УЧЕНЫЕ ЗАПИСКИ ПЕТРОЗАВОДСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
Proceedings of Petrozavodsk State University

Т. 46, № 3. С. 107–113 2024

Научная статья Теоретическая, прикладная и сравнительно-сопоставительная лингвистика
DOI: 10.15393/uchz.art.2024.1031
EDN: GHLQMO
УДК 81.33

АНТОНИНА НИКОЛАЕВНА ЛАПОШИНА
кандидат педагогических наук, научный сотрудник лаборатории когнитивных и лингвистических исследований
Государственный институт русского языка им. А. С. Пушкина
(Москва, Российская Федерация)
ORCID 0000-0003-0693-7657; antonina.laposhina@gmail.com

ЕДИНЫЙ ЧАСТОТНЫЙ ПРОФИЛЬ СЛОВА КАК ИНСТРУМЕНТ ОТБОРА ЛЕКСИКИ В ОБУЧЕНИИ ЯЗЫКУ

Аннотация. Частотность слова признается одним из важнейших показателей в ряде прикладных задач, в частности в отборе лексики для преподавания русского языка. С другой стороны, среди возможных проблем корпусно-ориентированного отбора лексики отмечаются тематические сдвиги имеющихся больших корпусов русского языка, использование в качестве источника информации корпуса текстов, нерелевантных целевой возрастной группе, расхождения частотности слова в текстах для носителей языка с его методической ценностью в иноязычной аудитории. Таким образом, актуальной проблемой остается поиск новых, гибких способов учета частотности слова в зависимости от конкретной аудитории и прикладной задачи. Цель настоящей статьи состоит в анализе истории учета данных о встречаемости слова в языке в контексте преподавания русского языка, описании основных современных источников информации о частотности слова, а также презентация концепции единого частотного профиля слова. Анализ истории вопроса учета частотности лексики для обучения языку выявил тенденции, во-первых, стремления получения более достоверных данных за счет увеличения объема корпуса, во-вторых, поиска новых форм представления частотной информации, объединяющих информацию из нескольких источников. В этой связи в статье предложена концепция единого частотного профиля слова, аккумулирующего информацию о его частотности по нескольким словарям и корпусам, а также уровне сложности слова. Подобный сервис дает возможность пользователю – преподавателю русского языка или исследователю получить корпусные данные, максимально релевантные конкретной задаче.

Ключевые слова: частотность, частотный словарь, отбор лексики, лексический минимум, корпусная лингводидактика, русский язык как иностранный

Благодарности. Работа выполнена при финансовой поддержке госзадания, проект FZNM-2020-0005 «Трансформация когнитивной и коммуникативной деятельности человека в условиях современной цифровой среды».

Для цитирования: Лапошина А. Н. Единый частотный профиль слова как инструмент отбора лексики в обучении языку // Ученые записки Петрозаводского государственного университета. 2024. Т. 46, № 3. С. 107–113. DOI: 10.15393/uchz.art.2024.1031

ВВЕДЕНИЕ

Частотность слова, его употребимость в языке, присутствует в большинстве современных классификаций критериев отбора лексики для изучающих РКИ. Предметом для дискуссии может являться его приоритетность в ряду других критериев (таких как, например, методическая ценность слова, стилистическая нейтральность, словообразовательный потенциал и др.), однако сама необходимость учета частотной информации не вызывает разногласий экспертов[1] [2]. Основная идея подхода учета информации об употребимости слова в контексте преподавания языка связана с предположением, что частотные слова более вероятно будут встречаться учащимся в аутентичных материалах, поэтому рациональнее всего изучить их в первую очередь [10]. Так, согласно исследованиям, знание 1000 наиболее употребимых слов русского языка позволяет понимать 70–80 % текста [1]; 3500 лексем из Лексического минимума 1985 года покрывают в среднем 82 % текста. С другой стороны, детальные корпусные исследования показывают, что частотная информация ценна ровно настолько, насколько репрезентативен, сбалансирован и релевантен конкретной задаче корпус текстов, на которых

