasov N. A., Modina V. V., Blekanov I. S. Modeling Lemma Frequency Bands for Lexical Complexity Assessment of Russian Texts. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2020"*. 2020, 19(26): 76–92.

Francois, Fairon 2012 — Francois T., Fairon C. An 'AI readability' formula for French as a foreign language. *Proceedings of the EMNLP and CoNLL 2012, Jeju Island, Korea, 12–14 July 2012*. P. 466–477.

Gries 2008 — Gries S. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*. 2008, (13): 403–437.

Kilgarriff 2012 — Kilgarriff, A. Getting to know your corpus. In *International conference on text, speech and dialogue.* Berlin, Heidelberg: Springer. 2012. P. 3–15.

Muñoz 2014 — Muñoz C. Contrasting Effects of Starting Age and Input on the Oral Performance of Foreign Language Learners. *Applied Linguistics*. 2014, (35): 463–482.

Nation, Waring 1997 — Nation P., Waring R. Vocabulary Size, Text Coverage and Word Lists. In *Vocabulary: Description, Acquisition, and Pedagogy*. Editors Schmitt N., McCarthy M. Cambridge: Cambridge University Press, 1997. P. 6–19.

Sharoff et al. 2013 — Sharoff S. A., Umanskaya E., Wilson J. *A Frequency Dictionary of Russian: Core vocabulary for learners.* New York: Routedge, 2013.

Volodina et al. 2013 — Volodina E., Pijetlovic D., Pilán I., Johansson S. Towards a gold standard for Swedish CEFR-based ICALL. *Proceedings of the second workshop on NLP for computer-assisted language learning at NODALIDA*. 2013. P. 48–65.