производится подсчет частотности. Среди трудностей корпусно-ориентированного подхода к отбору лексики в учебных целях называют возможные тематические сдвиги, которые могут приводить к искажению данных [6]; использование источников информации о частотности, нерелевантных возрастной группе учащихся [4]; несовпадение методической ценности слова и его употребимости в текстах для носителей языка. Все вышеперечисленные факторы доказывают опасность ориентации лишь на один частотный словарь или словник в качестве основной меры его методической ценности для обучения русскому языку. В качестве возможного решения предлагается принципиально новая форма презентации ключевой лексики для обучения РКИ: онлайн-ресурс, аккумулирующий информацию о частотности лексемы по нескольким словарям и корпусам и об уровне ее сложности в терминах ТРКИ, то есть создающий для каждого слова единый частотный профиль.

Целью настоящей статьи является анализ истории учета данных о встречаемости слова в языке в контексте преподавания РКИ, описание основных современных источников информации о частотности слова, релевантных для общего курса русского языка как иностранного, а также представление концепции единого частотного профиля слова.

КОРПУСНО-ОРИЕНТИРОВАННЫЙ ПОДХОД К ОТБОРУ ЛЕКСИКИ ДЛЯ ПРЕПОДАВАНИЯ РУССКОГО ЯЗЫКА: ИСТОРИЯ ВОПРОСА

История создания учебных списков наиболее востребованной лексики на основе частотной информации неразрывно связана с историей становления самой корпусной лингвистики и увеличением мощностей вычислительной техники. Первым частотным списком слов русского языка в литературе признается словарь Г. Йоссельсона, изданный для преподавания русского языка в США [8]. Он состоит из 1700 единиц «ходовых слов русского литературного языка», подсчет которых проводился на произведениях дореволюционной художественной литературы.

Значительное место в разработке лингвостатистического подхода к формированию лексических списков для инофонов занимают труды русистов 1960–1980-х годов по созданию учебных материалов для школьников республик, входящих в состав СССР. В 1963 году вышел словарь Э. Штейнфельд, созданный как основа для словаря-минимума, предназначенного ученикам начальной эстонской школы[2]. Для реализации идеи словаря была создана сбалансированная коллекция текстов, релевантных для конкретной аудитории школьников, поэтому выборка включала образцы оригинальной (А. Гайдар, Н. Носов) и переводной (М. Твен, Г. Х. Андерсен) детской художественной литературы, русской классической литературы, молодежных газет и журналов, а также транскрипции радиопередач для молодежи. Общий объем выборки текстов составил 400 тысяч слов.

П. И. Харакоз издает уникальный частотный словарь детской бытовой разговорной речи на выборке из 312 тыс. слов и частотный словарь учебников для русской начальной школы на базе 408 тыс. слов, на основе которых предлагает определять лексическую базу начального курса русского языка для детей в киргизских школах[3].

Важной вехой в развитии направления является создание частотного словаря под редакцией Л. Н. Засориной[4], впервые перешагнувшего планку объема текстов в 1 млн словоупотреблений. На его базе был создан учебный словарь для англоговорящих студентов, включающий 10 000 самых употребимых слов русского языка[5].

Отдельное место в истории создания корпусно-ориентированных лексических списков занимает работа коллектива под руководством В. В. Морковкина «Система лексических минимумов», созданная на основе сопоставления информации по восьми известным существующим на тот момент частотным словарям: как общим, так и специальным, например частотному словарю языка газеты и др. Издание состоит из двух частей: первая содержит сравнительные таблицы частотности слов по восьми словарям, вторая – полученный на основе этих данных лексический минимум, разделенный на отрезки по 1 тыс. слов[6].

Стоит отметить, что все вышеперечисленные словари и минимумы создавались в ручном режиме или с минимальной помощью вычислительных алгоритмов.

СОВРЕМЕННЫЕ ИСТОЧНИКИ ИНФОРМАЦИИ О ЧАСТОТНОСТИ СЛОВА, РЕЛЕВАНТНЫЕ ДЛЯ ЗАДАЧ РКИ

Современный этап создания корпусно-ориентированных списков начинается с 2000-х годов и характеризуется, во-первых, значительным скачком технологий автоматической обработки текстовых данных, во-вторых, появлением доступных программ для статистической обработки текста, что дало возможность собирать и анализировать узкоспециальные коллекции текстов под любые учебные задачи. В табл. 1 приведены примеры общих и узкоспециальных коллекций текстов и лексических списков на их основе, релевантных задаче обучения общему курсу русского языка как иностранного. Обычным шрифтом отмечены общие большие корпусы текстов, курсивом – малые узкоспециальные корпусы текстов.

Таблица 1. Возможные источники информации о частотности лексики для взрослой и детской аудитории

Table 1. Possible sources of information on word frequency for adult and child audiences

Источник	Объем источника, словоупотреблений	Частотный список / словарь на основе источника	Объем частотного списка, слов
Взрослая аудитория			
Национальный корпус русского языка	98 млн	Новый частотный словарь русской лексики[7]	20 000
Корпус интернет-страниц, Internet corpus	150 млн	A Frequency Dictionary of Russian: core vocabulary for learners[8]	5 000
Корпус интернет-страниц, Internet corpus	150 млн	Русский список KELLY, 2014[9]	9 000
Корпус текстов из учебников РКИ общего курса, RuFoLa	*665 тыс.*	*Частотный список RFLList, проходит редакцию*	*10 000*
Детская аудитория			
Корпус литературы для детей, ДетКорпус	68 млн	Частотный список на его основе, доступен в электронном виде[10]	50 000
Корпус учебников русского языка для младших школьников, TIRTEC	*1,7 млн*	*Частотный список, доступен в электронном виде*[11]	*5 000*

Создание в 2005 году Национального корпуса русского языка сделало возможным появление Нового частотного словаря русской лексики на его основе, базирующегося на выборке почти в 100 млн словоупотреблений, содержащей сбалансированную коллекцию текстов разных типов, жанров и стилей, в том числе и тексты русского зарубежья. Словарь содержит информацию о встречаемости лексемы по подкорпусам художественной литературы и публицистики по трем временным периодам создания: 1950–1960, 1970–1980 и 1990–2000-е годы, а также отдельно доступен частотный список наиболее употребимых слов живой устной речи. Пример данных, доступных в словаре, представлен в табл. 2.

Информация из Нового частотного словаря дает возможность получить, кроме обобщенного значения частотности слова в языке, более детальную информацию о росте (*проблема*) или спаде употребимости слова, а также сравнивать встречаемость слова в текстах различных стилей и форм (*проблема* более характерна для публицистической, *броситься* – для художественной, частица *вот* встречается в устной речи в девять раз чаще, чем в письменной). Однако, насколько нам известно, до настоящего времени не представлено учебных словарей для изучающих русский язык как иностранный, базирующихся на этом материале.

Таблица 2. Пример информации о частотности слова по разным разделам Нового частотного словаря русской лексики

Table 2. Example of word frequency information across different sections of the New Frequency Dictionary of Russian Vocabulary

Подкорпус	Нормализованная частотность лексемы, ipm		
	проблема (сущ.)	*броситься (глагол)*	*вот (частица)*
Общее значение ipm	475	60	1785
Худ. лит. 1950–1960-е годы	37	126	2815
Худ. лит. 1970–1980-е годы	70	115	2670
Худ. лит. 1990–2000-е годы	153	108	2177
Публицистика 1950–1960-е годы	142	55	1251
Публицистика 1970–1980-е годы	284	51	1665
Публицистика 1990–2000-е годы	701	33	1321
Живая устная речь	309	–	15699

Корпус интернет-страниц, Internet corpus, размером в 150 млн словоупотреблений стал источником сразу для двух учебно-методических продуктов: системы словников, маркированных уровнями шкалы CEFR, и учебного частотного словаря. A Frequency Dictionary of Russian: core vocabulary for learners (русский учебный словарь), созданный для англоговорящих студентов, изучающих русский язык, содержит 5 000 самых употребимых слов русского языка. Лексика в словаре презентуется в порядке убывания частотности и снабжена аутентичными примерами употребления, выбранными из корпуса Internet Corpus, и переводом на английский язык. Пример словарной статьи словаря представлен на рис. 1.

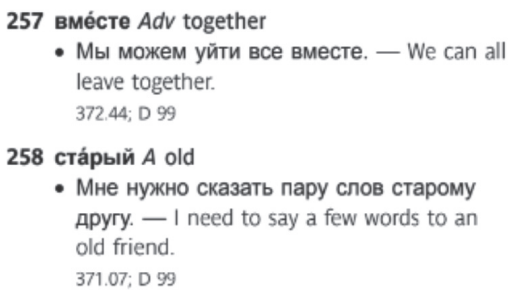
257 **вме́сте** *Adv* together
• Мы можем уйти все вместе. — We can all leave together.
372.44; D 99
258 **ста́рый** *A* old
• Мне нужно сказать пару слов старому другу. — I need to say a few words to an old friend.
371.07; D 99

Рис. 1. Пример словарной статьи A Frequency Dictionary of Russian: Core Vocabulary for Learners

Figure 1. Example of a dictionary entry in A Frequency Dictionary of Russian: Core Vocabulary for Learners

Русский список KELLY – часть проекта создания корпусно-ориентированных списков лексики для девяти языков, градуированных по уровням CEFR. При их создании центральным критерием стала частотность слова по корпусу интернет-страниц, а дополнительными – экспертная оценка и «международность» слова, то есть его наличие в списках для других языков [9].

Примером узкоспециального корпуса для задач преподавания РКИ является корпус Rufola (Russian as a Foreign Language corpus), который содержит тексты из пособий общего курса по РКИ для взрослых учащихся и отражает возможную методическую ценность и уровень сложности слова с позиции преподавания русского языка как иностранного [3]. На его основе были составлены списки, иллюстрирующие частотность лексемы в учебниках РКИ разных уровней сложности CEFR. Эта информация позволяет не только получить наглядную картину постепенного введения слова в лексикон студента, но и определить уровень согласованности мнений различных авторов о методической ценности того или иного слова [7]. Так, на рис. 2 иллюстрируется потенциал информации о частотности слова в учебниках РКИ разных уровней: представлен график встречаемости нескольких существительных разных уровней сложности – *автобус, страна, праздник* и *возможность*.

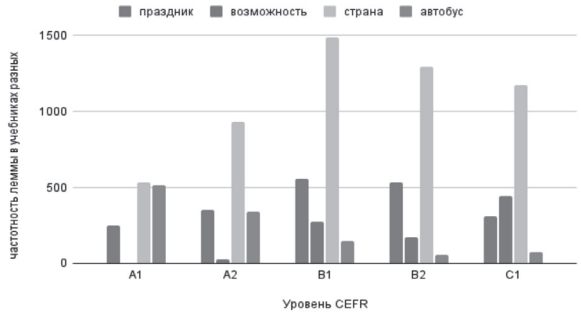

Рис. 2. Частотность слов по корпусу текстов из учебников РКИ

Figure 2. Word frequencies in the corpus of texts from textbooks on Russian as a foreign language

Анализ встречаемости слова в учебниках различных уровней помогает визуализировать момент введения слова в лексикон студента: *возможность* имеет единичные упоминания в текстах A2, однако частотность значительно повышается от уровня B1; так называемая лексика выживания, которая вводится и отрабатывается на начальных этапах, а затем имеет тенденцию к приближению частотности к нормальным значениям по корпусу для носителей языка (*автобус*); лексика, демонстрирующая более равномерное распределение в текстах различных уровней (*праздник*); лексика, вводимая с элементарных уровней и имеющая тенденцию повышения частотности до значений по корпусу для носителей языка (*страна*).

Отдельную группу составляют источники информации о частотности слов в текстах, адресованных детям и подросткам. Здесь примером большого, общего корпуса может служить Деткорпус – это аннотированный корпус русской литературы для детей, включающий более 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 отдельный подкорпус нехудожественной литературы для детей.

Наконец, корпус учебников русского языка для младших школьников TIRTEC составляет картину лексики, презентуемой в учебниках русского языка для начинающих изучать русский язык, однако с разной степенью начальной языковой подготовки: в коллекции содержатся как учебники для детей-инофонов, для детей-билингвов и для российских школьников – носителей языка. На основе этих текстов был создан единый частотный список самых употребимых слов для детей младшего школьного возраста [5].

ВОЗМОЖНЫЕ ФОРМЫ ПРЕДСТАВЛЕНИЯ ЧАСТОТНОЙ ИНФОРМАЦИИ О СЛОВЕ ДЛЯ ИЗУЧАЮЩИХ РКИ

Помимо вопросов качества частотных данных, для практического использования этой информации в преподавательской и исследовательской деятельности немаловажным является вопрос формы презентации этой информации и возможностей навигации по ней.

С одной стороны, «Система лексических минимумов» и ее переиздания представляют собой уникальную форму словаря, предоставляющего одновременно подробную частотную информацию и лексический минимум для иностранных учащихся, печатная форма издания обусловила двухступенчатую систему поиска слова: при поиске конкретного слова следует сначала найти его в алфавитном списке и узнать страницу, на которой расположена таблица со сравнительной информацией о частотности слова, что, безусловно, затрудняет поиск и сопоставительный анализ лексических единиц по словарю.

Наличие электронной версии Нового частотного словаря русской лексики значительно упростило поиск по словарю, однако все еще не дает полной картины востребованности слова в разных целевых аудиториях.

В последние десятилетия также отмечается выраженная тенденция совмещения информации из словарей нескольких типов для удобства пользования: так, новые версии крупнейших толковых словарей английского языка – Collins, Longman, Macmillan – визуализируют информацию об употребимости слова в виде специальных значков, упомянутый выше A Frequency Dictionary of Russian: core vocabulary for learners агрегирует информацию о частотности, толковании и переводе слова.

Наиболее важными составляющими удобства учета частотной информации для практических нужд РКИ представляются, помимо электронной формы списка и возможности поиска и сравнения информации о нескольких лексемах, ее совмещение с существующими уровневыми списками лексики для иностранных учащихся, а также возможность сопоставления данных о частотности из разных источников.

КОНЦЕПЦИЯ ЕДИНОГО ЧАСТОТНОГО ПРОФИЛЯ СЛОВА

В качестве возможной реализации новой формы частотного словаря ключевой лексики для обучения РКИ предлагается единый электронный частотный профиль слова, аккумулирующий информацию о частотности лексемы по нескольким источникам: Новому частотному словарю русской лексики, корпусу литературы для детей Детkорпус, корпусу текстов из учебников русского языка для различных групп учащихся (учебники для взрослых, для детей-инофонов, детей-билингвов, детей из школ с русским языком обучения), и уровне сложности слова в терминах CEFR.

Единый частотный профиль слова представляет собой открытый онлайн-сервис, который обобщает и демонстрирует пользователю информацию о присутствии лексемы в различных частотных словарях, нормализованные значения частотности и уровень сложности слова в терминах CEFR[12]. Интерфейс состоит из строки поиска, куда можно ввести до пяти слов. Если слово присутствует в базе информации, пользователю становится доступен результат с обобщенной информацией о слове. Пример возможного результата поиска приведен в табл. 3.

Организация информации, проиллюстрированная в табл. 3, позволяет получить как самую общую информацию о частотности слова: пункты «частотность по текстам для взрослых» и «частотность по текстам для детей» демонстрируют вхождение слова в верхушку частотных списков для взрослых и детей соответственно, так и более детальную информацию об употребимости слова по разным корпусам текстов. Так, несмотря на то что оба слова *страна* и *государство* входят в лексический минимум ТРКИ элементарного минимума, полезно понимать, что *государство* употребляется более чем в два раза реже, чем *страна*, в текстах для носителей, более чем в три раза реже в текстах для иностранных студентов, и в десятки раз меньше в текстах, адресованных детям. Слово *лагерь*, напротив, маркировано уровнем C1, однако его употребимость в литературе и учебных текстах для детей может служить аргументом для включения его на более ранних этапах обучения. Слово *мороженое* демонстрирует пример бытовой лексики, которая не является частотной в текстах для носителей языка,

Таблица 3. Виды информации, доступной в сервисе «Единый частотный профиль слова»

Table 3. Types of information available with the Unified Word Frequency Profile tool

Лемма	Страна	Государство	Лагерь	Мороженое
Часть речи	сущ.	сущ.	сущ.	сущ.
Частотность по текстам для взрослых	очень частотно (ТОП-1000)	очень частотно (ТОП-1000)	очень частотно (ТОП-2000)	частотно (ТОП-4000)
Частотность по текстам для детей	очень частотно (ТОП-1000)	частотно (ТОП-3000)	очень частотно (ТОП-1000)	очень частотно (ТОП-1000)
Уровень по лексическим минимумам ТРКИ	A1	A1	C1	A1
Уровень по русскому списку KELLY	A1	A2	B1	A1
Частотность по Новому частотному словарю русской лексики, ipm	725.7	326.4	88.6	18.8
Частотность по корпусу литературы для детей ДетКорпус, ipm	170	28	130	47
Частотность по корпусу учебников РКИ RuFoLa, ipm	887	190	29	56
Частотность по корпусу учебников русского языка для детей младшего школьного возраста TIRTEC, ipm				
Дети-инофоны	625.59	9.48	208.53	412.32
Дети-билингвы	733.33	64.86	73.87	111.71
Дети – российские школьники	302.86	18.72	31.94	33.04

однако признается методически ценной для формирования лексического ядра тематической области «Еда и напитки», поэтому отмечается уровнем А1 в лексических минимумах, встречается в учебниках РКИ в несколько раз чаще, чем в текстах для носителей, при этом частотность в текстах для детей у слова ожидаемо выше.

ЗАКЛЮЧЕНИЕ

Практика учета частотности слова в преподавании РКИ служит прекрасным примером развития научной области обучения языкам, основанного на данных: с одной стороны, технические возможности вычислительной техники уже позволяют собирать и обрабатывать большие корпусы текстов, с другой стороны, реальное использование информации о частотности лексики в обучении русскому языку может быть осложнено необходимостью ее поиска и сравнения по различным источникам.

Представленная концепция и первая версия единого частотного профиля слова призваны помогать принимать методические решения по отбору лексики к уроку или учебному пособию на основе корпусных данных, максимально релевантных для конкретной аудитории или учебной задачи. Сервис может быть полезен для исследователей, авторов словарей, преподавателей-практиков и авторов учебных материалов по русскому языку.

Среди дальнейших векторов развития отметим обогащение базы информацией о частотности слова в устной и письменной речи студентов-иностранцев для выявления наиболее востребованной продуцируемой студентами лексики.

ПРИМЕЧАНИЯ

[1] 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.
[2] Штейнфельдт Э. А. Частотный словарь современного русского литературного языка: 2500 наиболее употребительных слов: Пособие для преподавателей рус. яз. Таллин, 1963. 316 с.
[3] Харакоз П. И. Частотный словарь современного русского языка. Фрунзе, 1971. 180 с.
[4] Частотный словарь русского языка / Под ред. Л. Н. Засориной. М.: Русский язык, 1977. 936 с.
[5] Brown N. J. Russian learners' dictionary: 10 000 words in frequency order. London: Routledge, 1996. 429 p.
[6] Лексические минимумы современного русского языка / В. В. Морковкин, Ю. А. Сафьян, Е. М. Степанова, И. В. Дорофеева; Под ред. В. В. Морковкина. М.: Рус. яз., 1985. 608 с.
[7] Ляшевская О. Н., Шаров С. А. Новый частотный словарь русской лексики [Электронный ресурс]. Режим доступа: http://dict.ruslang.ru/freq.php (дата обращения 12.09.2023).
[8] Sharoff S., Umanskaya E., Wilson J. A frequency dictionary of Russian: core vocabulary for learners. London: Routledge, 2013. 400 p.
[9] Русский список KELLY [Электронный ресурс]. Режим доступа: http://corpus.leeds.ac.uk/serge/kelly/ (дата обращения 12.09.2023).
[10] Гитхаб [Электронный ресурс]. Режим доступа: https://github.com/Digital-Pushkin-Lab/Russian_frequency_lists (дата обращения 12.09.2023).
[11] DigitalPushkin [Электронный ресурс]. Режим доступа: https://digitalpushkin.tilda.ws/tirtec (дата обращения 12.09.2023).
[12] Текстометр. Проверка частотности слова [Электронный ресурс]. Режим доступа: https://textometr.ru/frequency-check (дата обращения 12.09.2023).

СПИСОК ЛИТЕРАТУРЫ

1. Алексеев П. М. Статистическая лексикография (типология, составление и применение частотных словарей): Учеб. пособие. Л.: ЛГПИ, 1975. 120 с.
2. Андрюшина Н. П. Лексические минимумы по русскому языку как иностранному: проблема отбора лексических и фразеологических единиц // Проблемы истории, филологии, культуры. 2001. № 3 (33). С. 648–652.
3. Лапошина А. Н. Корпус текстов учебников РКИ как инструмент анализа учебных материалов // Русский язык за рубежом. 2020. № 6 (283). С. 22–28.
4. Лапошина А. Н., Лебедева М. Ю. Смотря как считать: влияние типа корпуса на данные о частотности слова в контексте определения сложности учебных текстов для младшей школы // Языковое разнообразие в глобальном мире: Казанский международный лингвистический саммит (Казань, 15–19 ноября 2021 г.). Казань: Изд-во Казанского ун-та, 2022. Т. 1. С. 48–51.
5. Лапошина А. Н., Лебедева М. Ю. Формирование частотного словаря-минимума русского языка для детей-инофонов на основе корпусных данных // Мир русского слова. 2022. № 3. С. 90–99.
6. Шаров С. А. Не только размер имеет значение: аспекты создания частотных словарей на основе корпусов // Русский язык за рубежом. 2020. № 6 (283). С. 14–21. DOI: 10.37632/PI.2020.283.6.002
7. François T., Gala N., Watrin P., Fairon C. FLELex: a graded lexical resource for French foreign learners // Proceedings of the International Conference on Language Resources and Evaluation (LREC 2014). 2014. P. 3766–3773.
8. Josselson H. H. The Russian word count and frequency analysis of grammatical categories of standard literary Russian. Detroit, 1953. 277 p.
9. Kilgarriff A., Charalabopoulou F., Gavrilidou M., Johannessen J. B., Khalil S., Kokkinakis S. J., Lew R., Sharoff S., Vadlapudi R., Volodina E. Corpus-based

vocabulary lists for language learners for nine languages // Language Resources and Evaluation. 2014. № 48. P. 121–163. DOI: 10.1007/s10579-013-9251-2
10. Nation P., Waring R. Vocabulary: description, acquisition and pedagogy. Cambridge: Cambridge University Press, 1997. P. 6–19.

Поступила в редакцию 05.02.2024; принята к публикации 01.03.2024

Original article

Antonina N. Laposhina, Cand. Sc. (Pedagogics), Research Associate, Laboratory of Cognitive and Linguistic Studies, Pushkin State Russian Language Institute (Moscow, Russian Federation)
ORCID 0000-0003-0693-7657; antonina.laposhina@gmail.com

UNIFIED WORD FREQUENCY PROFILE AS A TOOL FOR VOCABULARY SELECTION FOR LANGUAGE LEARNING

Abstract. Word frequency is widely acknowledged as one of the most important factors in various practical applications, including vocabulary selection for teaching the Russian language. However, when it comes to the process of corpus-based lexicon selection, there exists a number of potential challenges. These challenges include shifts in themes within the vast existing corpora of the Russian language, the utilization of text corpora that may not be relevant to the target age group, and disparities between a word's frequency in texts for native speakers and its pedagogical value for a foreign-language audience. Consequently, the quest for innovative and adaptable approaches to incorporating word frequency into consideration depending on the specific audience and task at hand remains a pressing issue. The article aims to examine the history of utilizing data on word frequency in language instruction within the context of Russian language teaching, to outline the primary contemporary sources of information on word frequency, and to introduce the concept of a unified word frequency profile. Examining the historical progression of using word frequency data in Russian language teaching reveals two trends: firstly, efforts to acquire more reliable data by expanding corpora, and secondly, the pursuit of novel methods for representing frequency information that integrate data from multiple sources. In this regard, the article presents the concept of a unified word frequency profile that consolidates information on a word's frequency across various dictionaries and corpora, along with its difficulty level. Such a tool empowers users – be it Russian language teachers or researchers – to access corpus data that is most relevant to their specific tasks and requirements.

Keywords: word frequency, frequency dictionary, vocabulary selection, lexical minimum, corpus-oriented linguodidactics, Russian as a foreign language

Acknowledgements. The article was financially supported as part of the state research assignment (project FZNM-2020-0005 "Transformation of human cognitive and communicative activities in the modern digital environment").

For citation: Laposhina, A. N. Unified Word Frequency Profile as a tool for vocabulary selection for language learning. *Proceedings of Petrozavodsk State University.* 2024;46(3):107–113. DOI: 10.15393/uchz.art.2024.1031

REFERENCES

1. Alekseev, P. M. Statistical lexicography (typology, compilation, and application of frequency dictionaries): Textbook. Leningrad, 1975. 120 p. (In Russ.)
2. Andryushina, N. P. Basic dictionary for Russian as a second language (the choice of words and set phrases). *Problems of History, Philology and Culture.* 2001;3(33):648–652. (In Russ.)
3. Laposhina, A. N. A corpus of Russian textbook materials for foreign students as an instrument of an educational content analysis. *Russian Language Abroad.* 2020;6(283):22–28. (In Russ.)
4. Laposhina, A. N., Lebedeva, M. Yu. Depending on how you calculate it: the influence of corpus type on word frequency data in the context of determining the complexity of educational texts for primary school. *Language diversity in the global world: Kazan International Linguistics Summit (Kazan, 15–19 November 2021).* Kazan, 2022. Vol. 1. P. 48–51. (In Russ.)
5. Laposhina, A. N., Lebedeva, M. Yu. Developing a Russian frequency core vocabulary list for foreign children based on corpus data. *The World of Russian Word.* 2022;3:90–99. (In Russ.)
6. Sharoff, S. A. Not only size matters: issues in creating frequency dictionaries from corpora. *Russian Language Abroad.* 2020;6(283):14–21. DOI: 10.37632/PI.2020.283.6.002 (In Russ.)
7. François, T., Gala, N., Watrin, P., Fairon, C. FLELex: a graded lexical resource for French foreign learners. *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2014).* 2014:3766–3773.
8. Josselson, H. H. The Russian word count and frequency analysis of grammatical categories of standard literary Russian. Detroit, 1953. 277 p.
9. Kilgarriff, A., Charalabopoulou, F., Gavrilidou, M., Johannessen, J. B., Khalil, S., Kokkinakis, S. J., Lew, R., Sharoff, S., Vadlapudi, R., Volodina, E. Corpus-based vocabulary lists for language learners for nine languages. *Language Resources and Evaluation.* 2014;48:121–163. DOI: 10.1007/s10579-013-9251-2
10. Nation, P., Waring, R. Vocabulary: description, acquisition and pedagogy. Cambridge, 1997. P. 6–19.

Received: 5 February 2024; accepted: 1 March 2024